The Annals of Statistics
2015, Vol. 43, No. 4, 1498–1534
DOI: 10.1214/14-AOS1307

QUADRO: A SUPERVISED DIMENSION REDUCTION METHOD VIA RAYLEIGH QUOTIENT OPTIMIZATION

BY JIANQING FAN[*,1], ZHENG TRACY KE[†,1], HAN LIU[*,2] AND LUCY XIA[*,1]

Princeton University[*] and University of Chicago[†]

We propose a novel Rayleigh quotient based sparse quadratic dimension reduction method—named QUADRO (Quadratic Dimension Reduction via Rayleigh Optimization)—for analyzing high-dimensional data. Unlike in the linear setting where Rayleigh quotient optimization coincides with classification, these two problems are very different under nonlinear settings. In this paper, we clarify this difference and show that Rayleigh quotient optimization may be of independent scientific interests. One major challenge of Rayleigh quotient optimization is that the variance of quadratic statistics involves all fourth cross-moments of predictors, which are infeasible to compute for high-dimensional applications and may accumulate too many stochastic errors. This issue is resolved by considering a family of elliptical models. Moreover, for heavy-tail distributions, robust estimates of mean vectors and covariance matrices are employed to guarantee uniform convergence in estimating non-polynomially many parameters, even though only the fourth moments are assumed. Methodologically, QUADRO is based on elliptical models which allow us to formulate the Rayleigh quotient maximization as a convex optimization problem. Computationally, we propose an efficient linearized augmented Lagrangian method to solve the constrained optimization problem. Theoretically, we provide explicit rates of convergence in terms of Rayleigh quotient under both Gaussian and general elliptical models. Thorough numerical results on both synthetic and real datasets are also provided to back up our theoretical results.

1. Introduction. Rapid developments of imaging technology, microarray data studies and many other applications call for the analysis of high-dimensional binary-labeled data. We consider the problem of finding a "nice" projection $f : \mathbb{R}^d \to \mathbb{R}$ that embeds all data into the real line. A projection such as f has applications in many statistical problems for analyzing high-dimensional binary-labeled data, including:

Received November 2013; revised December 2014.

[1]Supported in part by NSF Grants DMS-12-06464 and DMS-14-06266 and NIH Grants R01-GM100474 and R01-GM072611.

[2]Supported in part by NSF Grants III-1116730, NSF III-1332109, an NIH sub-award and a FDA sub-award from Johns Hopkins University and an NIH-subaward from Harvard University.

MSC2010 subject classifications. Primary 62H30; secondary 62G20.

Key words and phrases. Classification, dimension reduction, quadratic discriminant analysis, Rayleigh quotient, oracle inequality.

- *Dimension reduction*: f provides a data reduction tool for people to visualize the high-dimensional data in a one-dimensional space.
- *Classification*: f can be used to construct classification rules. With a carefully chosen set $A \subset \mathbb{R}$, we can classify a new data point $\mathbf{x} \in \mathbb{R}^d$ by checking whether or not $f(\mathbf{x}) \in A$.
- *Feature selection*: when $f(\mathbf{x})$ only depends on a small number of coordinates of $\mathbf{x}$, this projection selects just a few features from numerous observed ones.

A natural question is what kind of f is a "nice" projection? It depends on the goal of statistical analysis. For classification, a good f should yield to a small classification error. In feature selection, different criteria select distinct features, and they may suit different real problems. In this paper, we propose using the following criterion for finding f:

> Under the mapping f, the data are as "separable" as possible between two classes, and as "coherent" as possible within each class.

It can be formulated as to maximize the *Rayleigh quotient* of f. Suppose all data are drawn independently from a joint distribution of $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^d$, and $Y \in \{0, 1\}$ is the label. The *Rayleigh quotient* of f is defined as

$$\mathrm{Rq}(f) \equiv \frac{\mathrm{var}\{\mathbb{E}[f(\mathbf{X})|Y]\}}{\mathrm{var}\{f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})|Y]\}}. \tag{1}$$

Here, the numerator is the variance of $\mathbf{X}$ explained by the class label, and the denominator is the remaining variance of $\mathbf{X}$. Simple calculation shows that $\mathrm{Rq}(f) = \pi(1-\pi)R(f)$, where $\pi \equiv \mathbb{P}(Y = 0)$ and

$$R(f) \equiv \frac{\{\mathbb{E}[f(\mathbf{X})|Y = 0] - \mathbb{E}[f(\mathbf{X})|Y = 1]\}^2}{\pi\,\mathrm{var}[f(\mathbf{X})|Y = 0] + (1-\pi)\,\mathrm{var}[f(\mathbf{X})|Y = 1]}. \tag{2}$$

Our goal is to develop a data-driven procedure to find $\hat{f}$ such that $\mathrm{Rq}(\hat{f})$ is large, and $\hat{f}$ is sparse in the sense that it depends on few coordinates of $\mathbf{X}$.

The Rayleigh quotient, as a criterion for finding a projection f, serves different purposes. First, for dimension reduction, it takes care of both variance explanation and label explanation. In contrast, methods such as principal component analysis (PCA) only consider variance explanation. Second, when the data are normally distributed, a monotone transform of the Rayleigh quotient approximates the classification error; see Section 6. Therefore, an f with a large Rayleigh quotient enables us to construct nice classification rules. In addition, it is a convex optimization to maximize the Rayleigh quotient among linear and quadratic f (see Section 3), while minimizing the classification error is not. Third, with appropriate regularization, this criterion provides a new feature selection tool for data analysis.

The criterion (1), initially introduced by Fisher (1936) for classification, is known as Fisher's linear discriminant analysis (LDA). In the literature of sufficient dimension reduction, the sliced inverse regression (SIR) proposed by Li (1991) can

also be formulated as maximizing (1), where Y can be any variable not necessarily binary. In both LDA and SIR, f is restricted to be a linear function, and the dimension d cannot be larger than n. In this sense, our work compares directly to various versions of LDA and SIR generalized to nonlinear, high-dimensional settings. We provide a more detailed comparison to the literature in Section 8, but preview here the uniqueness of our work. First, we consider a setting where $\mathbf{X}|Y$ has an elliptical distribution and f is a quadratic function, which allows us to derive a simplified version of (1) and gain extra statistical efficiency; see Section 2 for details. This simplified version of (1) was never considered before. Furthermore, the assumption of conditional elliptical distribution does not satisfy the requirement of SIR and many other dimension reduction methods [Cook and Weisberg (1991), Li (1991)]. In Section 1.2, we explain the motivation of the current setting. Second, we utilize robust estimators of mean and covariance matrix, while many generalizations of LDA and SIR are based on sample mean and sample covariance matrix. As shown in Section 4, the robust estimators adapt better to heavy tails on the data. It is worth noting that QUADRO only considers the projection to a one-dimensional subspace. In contrast, more sophisticated dimension reduction methods (e.g., the kernel SIR) are able to find multiple projections $f_1, \ldots, f_m$ for $m > 1$. This reflects a tradeoff between modeling tractability and flexibility. More specifically, QUADRO achieves better computational and theoretical properties at the cost of sacrificing some flexibility.

1.1. *Rayleigh quotient and classification error.* Many popular statistical methods for analyzing high-dimensional binary-labeled data are based on classification error minimization, which is closely related to the Rayleigh quotient maximization. We summarize their connections and differences as follows:

(a) In an "ideal" setting where two classes follow multivariate normal distributions with a common covariance matrix and the class of linear functions f is considered, the two criteria are exactly the same, with one being a monotone transform of the other.

(b) In a "relaxed" setting where two classes follow multivariate normal distributions but with nonequal covariance matrices and the class of quadratic functions f (including linear functions as special cases) is considered, the two criteria are closely related in the sense that a monotone transform of the Rayleigh quotient is an approximation of the classification error.

(c) In other settings, the two criteria can be very different.

We now show (a) and (c), and will discuss (b) in Section 6.

For each f, we define a family of classifiers $h_c(\mathbf{x}) = I\{f(\mathbf{x}) < c\}$ indexed by c, where $I(\cdot)$ is the indicator function. For each given c, we define the classification error of h_c to be $\operatorname{err}(h_c) \equiv \mathbb{P}(h_c(\mathbf{X}) \neq Y)$. The classification error of f is then defined by

$$\operatorname{Err}(f) \equiv \min_{c \in \mathbb{R}} \{\operatorname{err}(h_c)\}.$$

Most existing classification procedures aim at finding a data-driven projection $\hat{f}$ such that $\mathrm{Err}(\hat{f})$ is small (the threshold c is usually easy to choose). Examples include linear discriminant analysis (LDA) and its variations in high dimensions [e.g., Cai and Liu (2011), Fan and Fan (2008), Fan, Feng and Tong (2012), Guo, Hastie and Tibshirani (2005), Han, Zhao and Liu (2013), Shao et al. (2011), Witten and Tibshirani (2011)], quadratic discriminant analysis (QDA), support vector machine (SVM), logistic regression, boosting, etc.

We now compare $\mathrm{Rq}(f)$ and $\mathrm{Err}(f)$. Let $\pi = \mathbb{P}(Y = 0)$, $\boldsymbol{\mu}_1 = \mathbb{E}(\mathbf{X}|Y = 0)$, $\boldsymbol{\Sigma}_1 = \mathrm{cov}(\mathbf{X}|Y = 0)$, $\boldsymbol{\mu}_2 = \mathbb{E}(\mathbf{X}|Y = 1)$ and $\boldsymbol{\Sigma}_2 = \mathrm{cov}(\mathbf{X}|Y = 1)$. We consider linear functions $\{f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x} + b : \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}\}$, and write $\mathrm{Rq}(\mathbf{a}) = \mathrm{Rq}(\mathbf{a}^\top\mathbf{x})$, $\mathrm{Err}(\mathbf{a}) = \mathrm{Err}(\mathbf{a}^\top\mathbf{x})$ for short. By direct calculation, when the two classes have a common covariance matrix $\boldsymbol{\Sigma}$,

$$\mathrm{Rq}(\mathbf{a}) = \pi(1-\pi)\frac{[\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}}.$$

Hence, the optimal $\mathbf{a}_R = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. On the other hand, when data follow multivariate normal distributions, the optimal classifier is $h^*(\mathbf{x}) = I\{\mathbf{a}_E^\top\mathbf{x} < c\}$, where $\mathbf{a}_E = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $c = \frac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \log(\frac{1-\pi}{\pi})$. It is observed that $\mathbf{a}_R = \mathbf{a}_E$ and the two criteria are the same. In fact, for all vectors $\mathbf{a}$ such that $\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$,

$$\mathrm{Err}(\mathbf{a}) = 1 - \Phi\biggl(\frac{1}{2}\biggl[\frac{\mathrm{Rq}(\mathbf{a})}{\pi(1-\pi)}\biggr]^{1/2}\biggr),$$

where Φ is the distribution function of a standard normal random variable, and we fix $c = \mathbf{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Therefore, the classification error is a monotone transform of the Rayleigh quotient.

When we move away from these ideal assumptions, the above two criteria can be very different. We illustrate this point using a bivariate distribution, that is, $d = 2$, with different covariance matrices. Specifically, $\pi = 0.55$, $\boldsymbol{\mu}_1 = (0, 0)^\top$, $\boldsymbol{\mu}_2 = (1.28, 0.8)^\top$, $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, 1)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(3, 1/3)$. We still consider linear functions $f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x}$ but select only one out of the two features, X_1 or X_2. Then the maximum Rayleigh quotients, by using each of the two features alone, are 0.853 and 0.923, respectively, whereas the minimum classification errors are 0.284 and 0.295, respectively. As a result, under the criterion of maximizing Rayleigh quotient, Feature 2 is selected, whereas under the criterion of minimizing classification error, Feature 1 is selected. Figure 1 displays the distributions of data after being projected to each of the two features. It shows that since data from the second class has a much larger variability at Feature 1 than at Feature 2, the Rayleigh quotient maximization favors Feature 2, although Feature 1 yields a smaller classification error.

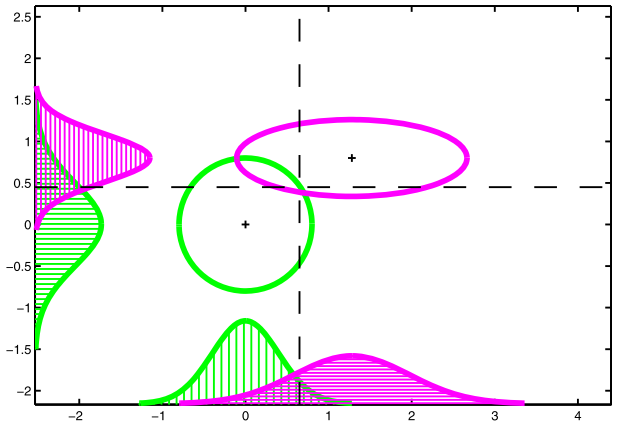


FIG. 1. *An example in* $\mathbb{R}^2$. *The green and purple represent class* 1 *and class* 2, *respectively. The ellipses are contours of distributions. Probability densities after being projected to* X_1 *and* X_2 *are also displayed. The dotted lines correspond to optimal thresholds for classification using each feature.*

1.2. *Objective of the paper.* In this paper, we consider the Rayleigh quotient maximization problem in the following setting:

- We consider sparse quadratic functions, that is, $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{\Omega}\mathbf{x} - 2\boldsymbol{\delta}^\top \mathbf{x}$, where $\mathbf{\Omega}$ is a sparse $d \times d$ symmetric matrix, and $\boldsymbol{\delta}$ is a sparse d-dimensional vector.
- The two classes can have different covariance matrices.
- Data from these two classes follow *elliptical distributions*.
- The dimension is large (it is possible that $d \gg n$).

Compared to Fisher's LDA, our setting has several new ingredients. First, we go beyond linear classifiers to enhance flexibility. It is well known that the linear classifiers are inefficient. For example, when two classes have the same mean, linear classifiers perform no better than random guesses. Instead of exploring arbitrary nonlinear functions, we consider the class of quadratic functions so that the Rayleigh quotient still has a nice parametric formulation, and at the same time it helps identify interaction effects between features. Second, we drop the requirement that the two classes share a common covariance matrix, which is a critical condition for Fisher's rule and many other high-dimensional classification methods [e.g., Cai and Liu (2011), Fan and Fan (2008), Fan, Feng and Tong (2012)]. In fact, by using quadratic discriminant functions, we take advantage of the difference of covariance matrices between the two classes to enhance classification power. Third, we generalize multivariate normal distributions to the elliptical family, which includes many heavy-tailed distributions, such as multivariate t-distributions, Laplace distributions, and Cauchy distributions. This family of distributions allows us to avoid estimating all $O(d^4)$ fourth cross-moments of d predictors in computing the variance of quadratic statistics and hence overcomes the computation and noise accumulation issues.

In our setting, Fisher's rule, that is, $\mathbf{a}_R = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, no longer maximizes the Rayleigh quotient. We propose a new method, called quadratic dimension reduction via Rayleigh optimization (QUADRO). It is a *Rayleigh-quotient-oriented procedure* and is a statistical tool for simultaneous dimension reduction and feature selection. QUADRO has several properties. First, it is a statistically efficient generalization of Fisher's linear discriminant analysis to the quadratic setting. A naive generalization involves estimation of all fourth cross-moments of the two underlying distributions. In contrast, QUADRO only requires estimating a one-dimensional kurtosis parameter. Second, QUADRO adopts rank-based estimators and robust M-estimators of the covariance matrices and the means. Therefore, it is robust to possibly heavy-tail distributions. Third, QUADRO can be formulated as a convex programming and is computationally efficient.

Theoretically, we prove that under elliptical models, the Rayleigh quotient of the estimated quadratic function $\hat{f}$ converges to population maximum Rayleigh quotient at rate $O_p(s\sqrt{\log(d)/n})$, where s is the number of important features (counting both single terms and interaction terms). In addition, we establish a connection between our method and quadratic discriminant analysis (QDA) under elliptical models.

The rest of this paper is organized as follows. Section 2 formulates Rayleigh quotient maximization as a convex optimization problem. Section 3 describes QUADRO. Section 4 discusses rank-based estimators and robust M-estimators used in QUADRO. Section 5 presents theoretical analysis. Section 6 discusses the application of QUADRO in elliptically distributed classification problems. Section 7 contains numerical studies. Section 8 concludes the paper. All proofs are collected in Section 9.

Notation. For $0 \le q \le \infty$, $|\mathbf{v}|_q$ denotes the L_q-norm of a vector $\mathbf{v}$, $|\mathbf{A}|_q$ denotes the elementwise L_q-norm of a matrix $\mathbf{A}$ and $\|\mathbf{A}\|_q$ denotes the matrix L_q-norm of $\mathbf{A}$. When $q = 2$, we omit the subscript q. $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the minimum and maximum eigenvalues of $\mathbf{A}$. $\det(\mathbf{A})$ denotes the determinant of $\mathbf{A}$. Let $I(\cdot)$ be the indicator function: for any event B, $I(B) = 1$ if B happens and $I(B) = 0$ otherwise. Let $\operatorname{sign}(\cdot)$ be the sign function, where $\operatorname{sign}(u) = 1$ when $u \ge 0$ and $\operatorname{sign}(u) = -1$ when $u < 0$.

2. Rayleigh quotient for quadratic functions. We first study the population form of Rayleigh quotient for an arbitrary quadratic function. We show that it has a simplified form under the elliptical family.

For a quadratic function

$$Q(\mathbf{X}) = \mathbf{X}^\top \mathbf{\Omega} \mathbf{X} - 2\boldsymbol{\delta}^\top \mathbf{X},$$

using (2), its Rayleigh quotient is

$$R(\mathbf{\Omega}, \boldsymbol{\delta}) = \frac{\{\mathbb{E}[Q(\mathbf{X})|Y=0] - \mathbb{E}[Q(\mathbf{X})|Y=1]\}^2}{\pi \operatorname{var}[Q(\mathbf{X})|Y=0] + (1-\pi)\operatorname{var}[Q(\mathbf{X})|Y=1]} \tag{3}$$

up to a constant multiplier. The Rayleigh quotient maximization can be expressed as

$$\max_{(\boldsymbol{\Omega},\boldsymbol{\delta})\,:\,\boldsymbol{\Omega}=\boldsymbol{\Omega}^\top} R(\boldsymbol{\Omega},\boldsymbol{\delta}).$$

2.1. *General setting.* Suppose $\mathbb{E}(\mathbf{Z}) = \boldsymbol{\mu}$ and $\operatorname{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}$. By direct calculation,

$$\begin{aligned}\mathbb{E}\big[Q(\mathbf{Z})\big] &= \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\Omega}\boldsymbol{\mu} - 2\boldsymbol{\delta}^\top\boldsymbol{\mu},\\ \operatorname{var}\big[Q(\mathbf{Z})\big] &= \mathbb{E}\big[\operatorname{tr}\big(\boldsymbol{\Omega}\mathbf{Z}\mathbf{Z}^\top\boldsymbol{\Omega}\mathbf{Z}\mathbf{Z}^\top\big)\big] - 4\mathbb{E}\big[\boldsymbol{\delta}^\top\mathbf{Z}\mathbf{Z}^\top\boldsymbol{\Omega}\mathbf{Z}\big]\\ &\quad + 4\boldsymbol{\delta}^\top\boldsymbol{\Sigma}\boldsymbol{\delta} + 4\big(\boldsymbol{\delta}^\top\boldsymbol{\mu}\big)^2 - \big\{\mathbb{E}\big[Q(\mathbf{Z})\big]\big\}^2.\end{aligned}$$

So $\mathbb{E}[Q(\mathbf{Z})]$ is a linear combination of the elements in $\{\Omega(i,j), 1 \le i \le j \le d;$ $\delta(i), 1 \le i \le d\}$, and $\operatorname{var}[Q(\mathbf{Z})]$ is a quadratic form of these elements. The coefficients in $\mathbb{E}[Q(\mathbf{Z})]$ are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ only. However, the coefficients in $\operatorname{var}[Q(\mathbf{Z})]$ also depend on all the fourth cross-moments of $\mathbf{Z}$, and there are $O(d^4)$ of them.

Let us define $M_1(\boldsymbol{\Omega},\boldsymbol{\delta}) = \mathbb{E}[Q(\mathbf{X})|Y=0]$, $L_1(\boldsymbol{\Omega},\boldsymbol{\delta}) = \operatorname{var}[Q(\mathbf{X})|Y=0]$ and $M_2(\boldsymbol{\Omega},\boldsymbol{\delta})$, $L_2(\boldsymbol{\Omega},\boldsymbol{\delta})$ similarly. Also, let $\kappa = (1-\pi)/\pi$. We have

$$R(\boldsymbol{\Omega},\boldsymbol{\delta}) = \frac{[M_1(\boldsymbol{\Omega},\boldsymbol{\delta}) - M_2(\boldsymbol{\Omega},\boldsymbol{\delta})]^2}{L_1(\boldsymbol{\Omega},\boldsymbol{\delta}) + \kappa L_2(\boldsymbol{\Omega},\boldsymbol{\delta})}.$$

Therefore, both the numerator and denominator are quadratic combinations of the elements in $\boldsymbol{\Omega}$ and $\boldsymbol{\delta}$. We can stack the $d(d+1)/2$ elements in $\boldsymbol{\Omega}$ (assuming it is symmetric) and the d elements in $\boldsymbol{\delta}$ into a long vector $\mathbf{v}$. Then $R(\boldsymbol{\Omega},\boldsymbol{\delta})$ can be written as

$$R(\mathbf{v}) = \frac{(\mathbf{a}^\top\mathbf{v})^2}{\mathbf{v}^\top\mathbf{A}\mathbf{v}},$$

where $\mathbf{a}$ is a $d' \times 1$ vector, $\mathbf{A}$ is a $d' \times d'$ positive semi-definite matrix and $d' = d(d+1)/2 + d$. $\mathbf{A}$ and $\mathbf{a}$ are determined by the coefficients in the denominator and numerator of $R(\boldsymbol{\Omega},\boldsymbol{\delta})$, respectively. Now, $\max_{(\boldsymbol{\Omega},\boldsymbol{\delta})} R(\boldsymbol{\Omega},\boldsymbol{\delta})$ is equivalent to $\max_{\mathbf{v}} R(\mathbf{v})$. It has explicit solutions. For example, when $\mathbf{A}$ is positive definite, the function $R(\mathbf{v})$ is maximized at $\mathbf{v}^* = \mathbf{A}^{-1}\mathbf{a}$. We can then reshape $\mathbf{v}^*$ to get the desired $(\boldsymbol{\Omega}^*, \boldsymbol{\delta}^*)$.

Practical implementation of the above idea is infeasible in high dimensions as it involves $O(d^4)$ cross moments of $\mathbf{Z}$. This not only poses computational challenges, but also accumulates noise in the estimation. Furthermore, good estimates of fourth moments usually require the existence of eighth moments, which is not realistic for many heavy tailed distributions. These problems can be avoided under the elliptical family, as we now illustrate in the next subsection.

2.2. *Elliptical distributions.* The elliptical family contains multivariate distributions whose densities have elliptical contours. It generalizes multivariate normal distributions and inherits many of their nice properties.

Given a $d \times 1$ vector $\boldsymbol{\mu}$ and a $d \times d$ positive definite matrix $\boldsymbol{\Sigma}$, a random vector $\mathbf{Z}$ that follows an elliptical distribution admits

$$\mathbf{Z} = \boldsymbol{\mu} + \xi \boldsymbol{\Sigma}^{1/2}\mathbf{U}, \tag{4}$$

where $\mathbf{U}$ is a random vector which follows the uniform distribution on unit sphere $\mathcal{S}^{d-1}$, and ξ is a nonnegative random variable independent of $\mathbf{U}$. Denote the elliptical distribution by $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, where g is the density of ξ. In this paper, we always assume that $\mathbb{E}\xi^4 < \infty$ and require that $\mathbb{E}(\xi^2) = d$ for the model identifiability. Then $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{Z}$.

PROPOSITION 2.1. *Suppose* $\mathbf{Z}$ *follows an elliptical distribution as in* (4). *Then*

$$\mathbb{E}\big[Q(\mathbf{Z})\big] = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\Omega}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{\delta},$$

$$\operatorname{var}\big[Q(\mathbf{Z})\big] = 2(1+\gamma)\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma}) + \gamma\big[\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma})\big]^2 + 4(\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\delta})^\top\boldsymbol{\Sigma}(\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\delta}),$$

where $\gamma = \frac{E(\xi^4)}{d(d+2)} - 1$ *is the kurtosis parameter.*

The proof is given in the online supplementary material [Fan et al. (2014)]. The variance of $Q(\mathbf{Z})$ does not involve any fourth cross-moments, but only the kurtosis parameter γ. For multivariate normal distributions, ξ^2 follows a χ^2-distribution with d degrees of freedom, and $\gamma = 0$. For multivariate t-distribution with degrees of freedom $\nu > 4$, we have $\gamma = 2/(\nu - 4)$.

2.3. *Rayleigh optimization.* We assume that the two classes both follow elliptical distributions: $\mathbf{X}|(Y=0) \sim \mathcal{E}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, g_1)$ and $\mathbf{X}|(Y=1) \sim \mathcal{E}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, g_2)$. To facilitate the presentation, we assume the quantity γ is the same for both classes of conditional distributions. Let

$$\begin{aligned} M(\boldsymbol{\Omega}, \boldsymbol{\delta}) &= -\boldsymbol{\mu}_1^\top\boldsymbol{\Omega}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^\top\boldsymbol{\Omega}\boldsymbol{\mu}_2 + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\delta} - \operatorname{tr}\big(\boldsymbol{\Omega}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\big), \\ L_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) &= 2(1+\gamma)\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}\boldsymbol{\Sigma}_k) + \gamma\big[\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k)\big]^2 \\ &\quad + 4(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta})^\top\boldsymbol{\Sigma}_k(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta}), \end{aligned} \tag{5}$$

for $k = 1$ and 2. Combining (3) with Proposition 2.1, we have

$$R(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \frac{[M(\boldsymbol{\Omega}, \boldsymbol{\delta})]^2}{L_1(\boldsymbol{\Omega}, \boldsymbol{\delta}) + \kappa L_2(\boldsymbol{\Omega}, \boldsymbol{\delta})}, \tag{6}$$

where $\kappa = (1-\pi)/\pi$.

Note that if we multiply both $\boldsymbol{\Omega}$ and $\boldsymbol{\delta}$ by a common constant, $R(\boldsymbol{\Omega}, \boldsymbol{\delta})$ remains unchanged. Therefore, maximizing $R(\boldsymbol{\Omega}, \boldsymbol{\delta})$ is equivalent to solving the following constrained minimization problem:

$$\min_{(\boldsymbol{\Omega}, \boldsymbol{\delta}) : M(\boldsymbol{\Omega}, \boldsymbol{\delta}) = 1, \boldsymbol{\Omega} = \boldsymbol{\Omega}^\top} \big\{L_1(\boldsymbol{\Omega}, \boldsymbol{\delta}) + \kappa L_2(\boldsymbol{\Omega}, \boldsymbol{\delta})\big\}. \tag{7}$$

We call problem (7) the *Rayleigh optimization*. It is a convex problem whenever $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are both positive semi-definite.

The formulation of the Rayleigh optimization only involves the means and covariance matrices, and the kurtosis parameter γ. Therefore, if we know γ (e.g., when we know which subfamily the distributions belong to) and have good estimates $(\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\mathbf{\Sigma}}_1, \widehat{\mathbf{\Sigma}}_2)$, we can solve the empirical version of (7) to obtain $(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}})$, which is the main idea of QUADRO. In addition, (7) is a convex problem, with a quadratic objective and equality constraints. Hence it can be solved efficiently by many optimization algorithms.

3. Quadratic dimension reduction via Rayleigh optimization. Now, we formally introduce the QUADRO procedure. We fix a model parameter $\gamma \geq 0$. Let $\widehat{M}$, $\widehat{L}_1$ and $\widehat{L}_2$ be the sample versions of M, L_1, L_2 in (5) by replacing $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ with their estimates. Details of these estimates will be given in Section 4. Let $\widehat{\pi} = n_1/(n_1 + n_2)$ and $\kappa = \widehat{\pi}/(1 - \widehat{\pi})$. Given tuning parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, we solve

$$\min_{(\mathbf{\Omega},\boldsymbol{\delta})\,:\,\widehat{M}(\mathbf{\Omega},\boldsymbol{\delta})=1,\mathbf{\Omega}=\mathbf{\Omega}^\top} \big\{\widehat{L}_1(\mathbf{\Omega},\boldsymbol{\delta}) + \kappa \widehat{L}_2(\mathbf{\Omega},\boldsymbol{\delta}) + \lambda_1|\mathbf{\Omega}|_1 + \lambda_2|\boldsymbol{\delta}|_1\big\}. \tag{8}$$

We propose a linearized augmented Lagrangian method to solve (8). To simplify the notation, we write $\widehat{L} = \widehat{L}_1 + \kappa \widehat{L}_2$, and omit the hat symbol on M and L when there is no confusion. The optimization problem is then

$$\min_{(\mathbf{\Omega},\boldsymbol{\delta})\,:\,M(\mathbf{\Omega},\boldsymbol{\delta})=1,\mathbf{\Omega}=\mathbf{\Omega}^\top} \big\{L(\mathbf{\Omega},\boldsymbol{\delta}) + \lambda_1|\mathbf{\Omega}|_1 + \lambda_2|\boldsymbol{\delta}|_1\big\}.$$

For an algorithm parameter $\rho > 0$, and a dual variable ν, we define the *augmented Lagrangian* as

$$F_\rho(\mathbf{\Omega},\boldsymbol{\delta},\nu) = L(\mathbf{\Omega},\boldsymbol{\delta}) + \nu\big[M(\mathbf{\Omega},\boldsymbol{\delta}) - 1\big] + (\rho/2)\big[M(\mathbf{\Omega},\boldsymbol{\delta}) - 1\big]^2.$$

Using zero as the initial value, we iteratively update:

- $\boldsymbol{\delta}^{(k)} = \operatorname{argmin}_{\boldsymbol{\delta}}\{F_\rho(\mathbf{\Omega}^{(k-1)}, \boldsymbol{\delta}, \nu^{(k-1)}) + \lambda_2|\boldsymbol{\delta}|_1\}$,
- $\mathbf{\Omega}^{(k)} = \operatorname{argmin}_{\mathbf{\Omega}\,:\,\mathbf{\Omega}=\mathbf{\Omega}^\top}\{F_\rho(\mathbf{\Omega}, \boldsymbol{\delta}^{(k)}, \nu^{(k-1)}) + \lambda_1|\mathbf{\Omega}|_1\}$,
- $\nu^{(k)} = \nu^{(k-1)} + \rho[M(\mathbf{\Omega}^{(k)}, \boldsymbol{\delta}^{(k)}) - 1]$.

Here, the first two steps are *primal updates*, and the third step is a *dual update*.

First, we consider the update of $\boldsymbol{\delta}$. When $\mathbf{\Omega}$ and ν are fixed, we can write

$$F_\rho(\mathbf{\Omega},\boldsymbol{\delta},\nu) = \boldsymbol{\delta}^\top \mathbf{A}\boldsymbol{\delta} - 2\boldsymbol{\delta}^\top \mathbf{b} + c_\rho(\mathbf{\Omega},\nu),$$

where

$$\begin{aligned}\mathbf{A} &= 4(\mathbf{\Sigma}_1 + \kappa\mathbf{\Sigma}_2) + 2\rho(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top,\\ \mathbf{b} &= 4(\mathbf{\Sigma}_1\mathbf{\Omega}\boldsymbol{\mu}_1 + \kappa\mathbf{\Sigma}_2\mathbf{\Omega}\boldsymbol{\mu}_2)\\ &\quad + \big[\rho\,\mathrm{tr}\big(\mathbf{\Omega}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\big) + \rho\boldsymbol{\mu}_1^\top\mathbf{\Omega}\boldsymbol{\mu}_1 - \rho\boldsymbol{\mu}_2^\top\mathbf{\Omega}\boldsymbol{\mu}_2 + (\rho - \nu)\big](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),\end{aligned} \tag{9}$$

and $c_\rho(\mathbf{\Omega}, \nu)$ does not depend on $\boldsymbol{\delta}$. Note that $\mathbf{A}$ is a positive semi-definite matrix. The update of $\boldsymbol{\delta}$ is indeed a Lasso problem.

Next, we consider the update of $\mathbf{\Omega}$. When $\boldsymbol{\delta}$ and ν are fixed, $F_\rho(\mathbf{\Omega}, \boldsymbol{\delta}, \nu)$ is a convex function of $\mathbf{\Omega}$. We propose an approximate update step: we first "linearize" F_ρ at $\mathbf{\Omega} = \mathbf{\Omega}^{(k-1)}$ to construct an upper envelope $\bar{F}_\rho$, and then minimize this upper envelope. In detail, at any $\mathbf{\Omega} = \mathbf{\Omega}_0$, we consider the following upper bound of $F_\rho(\mathbf{\Omega}, \boldsymbol{\delta}, \nu)$:

$$\bar{F}_\rho(\mathbf{\Omega}, \boldsymbol{\delta}, \nu) \equiv F_\rho(\mathbf{\Omega}_0, \boldsymbol{\delta}, \nu) + \sum_{1 \le i \le j \le d} \big[\Omega(i,j) - \Omega_0(i,j)\big] \frac{\partial F_\rho(\mathbf{\Omega}_0, \boldsymbol{\delta}, \nu)}{\partial \Omega(i,j)} + \frac{\tau}{2} \sum_{1 \le i \le j \le d} \big[\Omega(i,j) - \Omega_0(i,j)\big]^2,$$

where τ is a large enough constant [e.g., we can take $\tau = \sum_{1 \le i \le j \le d} \frac{\partial^2 F_\rho(\mathbf{\Omega}_0, \boldsymbol{\delta}, \nu)}{\partial \Omega(i,j)^2}$]. We then minimize $\bar{F}_\rho(\mathbf{\Omega}, \boldsymbol{\delta}, \nu) + \lambda_1 |\mathbf{\Omega}|_1$ to update $\mathbf{\Omega}$. This modified update step has an explicit solution,

$$\Omega^*(i,j) = \mathcal{S}\bigg(\Omega_0(i,j) - \frac{1}{\tau} \frac{\partial F_\rho(\mathbf{\Omega}_0, \boldsymbol{\delta}, \nu)}{\partial \Omega(i,j)}, \frac{\lambda_1}{\tau}\bigg),$$

where $\mathcal{S}(x, a) \equiv (|x| - a)_+ \operatorname{sign}(x)$ is the soft-thresholding function. We can write $\mathbf{\Omega}^*$ in a matrix form. Let

$$\begin{aligned} \mathbf{D} = {} & 4(1+\gamma)(\mathbf{\Sigma}_1 \mathbf{\Omega} \mathbf{\Sigma}_1 + \kappa \mathbf{\Sigma}_2 \mathbf{\Omega} \mathbf{\Sigma}_2) + 2\gamma\big[\operatorname{tr}(\mathbf{\Omega}\mathbf{\Sigma}_1)\mathbf{\Sigma}_1 + \kappa \operatorname{tr}(\mathbf{\Omega}\mathbf{\Sigma}_2)\mathbf{\Sigma}_2\big] \\ & + 4 \operatorname{sym}\big(\mathbf{\Sigma}_1(\mathbf{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})\boldsymbol{\mu}_1^\top + \kappa \mathbf{\Sigma}_2(\mathbf{\Omega}\boldsymbol{\mu}_2 - \boldsymbol{\delta})\boldsymbol{\mu}_2^\top\big), \end{aligned} \tag{10}$$

where $\operatorname{sym}(\mathbf{B}) = (\mathbf{B} + \mathbf{B}^\top)/2$ for any square matrix $\mathbf{B}$. By direct calculation,

$$\mathbf{\Omega}^* = \mathcal{S}\bigg(\mathbf{\Omega}_0 - \frac{1}{\tau}\mathbf{D}, \frac{\lambda_1}{\tau}\bigg).$$

We now describe our algorithm. Let us initialize $\mathbf{\Omega}^{(0)} = \mathbf{0}_{d \times d}$, $\boldsymbol{\delta}^{(0)} = \mathbf{0}$ and $\nu^{(0)} = 0$. At iteration k, the algorithm updates as follows:

- Compute $\mathbf{A} = \mathbf{A}(\mathbf{\Omega}^{(k-1)}, \boldsymbol{\delta}^{(k-1)}, \nu^{(k-1)})$ and $\mathbf{b} = \mathbf{b}(\mathbf{\Omega}^{(k-1)}, \boldsymbol{\delta}^{(k-1)}, \nu^{(k-1)})$ using (9). Update $\boldsymbol{\delta}^{(k)} = \operatorname{argmin}_{\boldsymbol{\delta}}\{\boldsymbol{\delta}^\top \mathbf{A} \boldsymbol{\delta} - 2\boldsymbol{\delta}^\top \mathbf{b} + \lambda_2 |\boldsymbol{\delta}|_1\}$.
- Compute $\mathbf{D} = \mathbf{D}(\mathbf{\Omega}^{(k-1)}, \boldsymbol{\delta}^{(k)}, \nu^{(k-1)})$ using (10). Update $\mathbf{\Omega}^{(k)} = \mathcal{S}(\mathbf{\Omega}^{(k-1)} - \frac{1}{\tau}\mathbf{D}, \frac{\lambda_1}{\tau})$.
- Update $\nu^{(k)} = \nu^{(k-1)} + \rho[M(\mathbf{\Omega}^{(k)}, \boldsymbol{\delta}^{(k)}) - 1]$.

Stop until $\max\{\rho|\mathbf{\Omega}^{(k)} - \mathbf{\Omega}^{(k-1)}|, \rho|\boldsymbol{\delta}^{(k)} - \boldsymbol{\delta}^{(k-1)}|, |\nu^{(k)} - \nu^{(k-1)}|/\rho\} \le \varepsilon$ for some pre-specified precision ε.

This is a modified version of the augmented Lagrangian method, where in the step of updating $\mathbf{\Omega}$, we minimize an upper envelope, which is obtained by locally linearizing the augmented Lagrangian.

REMARK. QUADRO can be extended to folded concave penalties, for example, to SCAD [Fan and Li (2001)] or to adaptive Lasso [Zou (2006)]. Using the Local Linear Approximation algorithm [Fan, Xue and Zou (2014), Zou and Li (2008)], we can solve the SCAD-penalized QUADRO and the adaptive-Lasso-penalized QUADRO by solving L_1-penalized QUADRO with multiple-step and one-step iterations, respectively.

4. Estimation of mean and covariance matrix. QUADRO requires estimates of the mean vector and covariance matrix for each class as inputs. We will show in Section 5 that the performance of QUADRO is closely related to the max-norm estimation error on mean vectors and covariance matrices. Sample mean and sample covariance matrix work well for Gaussian data. However, when data are from elliptical distributions, they may have inferior performance as we estimate nonpolynomially many of means and variances. In Sections 4.1–4.2, we suggest a robust M-estimator to estimate the mean and a rank-based estimator to estimate the covariance matrix, which are more appropriate for non-Gaussian data. Moreover, in Section 4.3 we discuss how to estimate the model parameter γ when it is unknown.

4.1. *Estimation of the mean.* Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. samples of a random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ from an elliptical distribution $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. Let us denote $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top$ for $i = 1, \ldots, n$. We estimate each μ_j marginally using the data $\{x_{1j}, \ldots, x_{nj}\}$.

One possible estimator is the sample median

$$\widehat{\mu}_{Mj} = \operatorname{median}(\{x_{1j}, \ldots, x_{nj}\}).$$

It can be shown that even under heavy-tailed distributions, $P(|\widehat{\mu}_{Mj} - \mu_j| > A\sqrt{\log(\delta^{-1})/n}) \le \delta$ for small $\delta \in (0, 1)$, where A is a constant determined by the probability density at μ_j, for each fixed j. This combined with the union bound gives that $|\widehat{\boldsymbol{\mu}}_M - \boldsymbol{\mu}|_\infty = O_p(\sqrt{\log(d)/n})$.

Catoni (2012) proposed another M-estimator for the mean of heavy-tailed distributions. It works for distributions where mean is not necessarily equal to median, which is essential for estimating covariance of random variables. We denote the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ as $\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2$, and the off-diagonal elements as σ_{kj} for $k \neq j$. The estimator $\widehat{\boldsymbol{\mu}}_C = (\widehat{\mu}_{C,1}, \ldots, \widehat{\mu}_{C,d})^\top$ is obtained as follows. For a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$ such that $-\log(1 - y + y^2/2) \le h(y) \le \log(1 + y + y^2/2)$, and a value $\delta \in (0, 1)$ such that $n > 2\log(1/\delta)$, we let

$$\alpha_\delta = \left\{ \frac{2\log(\delta^{-1})}{n[v + (2v\log(\delta^{-1}))/(n - 2\log(\delta^{-1}))]} \right\}^{1/2},$$

where v is an upper bound of $\max\{\sigma_1^2, \ldots, \sigma_d^2\}$. For each j, we define $\widehat{\mu}_{Cj}$ as the unique value that satisfies $\sum_{i=1}^n h(\alpha_\delta(x_{ij} - \widehat{\mu}_{Cj})) = 0$. It was shown in Catoni

(2012) that $P(|\widehat{\mu}_{Cj} - \mu_j| > \sqrt{\frac{2v\log(\delta^{-1})}{n(1-2\log(\delta^{-1})/n)}}) \le \delta$ when the variance of X_j exists. Therefore, by taking $\delta = 1/(n \vee d)^2$, $|\widehat{\boldsymbol{\mu}}_M - \boldsymbol{\mu}|_\infty \le C\sqrt{\log(d)/n}$ with probability at least $1 - (n \vee d)^{-1}$, which gives the desired convergence rate.

To implement this estimator, we take $h(y) = \operatorname{sgn}(y)\log(1 + |y| + y^2/2)$. For the choice of v, any value larger than $\max\{\sigma_1^2, \ldots, \sigma_d^2\}$ would work in theory. Catoni (2012) introduced a Lepski's adaptation method to choose v. For simplicity, we take $v = 3\max\{\widetilde{\sigma}_1^2, \ldots, \widetilde{\sigma}_d^2\}$, where $\widetilde{\sigma}_j^2$ is the sample covariance of X_j.

The two estimators, the median and the M-estimator, both have a convergence rate of $O_p(\sqrt{\log(d)/n})$ in terms of the max-norm error. In our numerical experiments, the M-estimator has a better numerical performance, and we stick to this estimator.

4.2. *Estimation of the covariance matrix.* To estimate the covariance matrix $\boldsymbol{\Sigma}$, we estimate the marginal covariances $\{\sigma_j^2, 1 \le j \le d\}$ and the correlation matrix $\mathbf{C}$ separately. Again, we need robust estimates even though the data have fourth moments, as we simultaneously estimate nonpolynomial number of covariance parameters.

First, we consider estimating σ_j^2. Note that $\sigma_j^2 = \mathbb{E}(X_j^2) - \mathbb{E}^2(X_j)$. We estimate $\mathbb{E}(X_j^2)$ and $\mathbb{E}(X_j)$ separately. To estimate $\mathbb{E}(X_j^2)$, we use the M-estimator described above on the squared data $\{x_{1j}^2, \ldots, x_{nj}^2\}$ and denote the estimator by $\widehat{\eta}_{Cj}$. This works as $\mathbb{E}(X_j^4)$ is finite for each j in our setting; in addition, the M-estimator applies to asymmetric distributions. We then define

$$\widehat{\sigma}_{Cj}^2 = \max\{\widehat{\eta}_{Cj} - \widehat{\mu}_{Cj}^2, \delta_0\},$$

where $\widehat{\mu}_{Cj}$ is the M-estimator of $\mathbb{E}(X_j)$ and $\delta_0 > 0$ is a small constant ($\delta_0 < \min\{\sigma_1^2, \ldots, \sigma_d^2\}$). It is easy to see that when the fourth moments of X_j are uniformly upper bounded by a constant and $n \ge 4\log(d^2)$, $\max\{|\widehat{\sigma}_{Cj} - \sigma_j|, 1 \le j \le d\} = O_p(\sqrt{\log(d)/n})$.

Next, we consider estimating the correlation matrix $\mathbf{C}$. For this, we use Kendall's tau correlation matrix proposed by Han and Liu (2012). Kendall's tau correlation coefficients [Kendall (1938)] are defined as

$$\tau_{jk} = \mathbb{P}\big((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0\big) - \mathbb{P}\big((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0\big),$$

where $\widetilde{\mathbf{X}}$ is an independent copy of $\mathbf{X}$. They have the following relationship to the true coefficients: $C_{jk} = \sin(\frac{\pi}{2}\tau_{jk})$ for the elliptical family. Based on this equality, we first estimate Kendall's tau correlation coefficients using rank-based estimators

$$\widehat{\tau}_{jk} = \begin{cases} \dfrac{2}{n(n-1)} \displaystyle\sum_{1 \le i < i' \le n} \operatorname{sign}\big((x_{ij} - x_{i'j})(x_{ik} - x_{i'k})\big), & j \ne k, \\ 1, & j = k, \end{cases}$$

and then estimate the correlation matrix by $\widehat{\mathbf{C}} = (\widehat{C}_{jk})$ with

$$\widehat{C}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right).$$

It is shown in Han and Liu (2012) that $|\widehat{\mathbf{C}} - \mathbf{C}|_\infty = O_p(\sqrt{\log(d)/n})$.

Finally, we combine $\{\widehat{\sigma}_j^2, 1 \le j \le d\}$ and $\widehat{\mathbf{C}}$ to get $\widehat{\mathbf{\Sigma}}$. Let

$$\widetilde{\Sigma}_{jk} = \widehat{\sigma}_j \widehat{\sigma}_k \widehat{C}_{jk}, \qquad 1 \le j, k \le d.$$

It follows immediately that $|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty = O_p(\sqrt{\log(d)/n})$. However, this estimator is not necessarily positive semi-definite. To implement QUADRO, we need $\widehat{\mathbf{\Sigma}}$ to be positive semi-definite so that the optimization in (8) is a convex problem. We obtain $\widehat{\mathbf{\Sigma}}$ by projecting $\widetilde{\mathbf{\Sigma}}$ onto the cone of positive semi-definite matrices through the convex optimization

$$\widehat{\mathbf{\Sigma}} = \underset{\mathbf{A}:\,\mathbf{A} \text{ is positive semidefinite}}{\operatorname{argmin}} \{|\mathbf{A} - \widetilde{\mathbf{\Sigma}}|_\infty\}. \tag{11}$$

Note that $|\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}|_\infty \le |\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}|_\infty$ by definition. Therefore, $|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \le |\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}|_\infty + |\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \le 2|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty = O_p(\sqrt{\log(d)/n})$. To compute $\widehat{\mathbf{\Sigma}}$, we note that the optimization problem in (11) can be formulated as the dual of a graphical lasso problem corresponding to the smallest possible tuning parameter that still guarantees a feasible solution [Liu et al. (2012)]. Zhao, Roeder and Liu (2013) provide more algorithmic details.

4.3. *Estimation of kurtosis parameter.* When the kurtosis parameter γ is unknown, we can estimate it from data. Recall that $\gamma = \frac{1}{d(d+2)}\mathbb{E}(\xi^4) - 1$. Using decomposition (4) and the properties of $\mathbf{U}$, we have

$$\mathbb{E}(\xi^4) = \mathbb{E}\{[(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})]^2\}.$$

Motivated by this equality, we propose the estimator

$$\widehat{\gamma} = \max\left\{\frac{1}{d(d+2)}\frac{1}{n}\sum_{i=1}^n [(\mathbf{x}_i - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\mathbf{\Omega}}(\mathbf{x}_i - \widetilde{\boldsymbol{\mu}})]^2 - 1, 0\right\},$$

where $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\mathbf{\Omega}}$ are estimators of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}^{-1}$, respectively. Maruyama and Seo (2003) considered a similar estimator in low-dimensional settings, where they used the sample mean and sample covariance matrix. In high dimensions, we a robust estimate to guarantee uniform convergence. In particular, we take $\widetilde{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_C$ and $\widetilde{\mathbf{\Omega}} = \widehat{\mathbf{\Omega}}_{\text{clime}}$ where $\widehat{\mathbf{\Omega}}_{\text{clime}}$ is the CLIME estimator proposed in Cai, Liu and Luo (2011). We can also take the covariance estimator in Section 4.2, but we will then need to establish its sampling property as a precision matrix estimator. We decide

to use the CLIME estimator since such a property has already been established by Cai, Liu and Luo (2011). Denote by $\mathbf{\Sigma}^{-1} = (\Omega_{jk})_{d\times d}$. From simple algebra,

$$|\widehat{\gamma} - \gamma| \le \max_{1\le j,k\le d} |\widetilde{\mu}_j \widetilde{\Omega}_{jk} \widetilde{\mu}_k - \mu_j \Omega_{jk} \mu_k|$$

$$\le C \max\{|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}|_\infty, |\widetilde{\mathbf{\Omega}} - \mathbf{\Sigma}^{-1}|_\infty\}.$$

In Section 4.1, we have seen that $\|\widehat{\boldsymbol{\mu}}_C - \boldsymbol{\mu}\|_\infty = O_p(\sqrt{\log(d)/n})$. Moreover, Cai, Liu and Luo (2011) showed that $|\widetilde{\mathbf{\Omega}} - \mathbf{\Sigma}^{-1}|_\infty = \|\mathbf{\Sigma}^{-1}\|_1 \cdot O_p(\sqrt{\log(d)/n})$ under mild conditions, where $\|\cdot\|_1$ is the matrix L_1-norm. Therefore, provided that $\|\mathbf{\Sigma}^{-1}\|_1 \le C$, we immediately have $|\widehat{\gamma} - \gamma| = O_p(\sqrt{\log(d)/n})$.

5. Theoretical properties. In this section, we establish an oracle inequality for the Rayleigh quotient of the QUADRO estimates $(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}})$. We assume that π and γ are known. For notational simplicity, we set $\lambda_1 = \lambda_2 = \lambda$. The results can be easily generalized to the case $\lambda_1 \ne \lambda_2$. Moreover, we drop the symmetry constraint $\mathbf{\Omega} = \mathbf{\Omega}^\top$ in all optimization problems involved. This simplifies the expression of the regularity conditions. The analysis with the symmetry constraint is a trivial extension of current analysis.

Recall the definition of M, L_1 and L_2 in (5) and $\kappa = (1-\pi)/\pi$ and $L = L_1 + \kappa L_2$, the Rayleigh quotient of $(\mathbf{\Omega}, \boldsymbol{\delta})$ is equal to (up to a multiplicative constant)

$$R(\mathbf{\Omega}, \boldsymbol{\delta}) = \frac{[M(\mathbf{\Omega}, \boldsymbol{\delta})]^2}{L(\mathbf{\Omega}, \boldsymbol{\delta})}.$$

The QUADRO estimates are

$$(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}}) = \operatorname*{argmin}_{(\mathbf{\Omega},\boldsymbol{\delta}):\, \widehat{M}(\mathbf{\Omega},\boldsymbol{\delta})=1} \{\widehat{L}(\mathbf{\Omega}, \boldsymbol{\delta}) + \lambda|\mathbf{\Omega}|_1 + \lambda|\boldsymbol{\delta}|_1\}.$$

We shall compare the Rayleigh quotient of $(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}})$ with the Rayleigh quotients of a class of "oracle solutions." This class includes the one that maximizes the true Rayleigh quotient, which we denote by $(\mathbf{\Omega}_0^*, \boldsymbol{\delta}_0^*)$. Here we adopt a class of solutions as the "oracle" instead of only $(\mathbf{\Omega}_0^*, \boldsymbol{\delta}_0^*)$, because we want the results not tied to the sparsity assumption on $(\mathbf{\Omega}_0^*, \boldsymbol{\delta}_0^*)$ but a weaker assumption: at least one solution in this class is sparse.

Our theoretical development is technically nontrivial. Conventional oracle inequalities are derived in a setting of minimizing a data-dependent loss without constraint, and the risk function is the expectation of the loss. Here we minimize a data-dependent loss with a data-dependent equality constraint, and the risk function—the Rayleigh quotient—is not equal to the expectation of the loss. A similar setting was considered in Fan, Feng and Tong (2012), where they introduced a data-dependent intermediate solution to deal with such equality constraint. However, the rate they obtained depends on this intermediate solution, which is very hard to quantify. In contrast, the rate in our results purely depends on the oracle solution. To get rid of the intermediate solution in the rate, we need to carefully

quantify its difference from both the QUADRO solution and the oracle solution. The technique is new, and potentially useful for other problems.

5.1. *Oracle solutions, the restricted eigenvalue condition.* For any $\lambda_0 \geq 0$, we define the *oracle solution associated with* λ_0 to be

$$(\boldsymbol{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0}) = \underset{(\boldsymbol{\Omega},\boldsymbol{\delta})\,:\, M(\boldsymbol{\Omega},\boldsymbol{\delta})=1}{\operatorname{argmin}} \big\{L(\boldsymbol{\Omega}, \boldsymbol{\delta}) + \lambda_0|\boldsymbol{\Omega}|_1 + \lambda_0|\boldsymbol{\delta}|_1\big\}. \tag{12}$$

We shall compare the Rayleigh quotient of $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ to that of $(\boldsymbol{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})$, for an arbitrary λ_0. In particular, when $\lambda_0 = 0$, the associated oracle solution (may not be unique) becomes

$$(\boldsymbol{\Omega}^*_0, \boldsymbol{\delta}^*_0) = \underset{(\boldsymbol{\Omega},\boldsymbol{\delta})\,:\, M(\boldsymbol{\Omega},\boldsymbol{\delta})=1}{\operatorname{argmin}} \big\{L(\boldsymbol{\Omega}, \boldsymbol{\delta})\big\}.$$

It maximizes the true Rayleigh quotient.

Next, we introduce a restricted eigenvalue (RE) condition jointly on $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. For any matrices $\mathbf{A}$ and $\mathbf{B}$, let $\operatorname{vec}(\mathbf{A})$ be the vectorization of $\mathbf{A}$ by stacking all the elements of $\mathbf{A}$ column by column, and $\mathbf{A} \otimes \mathbf{B}$ be the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$. We define the matrices

$$\mathbf{Q}_k = \begin{bmatrix} (2(1+\gamma)\boldsymbol{\Sigma}_k + 4\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top) \otimes \boldsymbol{\Sigma}_k + \gamma \operatorname{vec}(\boldsymbol{\Sigma}_k)\operatorname{vec}(\boldsymbol{\Sigma}_k)^\top & -4\boldsymbol{\mu}_k \otimes \boldsymbol{\Sigma}_k \\ -4\boldsymbol{\mu}_k^\top \otimes \boldsymbol{\Sigma}_k & 4\boldsymbol{\Sigma}_k \end{bmatrix},$$

for $k = 1, 2$. We note that there are $(d^2 + d)$ coefficients to decide when maximizing $R(\boldsymbol{\Omega}, \boldsymbol{\delta})$: d^2 elements of $\boldsymbol{\Omega}$ and d elements of $\boldsymbol{\delta}$. We can stack all these coefficients into a long vector $\mathbf{x} = \mathbf{x}(\boldsymbol{\Omega}, \boldsymbol{\delta})$ in $\mathbb{R}^{d^2+d}$ defined as

$$\mathbf{x}(\boldsymbol{\Omega}, \boldsymbol{\delta}) \equiv \big[\operatorname{vec}(\boldsymbol{\Omega})^\top, \boldsymbol{\delta}^\top\big]^\top. \tag{13}$$

It can be shown that $L_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \mathbf{x}^\top \mathbf{Q}_k \mathbf{x}$, for $k = 1, 2$; see Lemma 9.1. Therefore, $L(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \mathbf{x}^\top \mathbf{Q}\mathbf{x}$, where $\mathbf{Q} = \mathbf{Q}_1 + \kappa\mathbf{Q}_2$. Our RE condition is then imposed on the $(d^2 + d) \times (d^2 + d)$ matrix $\mathbf{Q}$, and hence implicitly on $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$.

We now formally introduce the RE condition. For a set $S \subset \{1, 2, \ldots, d^2 + d\}$ and a nonnegative value $\bar{c}$, we define the *restricted eigenvalue* in the following way:

$$\Theta(S; \bar{c}) = \min_{\mathbf{v}\,:\, |\mathbf{v}_{S^c}|_1 \leq \bar{c}|\mathbf{v}_S|_1} \frac{\mathbf{v}^\top \mathbf{Q}\mathbf{v}}{|\mathbf{v}_S|^2}.$$

Generally speaking, $\Theta(S; \bar{c})$ depends on $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ in a complicated way. For $\bar{c} = 0$, the following proposition builds a connection between $\Theta(S; 0)$ and $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. For each $S \subset \{1, 2, \ldots, d^2 + d\}$, there exist sets $U \subset \{1, \ldots, d\} \times \{1, \ldots, d\}$ and $V \subset \{1, \ldots, d\}$ such that the support of $\mathbf{x}(\boldsymbol{\Omega}, \boldsymbol{\delta})$ is S if and only if the support of $\boldsymbol{\Omega}$ is U and the support of $\boldsymbol{\delta}$ is V. Let

$$U' = \bigcup_{(i,j)\in U} \{i, j\}.$$

Then $U \subset U' \times U'$. The following result is proved in Fan et al. (2014).

PROPOSITION 5.1. *For any set $S \subset \{1, \ldots, d^2 + d\}$, suppose U' and V are defined as above. Let $\widetilde{\mathbf{\Sigma}}_k$ be the submatrix of $\mathbf{\Sigma}_k$ by restricting rows and columns to $U' \cup V$, $\widetilde{\boldsymbol{\mu}}_k$ be the subvector of $\boldsymbol{\mu}_k$ by constraining elements to $U' \cup V$, for $k = 1, 2$. If there exist constants $v_1, v_2 > 0$ such that $\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_k - v_1 \widetilde{\boldsymbol{\mu}}_k \widetilde{\boldsymbol{\mu}}_k^\top) \geq \frac{1}{2}\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_k) \geq \frac{v_2}{2}$ for $k = 1, 2$, then*

$$\Theta(S, 0) \geq (1 + \gamma)(1 + \kappa) v_2 \min\left\{ v_2, \frac{4 v_1}{2 + v_1(1 + \gamma)} \right\} > 0.$$

5.2. *Oracle inequality on Rayleigh's quotient.* Suppose $\max\{|\mathbf{\Sigma}_k|_\infty, |\boldsymbol{\mu}_k|_\infty, k = 1, 2\} \leq 1$ and $|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty \leq |\mathbf{\Sigma}_k|_\infty$, $|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty \leq |\boldsymbol{\mu}_k|_\infty$ for $k = 1, 2$, without loss of generality. For any $\lambda_0 \geq 0$, let $(\mathbf{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})$ be the associated oracle solution and S be the support of $\mathbf{x}^*_{\lambda_0} = [\mathrm{vec}(\mathbf{\Omega}^*_{\lambda_0})^\top, (\boldsymbol{\delta}^*_{\lambda_0})^\top]^\top$. Let $\Delta_n = \max\{|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty, |\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty, k = 1, 2\}$. We have the following result for any given estimators, the proof of which we postpone to Section 9.

THEOREM 5.1. *Given $\lambda_0 \geq 0$, let S be the support of $\mathbf{x}^*_{\lambda_0}$, $s_0 = |S|$ and $k_0 = \max\{s_0, R(\mathbf{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})\}$. Suppose that $\Theta(S, 0) \geq c_0$, $\Theta(S, 3) \geq a_0$ and $R(\mathbf{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0}) \geq u_0$, for some positive constants a_0, c_0 and u_0. We assume $4 s_0 \Delta_n^2 \leq a_0 c_0$ and $\max\{s_0 \Delta_n, s_0^{1/2} k_0^{1/2} \lambda_0\} < 1$ without loss of generality. Then there exist positive constants $C = C(a_0, c_0, u_0)$ and $A = A(a_0, c_0, u_0)$ such that for any $\eta > 1$,*

$$\frac{R(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}})}{R(\mathbf{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})} \geq 1 - A\eta^2 \max\{s_0 \Delta_n, s_0^{1/2} k_0^{1/2} \lambda_0\},$$

by taking $\lambda = C\eta \max\{s_0^{1/2} \Delta_n, k_0^{1/2} \lambda_0\} [R(\mathbf{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})]^{-1/2}$.

In Theorem 5.1, the rate of convergence has two parts. The term $s_0 \Delta_n$ reflects how the stochastic errors of estimating $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ affect the Rayleigh quotient. The term $s_0^{1/2} k_0^{1/2} \lambda_0$ is an extra term that depends on the oracle solution we aim to use for comparison. In particular, if we compare $R(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}})$ with $R_{\max} \equiv R(\mathbf{\Omega}^*_0, \boldsymbol{\delta}^*_0)$, the population maximum Rayleigh quotient with $\lambda_0 = 0$, this extra term disappears. If we further use the estimators in Section 4, $\Delta_n = O_p(\sqrt{\log(d)/n})$. We summarize the result as follows.

COROLLARY 5.1. *Suppose that the condition of Theorem 5.1 holds with $\lambda_0 = 0$. Then for some positive constants A and C, when $\lambda > C s_0^{1/2} R_{\max}^{-1/2} \Delta_n$, we have*

$$R(\widehat{\mathbf{\Omega}}, \widehat{\boldsymbol{\delta}}) \geq (1 - A s_0 \Delta_n) R_{\max}.$$

Furthermore, if the mean vectors and covariance matrices are estimated by using the robust methods in Section 4, then when $\lambda > Cs_0^{1/2} R_{\max}^{-1/2} \sqrt{\log(d)/n}$,

$$R(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}}) \geq \bigl(1 - As_0 \sqrt{\log(d)/n}\bigr) R_{\max},$$

with probability at least $1 - (n \vee d)^{-1}$.

From Corollary 5.1, when $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$ is truly sparse, $R(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ is close to the population maximum Rayleigh quotient $R_{\max}$. However, we note that Theorem 5.1 considers more general situations, including cases where $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$ is not sparse. As long as there exists an "approximately optimal" and sparse solution, that is, for a small λ_0 the associated oracle solution $(\boldsymbol{\Omega}_{\lambda_0}^*, \boldsymbol{\delta}_{\lambda_0}^*)$ is sparse, Theorem 5.1 guarantees that $R(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ is close to $R(\boldsymbol{\Omega}_{\lambda_0}^*, \boldsymbol{\delta}_0^*)$ and hence close to $R_{\max}$.

REMARK. Our results are analogous to oracle inequalities for prediction error in linear regressions; therefore, the condition $\Theta(S, \bar{c})$ is similar to the RE condition in linear regressions [Bickel, Ritov and Tsybakov (2009)]. To recover the support of $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$, conditions similar to the "irrepresentable condition" for Lasso [Zhao and Yu (2006)] are needed.

6. Application to classification. One important application of QUADRO is high-dimensional classification for elliptically-distributed data. Suppose $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ are the QUADRO estimates. This yields the classification rule

$$\widehat{h}(\mathbf{x}) = I\bigl\{\mathbf{x}^\top \widehat{\boldsymbol{\Omega}} \mathbf{x} - 2\widehat{\boldsymbol{\delta}}^\top \mathbf{x} < c\bigr\}.$$

In this section, we first show that for normally distributed data, the Rayleigh quotient is a proxy of the classification error, and then derive an analytic choice of c. Comparing with many other high-dimensional classification methods, QUADRO produces quadratic boundaries and can handle both non-Gaussian distributions and nonequal covariance matrices.

6.1. *Approximation of classification errors.* Given $(\boldsymbol{\Omega}, \boldsymbol{\delta})$ and a threshold c, a general quadratic rule $h(\mathbf{x}) = h(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\delta}, c)$ is defined as

$$h(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\delta}, c) = I\bigl\{\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} - 2\mathbf{x}^\top \boldsymbol{\delta} < c\bigr\}. \tag{14}$$

We reparametrize c as

$$c = t M_1(\boldsymbol{\Omega}, \boldsymbol{\delta}) + (1 - t) M_2(\boldsymbol{\Omega}, \boldsymbol{\delta}). \tag{15}$$

Here $M_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \boldsymbol{\mu}_k^\top \boldsymbol{\Omega} \boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k^\top \boldsymbol{\delta} + \operatorname{tr}(\boldsymbol{\Omega} \boldsymbol{\Sigma}_k)$ is the mean of $Q(\mathbf{X})$ in class k, for $k = 1, 2$. After the reparametrization, t is *scale-free*. As we will see below, in most cases, given $\boldsymbol{\Omega}$ and $\boldsymbol{\delta}$, the optimal t that minimizes the classification error takes values on $(0, 1)$.

From now on, we write $h(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\delta}, c) = h(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\delta}, t)$. Let $\mathrm{Err}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t)$ be the classification error of $h(\cdot; \boldsymbol{\Omega}, \boldsymbol{\delta}, t)$. Due to technical difficulties, we only give results for Gaussian distributions. Suppose $\mathbf{X}|(Y=0) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{X}|(Y=1) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. For $k = 1, 2$, we write

$$\boldsymbol{\Sigma}_k^{1/2} \boldsymbol{\Omega} \boldsymbol{\Sigma}_k^{1/2} = \mathbf{K}_k \mathbf{S}_k \mathbf{K}_k^T,$$

where $\mathbf{S}_k$ is a diagonal matrix containing the nonzero eigenvalues, and the columns of $\mathbf{K}_k$ are corresponding eigenvectors. Let $\boldsymbol{\beta}_k = \mathbf{K}_k^T \boldsymbol{\Sigma}_k (\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta})$. When $\max\{|\mathbf{S}_k|_\infty, |\boldsymbol{\beta}_k|_\infty, k = 1, 2\}$ is bounded, the following proposition shows that an approximation of $\mathrm{Err}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t)$ is

$$\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) \equiv \pi \bar{\Phi}\biggl(\frac{(1-t)M(\boldsymbol{\Omega}, \boldsymbol{\delta})}{\sqrt{L_1(\boldsymbol{\Omega}, \boldsymbol{\delta})}}\biggr) + (1-\pi)\bar{\Phi}\biggl(\frac{tM(\boldsymbol{\Omega}, \boldsymbol{\delta})}{\sqrt{L_2(\boldsymbol{\Omega}, \boldsymbol{\delta})}}\biggr),$$

where M, L_1 and L_2 are defined in (5), Φ is the distribution function of a standard normal variable and $\bar{\Phi} = 1 - \Phi$. Its proof is contained in Section 9.

PROPOSITION 6.1. *Suppose that* $\max\{|\mathbf{S}_k|_\infty, |\boldsymbol{\beta}_k|_\infty, k = 1, 2\} \le C_0$ *for some constant* $C_0 > 0$, *and let* q *be the rank of* $\boldsymbol{\Omega}$. *Then as* d *goes to infinity*,

$$\bigl|\mathrm{Err}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) - \overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t)\bigr| = \frac{O(q) + o(d)}{[\min\{L_1(\boldsymbol{\Omega}, \boldsymbol{\delta}), L_2(\boldsymbol{\Omega}, \boldsymbol{\delta})\}]^{3/2}}.$$

In particular, if we consider all such $(\boldsymbol{\Omega}, \boldsymbol{\delta})$ that the variance of $Q(\mathbf{X}; \boldsymbol{\Omega}, \boldsymbol{\delta})$ under both classes are lower bounded by $c_0 d^\theta$ for some constants $\theta > 2/3$ and $c_0 > 0$, then we have $|\mathrm{Err} - \overline{\mathrm{Err}}| = o(1)$.

We now take a closer look at $\overline{\mathrm{Err}}$. Let $H(x) = \bar{\Phi}(1/\sqrt{x})$, which is monotone increasing on $(0, \infty)$. Writing for short $M = M_1 - M_2$, $M_k = M_k(\boldsymbol{\Omega}, \boldsymbol{\delta})$ and $L_k = L_k(\boldsymbol{\Omega}, \boldsymbol{\delta})$ for $k = 1, 2$, we have

$$\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) = \pi H\biggl(\frac{L_1}{(1-t)^2 M^2}\biggr) + (1-\pi) H\biggl(\frac{L_2}{t^2 M^2}\biggr).$$

Figure 2 shows that $H(\cdot)$ is nearly linear on an important range. This suggests the following approximation:

$$\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) \approx H\biggl(\pi \frac{L_1}{(1-t)^2 M^2} + (1-\pi)\frac{L_2}{t^2 M^2}\biggr) = H\biggl(\frac{\pi}{(1-t)^2}\frac{1}{R^{(t)}}\biggr), \tag{16}$$

where $R^{(t)} = R^{(t)}(\boldsymbol{\Omega}, \boldsymbol{\delta})$ is the $R(\boldsymbol{\Omega}, \boldsymbol{\delta})$ in (6) corresponding to the κ value

$$\kappa(t) \equiv \frac{1-\pi}{\pi}\frac{(1-t)^2}{t^2}.$$

The approximation in (16) is quantified in the following proposition, which is proved in Fan et al. (2014).

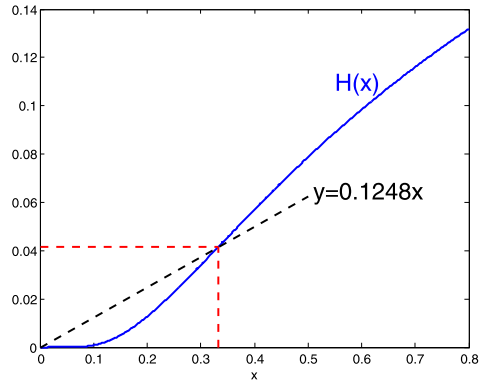


FIG. 2. *Function* $H(x) = \bar{\Phi}(1/\sqrt{x})$.

PROPOSITION 6.2. *Given* $(\boldsymbol{\Omega}, \boldsymbol{\delta}, t)$, *we write for short* $R_k = R_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) = [M(\boldsymbol{\Omega}, \boldsymbol{\delta})]^2 / L_k(\boldsymbol{\Omega}, \boldsymbol{\delta})$, *for* $k = 1, 2$, *and define*

$$V_1 = V_1(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) = \min\left\{(1-t)^2 R_1, \frac{1}{(1-t)^2 R_1}\right\},$$

$$V_2 = V_2(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) = \min\left\{t^2 R_2, \frac{1}{t^2 R_2}\right\},$$

$$V = V(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) = \max\{V_1/V_2, V_2/V_1\}.$$

Then there exists a constant $C > 0$ *such that*

$$\left|\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) - H\left(\frac{\pi}{(1-t)^2 R^{(t)}(\boldsymbol{\Omega}, \boldsymbol{\delta})}\right)\right| \le C\big[\max\{V_1, V_2\}\big]^{1/2} \cdot |V - 1|^2.$$

In particular, when $t = 1/2$,

$$\left|\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) - H\left(\frac{\pi}{(1-t)^2 R^{(t)}(\boldsymbol{\Omega}, \boldsymbol{\delta})}\right)\right| \le C R_0^{1/2} \cdot \left(\frac{\Delta R}{R_0}\right)^2,$$

where $R_0 = \max\{\min\{R_1, 1/R_1\}, \min\{R_2, 1/R_2\}\}$ *and* $\Delta R = |R_1 - R_2|$.

Note that L_1 and L_2 are the variances of $Q(\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X} - 2\mathbf{X}^\top \boldsymbol{\delta}$ for two classes, respectively. In cases where $|L_1 - L_2| \ll \min\{L_1, L_2\}$, $\Delta R \ll R_0$. Also, R_0 is always bounded by 1, and it tends to 0 in many situations, for example, when $R_1, R_2 \to \infty$, or $R_1, R_2 \to 0$, or $R_1 \to 0$, $R_2 \to \infty$. Proposition 6.2 then implies that the approximation in (16) when $t = 1/2$ is good.

Combining Propositions 6.1 and 6.2, the classification error of a general quadratic rule $h(\cdot; \boldsymbol{\Omega}, \boldsymbol{\delta}, t)$ is approximately a monotone decreasing transform of the Rayleigh quotient $R^{(t)}(\boldsymbol{\Omega}, \boldsymbol{\delta})$, corresponding to $\kappa = \kappa(t)$. In particular, when $t = 1/2$ [i.e., $c = (M_1 + M_2)/2$], $R^{(1/2)}(\boldsymbol{\Omega}, \boldsymbol{\delta})$ is exactly the one used in QUADRO. Consequently, if we fix the threshold to be $c = (M_1 + M_2)/2$, then the Rayleigh

quotient (upon with a monotone transform) is a good proxy for classification error. This explains why Rayleigh-quotient based procedures can be used for classification.

REMARK. Even in the region that $H(\cdot)$ is far from being linear such that the upper bound in Proposition 6.2 is not $o(1)$, we can still find a monotone transform of the Rayleigh quotient as an *upper bound* of the classification error. To see this, note that for $x \in [1/3, \infty)$, $H(x)$ is a concave function. Therefore, the approximation in (16) becomes an inequality, that is, $\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) \le H(\frac{\pi R^{(t)}}{(1-t)^2})$. For $x \in (0, 1/3)$, $H(x) \le 0.1248x$. It follows that $\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) \le 0.1248 \cdot \frac{\pi R^{(t)}}{(1-t)^2}$.

REMARK. In the current setting, the Bayes classifier is a quadratic rule $h(\mathbf{x}; \boldsymbol{\Omega}_B, \boldsymbol{\delta}_B, c_B)$ with $\boldsymbol{\Omega}_B = \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$, $\boldsymbol{\delta}_B = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$ and $c_B = \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1$. Let $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$ be the population solution of QUADRO when $\lambda = 0$. We note that $(\boldsymbol{\Omega}_B, \boldsymbol{\delta}_B)$ and $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$ are different: the former minimizes $\inf_t \mathrm{Err}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t)$, while the latter minimizes $\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, 1/2)$.

6.2. *QUADRO as a classification method.* Results in Section 6.1 suggest an analytic method to choose the threshold c, or equivalently t, with given $(\boldsymbol{\Omega}, \boldsymbol{\delta})$. Let

$$\widehat{t} \in \min_t \left\{ \pi \bar{\Phi}\left(\frac{(1-t)\widehat{M}(\boldsymbol{\Omega}, \boldsymbol{\delta})}{\sqrt{\widehat{L}_1(\boldsymbol{\Omega}, \boldsymbol{\delta})}} \right) + (1-\pi)\bar{\Phi}\left(\frac{t\widehat{M}(\boldsymbol{\Omega}, \boldsymbol{\delta})}{\sqrt{\widehat{L}_2(\boldsymbol{\Omega}, \boldsymbol{\delta})}} \right) \right\}, \tag{17}$$

and set

$$\widehat{c} = (1-\widehat{t})\widehat{M}_1(\boldsymbol{\Omega}, \boldsymbol{\delta}) + \widehat{t}\widehat{M}_2(\boldsymbol{\Omega}, \boldsymbol{\delta}). \tag{18}$$

Here (17) is a one-dimensional optimization problem and can be solved easily. The resulting QUADRO classification rule is

$$\widehat{h}^{\mathrm{Quad}}(\mathbf{x}) = I\{\mathbf{x}^\top \widehat{\boldsymbol{\Omega}}\mathbf{x} - 2\mathbf{x}^\top \widehat{\boldsymbol{\delta}} - \widehat{c} < 0\}.$$

As a by-product, the method to decide c, described in (17) and (18), can be used in other classification procedures on Gaussian data, such as logistic regression, quadratic discriminant analysis (QDA) and kernel support vector machine, once $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ are given. It provides a fast and purely data-driven way to decide the threshold value in quadratic classification rules. In our numerical experiments, it performs well.

7. Numerical studies. In this section, we investigate the performance of QUADRO in several simulation examples and a real data example. The simulation studies contain both Gaussian models and general elliptical models. We compare QUADRO with several *classification-oriented procedures*. Performances are evaluated in terms of classification errors.

7.1. *Simulations under Gaussian models.* Let $n_1 = n_2 = 50$ and $d = 40$. For each given $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we generate 100 training datasets independently, each with n_1 data from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and n_2 data from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. In QUADRO, we input the sample means and sample covariance matrices. We set $\lambda_2 = r\lambda_1$ and work with λ_1 and r from now on. The two tuning parameters $\lambda_1 \geq 0$ and $r > 0$ are selected in the following way. For various pairs of (λ_1, r), we apply QUADRO for each pair and evaluate the classification error via 4000 newly generated testing data; we then choose the (λ_1, r) that minimize the classification error.

We compare QUADRO with five *classification-oriented procedures*:

- Sparse logistic regression (SLR): We apply the sparse logistic regression to the augmented feature space $\{X_i, 1 \leq i \leq d; X_i X_j, 1 \leq i \leq j \leq d\}$. The resulting estimator then gives a quadratic projection with $(\boldsymbol{\Omega}, \boldsymbol{\delta}, c)$ decided from the fitted regression coefficients. We implement the sparse logistic regression using the R package glmnet.
- Linear sparse logistic regression (L-SLR): We apply the sparse logistic regression directly to the original feature space $\{X_i, 1 \leq i \leq d\}$.
- ROAD [Fan, Feng and Tong (2012)]: This is a linear classification method, which can be formulated equivalently as a modified version of QUADRO by enforcing $\widehat{\Omega}$ as the zero matrix and plugging in the pooled sample covariance matrix.
- Penalized-LDA (P-LDA) [Witten and Tibshirani (2011)]: This is a variant of LDA, which solves an optimization problem with a nonconvex objective and L_1 penalties. Also, P-LDA only uses diagonals of the sample covariance matrices.
- FAIR [Fan and Fan (2008)]: This is a variant of LDA for high-dimensional settings, where screening is adopted to pre-select features and only the diagonals of the sample covariance matrices are used.

To make a fair comparison, the tuning parameters in SLR and L-SLR are selected in the same way as in QUADRO based on 4000 testing data. ROAD and P-LDA are self-tuned by its package. The number of features chosen in FAIR is calculated in the way suggested in [Fan and Fan (2008)].

We consider four models:

– *Model* 1: $\boldsymbol{\Sigma}_1$ is the identity matrix. $\boldsymbol{\Sigma}_2$ is a diagonal matrix in which the first 10 elements are equal to 1.3 and the rest are equal to 1. $\boldsymbol{\mu}_1 = \mathbf{0}$, and $\boldsymbol{\mu}_2 = (0.7, \ldots, 0.7, 0, \ldots, 0)^\top$ with the first 10 elements of $\boldsymbol{\mu}_2$ being nonzero.
– *Model* 1L: $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ are the same as in model 1, and both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the identity matrix.
– *Model* 2: $\boldsymbol{\Sigma}_1$ is a block-diagonal matrix. Its upper left 20×20 block is an equal correlation matrix with $\rho = 0.4$, and its lower right 20×20 block is an identity matrix. $\boldsymbol{\Sigma}_2 = (\boldsymbol{\Sigma}_1^{-1} + \mathbf{I})^{-1}$. We also set $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. In this model, neither $\boldsymbol{\Sigma}_1^{-1}$ nor $\boldsymbol{\Sigma}_2^{-1}$ is sparse, but $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$ is.

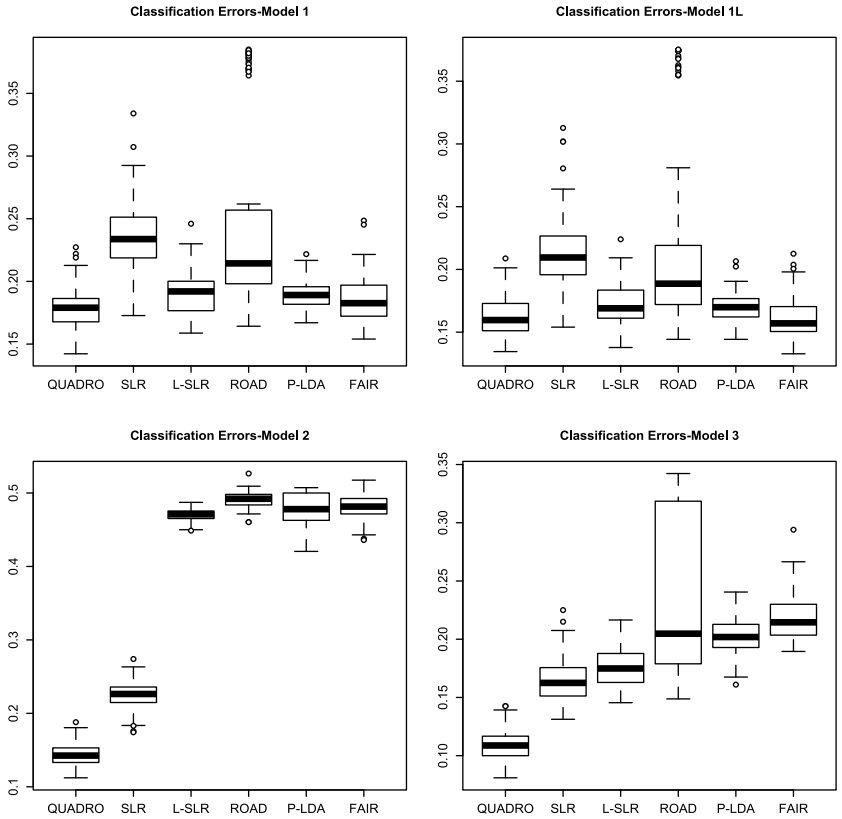


FIG. 3. *Distributions of minimum classification error based on* 100 *replications for four different normal models. The tuning parameters for QUADRO, SLR and L-SLR are chosen to minimize the classification errors of* 4000 *testing samples. See Fan et al.* (2014) *for detailed numerical tables.*

– *Model* 3: $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ and $\boldsymbol{\mu}_1$ are the same as in model 2, and $\boldsymbol{\mu}_2$ is taken from model 1.

Figure 3 contains the boxplots for the classification errors of all methods. In all four models, QUADRO outperforms other methods in terms of classification error. In model 1L, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, so the Bayes classifier is linear. In this case which favors linear methods, QUADRO is still competitive with the best of all linear classifiers. In model 2, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, so linear methods can do no better than random guessing. Therefore, ROAD, L-SLR, P-LDA and FAIR all have very poor performances. For the two quadratic methods, QUADRO is significantly better than SLR. In models 1 and 3, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, so in the Bayes classifier, both "linear" parts and "quadratic" parts play important roles. In model 1, both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal, and the setting favors methods using only diagonals of sample covariance matrices. As a result, P-LDA and FAIR perform quite well. In model 3, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are both nondiagonal and nonsparse (but $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ is sparse). We see that the performances

of P-LDA and FAIR are unsatisfactory. QUADRO outperforms other methods in both models 1 and 3.

Comparing SLR and L-SLR, we see the former considers a broader class, while the latter is more robust, but neither of them perform uniformly better. However, QUADRO performs well in all cases. In terms of Rayleigh quotients, QUADRO also outperforms other methods in most cases.

7.2. *Simulations under elliptical models*. Let $n_1 = n_2 = 50$ and $d = 40$. For each given $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, data are generated from multivariate t distribution with degrees of freedom 5. In QUADRO, we input the robust M-estimators for means and the rank-based estimators for covariance matrices as described in Section 4. We compare the performance of QUADRO with the five methods compared under Gaussian settings. We also implement QUADRO with inputs of sample means and sample covariance matrices. We name this method QUADRO-0 to differentiate it from QUADRO.

We consider three models:

- *Model* 4: Here we use same parameters as those in model 1.
- *Model* 5: $\boldsymbol{\Sigma}_1$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the same as in model 1. $\boldsymbol{\Sigma}_2$ is the covariance matrix of a fractional white noise process, where the difference parameter $l = 0.2$. In other words, $\boldsymbol{\Sigma}_2$ has the polynomial off-diagonal decay $|\Sigma_2(i, j)| = O(|i - j|^{1-2l})$.
- *Model* 6: $\boldsymbol{\Sigma}_1$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the same as in model 1. $\boldsymbol{\Sigma}_2$ is a matrix such that $\Sigma_2(i, j) = 0.6^{|i-j|}$; that is, $\boldsymbol{\Sigma}_2$ has an exponential off-diagonal decay.

Figure 4 contains the boxplots of average classification error over 100 replications. QUADRO outperforms the other methods in all settings. Also, QUADRO is better than QUADRO-0 (e.g., 0.161 versus 0.173, of the average classification error in model 5), which illustrates the advantage of using the robust estimators for means and covariance matrices.

7.3. *Real data analysis*. We apply QUADRO to a large-scale genomic dataset, GPL96, and compare the performance of QUADRO with SLR, L-SLR, ROAD, P-LDA and FAIR. The GPL96 data set contains 20,263 probes and 8124 samples from 309 tissues. Among the tissues, breast tumor has 1142 samples, which is the largest set. We merge the probes from the same gene by averaging them, and finally get 12,679 genes and 8124 samples. We divide all samples into two groups: breast tumor or nonbreast tumor.

First, we look at the classification errors. We replicate our experiment 100 times. Each time, we proceed with the following steps:

- Randomly choose a training set of 400 samples, 200 from breast tumor and 200 from nonbreast tumor.
- For each training set, we use half of the samples to compute $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ and the other half to select the tuning parameters by minimizing the classification error.

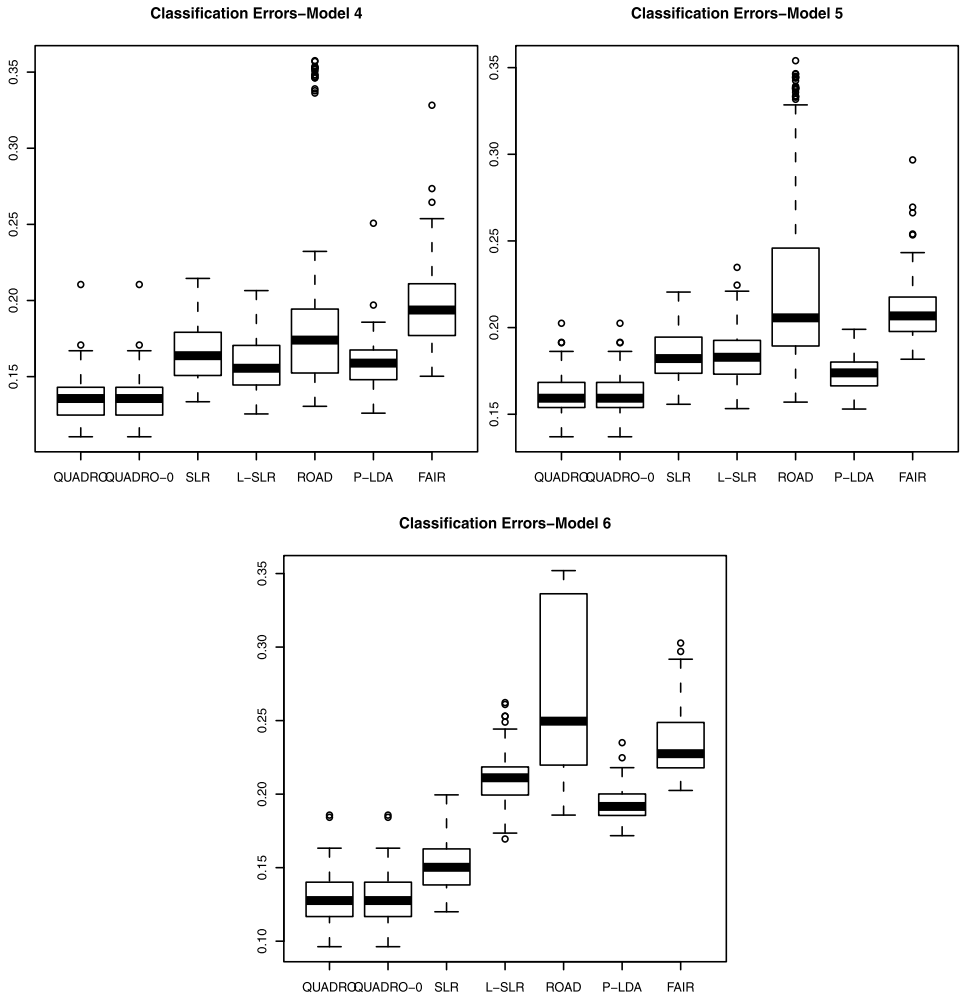


FIG. 4. *Distributions of minimum classification error based on* 100 *replications across different elliptical distribution models. The tuning parameters for QUADRO, SLR and L-SLR are chosen to minimize the classification errors. See Fan et al.* (2014) *for detailed numerical tables.*

- Use the remaining 942 samples from breast tumor and another randomly chosen 942 samples from nonbreast tumor as testing set, and calculate the testing error.

FAIR does not have any tuning parameters, so we use the whole training set to calculate classification frontier, and the rest to calculate testing error. The results are summarized in Table 1. We see that QUADRO outperforms all other methods.

Next, we look at gene selection and focus on the two quadratic methods, QUADRO and SLR. We apply two-fold cross-validation to both QUADRO and SLR. In the results, QUADRO selects 139 genes and SLR selects 128 genes. According to KEGG database, genes selected by QUADRO belong to 5 of the

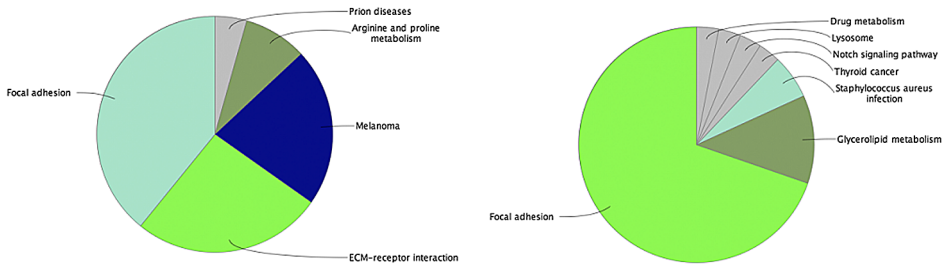
TABLE 1
Classification errors on GPL96 dataset, across methods QUADRO, SLR and L-SLR. Means and standard deviations (in the parenthesis) of 100 replications are reported

QUADRO	SLR	L-SLR	ROAD	Penalized-LDA	FAIR
0.014 (0.007)	0.025 (0.007)	0.025 (0.009)	0.016 (0.007)	0.060 (0.011)	0.046 (0.009)

pathways that contain more than two genes; correspondingly, genes selected by SLR belong to 7 pathways. Using the ClueGo tool [Bindea et al. (2009)], we display the overall KEGG enrichment chart in Figure 5. We see from Figure 5 that both QUADRO and SLR have *focal adhesion* as its most important functional group. Nevertheless, QUADRO finds *ECM-receptor interaction* as another important functional group. *ECM-receptor interaction* is a class consisting of a mixture of structural and functional macromolecules, and it plays an important role in maintaining cell and tissue structures and functions. Massive studies [Luparello (2013), Wei and Li (2007)] have found evidence that this class is closely related to breast cancer.

Besides the pathway analysis, we also perform the Gene Ontology (GO) enrichment analysis on genes selected by QUADRO. This analysis was completed by DAVID Bioinformatics Resources, and the results are shown in Table 2. We present the biological processes with p-values smaller than 10^{-3}. According to the table, we see that many biological processes are significantly enriched, and they are related to previously selected pathways. For instance, the biological process *cell adhesion* is known to be highly related to *cell communication pathways*, including *focal adhesion* and *ECM-receptor interaction*.

8. Conclusions and extensions. QUADRO is a robust sparse high-dimensional classifier, which allows us to use differences in covariance matrices to enhance discriminability. It is based on Rayleigh quotient optimization. The variance of



(a) QUADRO pathways (b) SLR pathways

FIG. 5. *Overall KEGG enrichment chart, using* (a) *QUADRO*; (b) *SLR*.

TABLE 2

Enrichment analysis results according to Gene Ontology for genes selected by QUADRO. The four columns represent GO ID, GO attribute, number of selected genes having the attribute and their corresponding p-values. We rank them according to p-values in increasing order

GO ID	GO attribute	No. of genes	*p*-value
0048856	Anatomical structure development	58	3.7E–12
0032502	Developmental process	62	2.9E–10
0048731	System development	52	3.1E–10
0007275	Multicellular organismal development	55	1.8E–8
0001501	Skeletal system development	15	1.3E–6
0032501	Multicellular organismal process	66	1.4E–6
0048513	Organ development	37	1.4E–6
0009653	Anatomical structure morphogenesis	28	8.7E–6
0048869	Cellular developmental process	34	1.9E–5
0030154	Cell differentiation	33	2.1E–5
0007155	Cell adhesion	18	2.4E–4
0022610	Biological adhesion	18	2.2E–4
0042127	Regulation of cell proliferation	19	2.9E–4
0009888	Tissue development	17	3.7E–4
0007398	Ectoderm development	9	4.8E–4
0048518	Positive regulation of biological process	34	5.6E–4
0009605	Response to external stimulus	20	6.3E–4
0043062	Extracellular structure organization	8	7.4E–4
0007399	Nervous system development	22	8.4E–4

quadratic statistics involves all fourth cross moments, and this can create both computational and statistical problems. These problems are avoided by limiting our applications to the elliptical class of distributions. Robust M-estimator and rank-based estimation of correlations allow us to obtain the uniform convergence for nonpolynomially many parameters, even when the underlying distributions have the finite fourth moments. This allows us to establish oracle inequalities under relatively weaker conditions.

Existing methods in the literature about constructing high-dimensional quadratic classifiers can be divided into two types. One is the regularized QDA, where regularized estimates of $\mathbf{\Sigma}_1^{-1}$ and $\mathbf{\Sigma}_2^{-1}$ are plugged into the Bayes classifier; see, for example, Friedman (1989). QUADRO avoids directly estimating inverse covariance matrices, which requires strong assumptions in high dimensions. The other is to combine linear classifiers with the inner-product kernel. The main difference between QUADRO and this approach is the simplification in Proposition 2.1. Due to this simplification, QUADRO avoids incorporating all fourth cross moments from the data and gains extra statistical efficiency.

QUADRO also has deep connections with the literature of sufficient dimension reduction. Dimension reduction methods, such as SIR [Li (1991)], SAVE [Cook and Weisberg (1991)] and Directional Regression [Li and Wang (2007)], can be

equivalently formulated as maximizing some "quotients." The population objective of SIR is to maximize $\operatorname{var}\{\mathbb{E}[f(\mathbf{X}|Y)]\}$ subject to $\operatorname{var}[f(\mathbf{X})] = 1$. Using the same constraint, SAVE and directional regression combine $\operatorname{var}\{\mathbb{E}[f(\mathbf{X}|Y)]\}$ and $\mathbb{E}[\operatorname{var}(f(\mathbf{X}|Y))]$ in the objective. An interesting observation is that the Rayleigh quotient maximization is equivalent to the population objective of SIR, by noting that the denominator of (1) is equal to $\mathbb{E}[\operatorname{var}(f(\mathbf{X}|Y))]$ and $\operatorname{var}[f(\mathbf{X})] = \mathbb{E}[\operatorname{var}(f(\mathbf{X}|Y))] + \operatorname{var}\{\mathbb{E}[f(\mathbf{X}|Y)]\}$. This is not a coincidence, but due instead to the known equivalence between SIR and LDA in classification [Kent (1991), Li (2000)].

Despite similar population objectives, QUADRO and the aforementioned dimension reduction methods are different in important ways. First, we clarify that even when λ_1, λ_2 are 0, QUADRO is not the same procedure as SIR combined with the inner-product kernel [Wu (2008)], although they share the same population objective. The difference is that QUADRO utilizes a simplification of the Rayleigh quotient for quadratic f, relying on the assumption that $\mathbf{X}|Y$ is always elliptically distributed; moreover, it adopts robust estimators of the mean vectors and covariance matrices. Second, QUADRO is designed for high-dimensional settings, in which neither SIR, SAVE nor Directional Regression can be directly implemented. These methods need to either standardize the original data $\mathbf{X} \mapsto \widehat{\mathbf{\Sigma}}^{-1}(\mathbf{X} - \bar{\mathbf{X}})$ or solve a generalized eigen-decomposition problem $\mathbf{A}\mathbf{v} = \lambda\widehat{\mathbf{\Sigma}}\mathbf{v}$ for some matrix $\mathbf{A}$. Both methods require that the sample covariance matrix is well conditioned, which is often not the case in high dimensions. Possible solutions include Regularized SIR [Li and Yin (2008), Zhong et al. (2005)], solving generalized eigen-decomposition for an undetermined system [Coudret, Liquet and Saracco (2014)] and variable selection approaches [Chen, Zou and Cook (2010), Jiang and Liu (2013)]. However, these methods are not designed for Rayleigh quotient maximization. Third, our assumption on the model is different from that in dimension reduction. We require $\mathbf{X}|Y$ to be elliptically distributed, while many dimension reduction methods "implicitly" require $\mathbf{X}$ to be marginally elliptically distributed. Neither method is stronger than the other. Assuming conditional elliptical distribution is more natural in classification. In addition, our assumption is used only to simplify the variances of quadratic statistics, whereas the elliptical assumption is critical to SIR.

The Rayleigh optimization framework developed in this paper can be extended to the multi-class case. Suppose the data are drawn independently from a joint distribution of $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^d$ and Y takes values in $\{0, 1, \ldots, K-1\}$. Definition (1) for the Rayleigh quotient of a projection $f : \mathbb{R}^d \to \mathbb{R}$ is still well defined. Let $\pi_k = \mathbb{P}(Y = k)$, for $k = 0, 1, \ldots, K-1$. In this K-class situation,

$$\operatorname{Rq}(f) = \frac{\sum_{0 \le k < l \le K-1} \pi_k \pi_l \{\mathbb{E}[f(\mathbf{X})|Y = k] - \mathbb{E}[f(\mathbf{X})|Y = l]\}^2}{\sum_{0 \le k \le K-1} \pi_k \operatorname{var}[f(\mathbf{X})|Y = k]}. \tag{19}$$

Let $M_k(f) = \mathbb{E}[f(\mathbf{X})|Y = k]$ and $L_k(f) = \operatorname{var}[f(\mathbf{X})|Y = k]$. Similar to the two-class case, maximizing $\operatorname{Rq}(f)$ is equivalent to solving the following optimization

problem:

$$\min_f \sum_{k=0}^{K-1} \pi_k L_k(f) \quad \text{s.t.} \quad \sum_{0\le k<l\le K-1} \pi_k\pi_l \big|M_k(f) - M_l(f)\big|^2 = 1.$$

However, this is not a convex problem. We consider an approximate Rayleigh-quotient-maximization problem as follows:

$$\min_f \sum_{k=0}^{K-1} \pi_k L_k(f) \quad \text{s.t.} \quad \sqrt{\pi_k\pi_l}\big|M_k(f) - M_l(f)\big| \ge 1, \qquad 0 \le k < l \le K-1.$$

To solve this problem, we first pick an order of $M_1(f), \ldots, M_K(f)$ to remove the absolute values in the constraints. Then it becomes a convex problem. Therefore, the whole optimization can be carried out by simultaneously solving $K!$ convex problems. When K is small, the computational cost is reasonable. In practice, we can apply more efficient algorithms to speed up the computation.

9. Proofs.

9.1. *Proof of Theorem* 5.1. We prove the claim by first rewriting optimization problem (8) into a vector form. For any $(\boldsymbol{\Omega}, \boldsymbol{\delta})$, write $\mathbf{x} = [\mathrm{vec}(\boldsymbol{\Omega})^\top, \boldsymbol{\delta}^\top]^\top$. Let $\mathbf{Q}$ be as defined in Section 5, and

$$\mathbf{q} = \big[\,\mathrm{vec}\big(\boldsymbol{\Sigma}_2 + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top - \boldsymbol{\Sigma}_1 - \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top\big)^\top, 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\,\big]^\top.$$

We introduce the following lemma which is proved in the supplementary material [Fan et al. (2014)].

LEMMA 9.1. *$M(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \mathbf{q}^\top\mathbf{x}$ and $L(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \mathbf{x}^\top\mathbf{Q}\mathbf{x}$.*

Let $\mathbf{x}^*_{\lambda_0} = [\mathrm{vec}(\boldsymbol{\Omega}^*_{\lambda_0})^\top, (\boldsymbol{\delta}^*_{\lambda_0})^\top]^\top$ and $\widehat{\mathbf{x}} = [\mathrm{vec}(\widehat{\boldsymbol{\Omega}})^\top, \widehat{\boldsymbol{\delta}}^\top]^\top$. Using Lemma 9.1,

$$\mathbf{x}^*_{\lambda_0} = \min_{\mathbf{x}\in\mathbb{R}^d:\, \mathbf{q}^\top\mathbf{x}=1} \big\{\mathbf{x}^\top\mathbf{Q}\mathbf{x} + \lambda_0|\mathbf{x}|_1\big\},$$

$$\widehat{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d:\, \widehat{\mathbf{q}}^\top\mathbf{x}=1} \big\{\mathbf{x}^\top\widehat{\mathbf{Q}}\mathbf{x} + \lambda|\mathbf{x}|_1\big\},$$

where $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{q}}$ are counterparts of $\mathbf{Q}$ and $\mathbf{q}$, respectively, by replacing $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ with their estimates. Moreover, we have the Rayleigh quotient

$$R(\boldsymbol{\Omega}, \boldsymbol{\delta}) = R(\mathbf{x}) \equiv \frac{(\mathbf{q}^\top\mathbf{x})^2}{\mathbf{x}^\top\mathbf{Q}\mathbf{x}}.$$

In addition, we have the following lemma, which is proved in the supplementary material [Fan et al. (2014)].

LEMMA 9.2. *$\max\{|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty, |\widehat{\mathbf{q}} - \mathbf{q}|_\infty\} \le C_0 \max\{|\widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k|_\infty, |\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty, k = 1, 2\}$ for some constant $C_0 > 0$.*

Combining the above results, the claim follows immediately from the following theorem:

THEOREM 9.1. *For any* $\lambda_0 \geq 0$, *let* S *be the support of* $\mathbf{x}^*_{\lambda_0}$. *Suppose* $\Theta(S,0) \geq c_0$, $\Theta(S,3) \geq a_0$ *and* $R(\mathbf{x}^*_{\lambda_0}) \geq u_0$, *for positive constants* a_0, c_0 *and* u_0. *Let* $\Delta_n = \max\{|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty, |\widehat{\mathbf{q}} - \mathbf{q}|_\infty\}$, $s_0 = |S|$ *and* $k_0 = \max\{s_0, R(\mathbf{x}^*_{\lambda_0})\}$. *Suppose* $4s_0\Delta_n^2 < c_0u_0$ *and* $\max\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\} < 1$. *Then there exist positive constants* $C = C(a_0, c_0, u_0)$ *and* $A = A(a_0, c_0, u_0)$, *such that for any* $\eta > 1$, *by taking* $\lambda = C\eta \max\{s_0^{1/2}\Delta_n, k_0^{1/2}\lambda_0\}[R(\mathbf{x}^*_{\lambda_0})]^{-1/2}$,

$$\frac{R(\widehat{\mathbf{x}})}{R(\mathbf{x}^*_{\lambda_0})} \geq 1 - A\eta^2 \max\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\}.$$

The main part of the proof is to show Theorem 9.1. Write for short $\mathbf{x}^* = \mathbf{x}^*_{\lambda_0}$, $R^* = R(\mathbf{x}^*)$, $V^* = (R^*)^{-1} = (\mathbf{x}^*)^\top \mathbf{Q}\mathbf{x}^*$, $\bar{V}^* = (V^*)^{1/2}$. Let $\alpha_n = \Delta_n|\mathbf{x}^*|_0^{1/2}$, $\beta_n = \Delta_n|\mathbf{x}^*|_0$ and $T_n(\mathbf{x}^*) = \max\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\}$. We define the quantity

$$\Gamma(\mathbf{x}) = \frac{|\mathbf{Q}\mathbf{x} - (\mathbf{x}^\top\mathbf{Q}\mathbf{x})\mathbf{q}|_\infty}{(\mathbf{x}^\top\mathbf{Q}\mathbf{x})^{1/2}} \qquad \text{for any } \mathbf{x}.$$

Step 1. We introduce $\mathbf{x}^*_1$, a multiple of $\mathbf{x}^*$, and use it to bound $|\widehat{\mathbf{x}}|_1$.

Let $\mathbf{Q}_{SS}$ be the submatrix of $\mathbf{Q}$ formed by rows and columns corresponding to S. Since $\lambda_{\min}(\mathbf{Q}_{SS}) = \Theta(S,0) \geq c_0$, we have $(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^* \geq c_0|\mathbf{x}^*|^2$. Using this fact and by the Cauchy–Schwarz inequality,

$$|\mathbf{x}^*|_1 \leq \sqrt{|\mathbf{x}^*|_0}|\mathbf{x}^*| \leq c_0^{-1/2}\sqrt{|\mathbf{x}^*|_0}\bar{V}^*. \tag{20}$$

It follows that

$$|\widehat{\mathbf{q}}^\top\mathbf{x}^* - \mathbf{q}^\top\mathbf{x}^*| \leq |\widehat{\mathbf{q}} - \mathbf{q}|_\infty|\mathbf{x}^*|_1 \leq c_0^{-1/2}\Delta_n\sqrt{|\mathbf{x}^*|_0}\bar{V}^* = c_0^{-1/2}\alpha_n\bar{V}^*. \tag{21}$$

Let $t_n = \widehat{\mathbf{q}}^\top\mathbf{x}^*$. Then (21) says that $|t_n - 1| \leq c_0^{-1/2}\alpha_n\bar{V}^*$. Noting that $\bar{V}^* = (R^*)^{1/2} \leq u_0^{-1/2}$, we have $|t_n - 1| \leq (c_0u_0)^{-1/2}s_0^{1/2}\Delta_n < 1/2$ by assumption. In particular, $t_n > 0$. Let

$$\mathbf{x}^*_1 = t_n^{-1}\mathbf{x}^*.$$

Then $\widehat{\mathbf{q}}^\top\mathbf{x}^*_1 = 1$. From the definition of $\widehat{\mathbf{x}}$,

$$\widehat{\mathbf{x}}^\top\widehat{\mathbf{Q}}\widehat{\mathbf{x}} + \lambda|\widehat{\mathbf{x}}|_1 \leq (\mathbf{x}^*_1)^\top\widehat{\mathbf{Q}}\mathbf{x}^*_1 + \lambda|\mathbf{x}^*_1|_1. \tag{22}$$

By direct calculation,

$$\begin{aligned}
\widehat{\mathbf{x}}^\top\widehat{\mathbf{Q}}\widehat{\mathbf{x}} - (\mathbf{x}^*_1)^\top\widehat{\mathbf{Q}}\mathbf{x}^*_1 &= (\widehat{\mathbf{x}} - \mathbf{x}^*_1)^\top\widehat{\mathbf{Q}}(\widehat{\mathbf{x}} - \mathbf{x}^*_1) + 2(\widehat{\mathbf{x}} - \mathbf{x}^*_1)^\top\widehat{\mathbf{Q}}\mathbf{x}^*_1 \\
&= (\widehat{\mathbf{x}} - \mathbf{x}^*_1)^\top\widehat{\mathbf{Q}}(\widehat{\mathbf{x}} - \mathbf{x}^*_1) + 2(\widehat{\mathbf{x}} - \mathbf{x}^*_1)^\top(\widehat{\mathbf{Q}}\mathbf{x}^*_1 - V^*\widehat{\mathbf{q}}) \\
&\geq 2(\widehat{\mathbf{x}} - \mathbf{x}^*_1)^\top(\widehat{\mathbf{Q}}\mathbf{x}^*_1 - V^*\widehat{\mathbf{q}}),
\end{aligned} \tag{23}$$

where the second equality is due to $\widehat{\mathbf{q}}^\top \widehat{\mathbf{x}} = \widehat{\mathbf{q}}^\top \mathbf{x}_1^* = 1$. We aim to bound $|\widehat{\mathbf{Q}}\mathbf{x}_1^* - V^*\widehat{\mathbf{q}}|_\infty$. The following lemma is proved in the supplementary material [Fan et al. (2014)].

LEMMA 9.3. *When $\Theta(S, 0) \geq c_0$, there exists a positive constant $C_1 = C_1(c_0)$ such that $\Gamma(\mathbf{x}_{\lambda_0}^*) \leq C_1\lambda_0[\max\{s_0, R(\mathbf{x}_{\lambda_0}^*)\}]^{1/2}$ for any $\lambda_0 \geq 0$.*

Since $\mathbf{x}_1^* = t_n^{-1}\mathbf{x}^*$ and $t_n^{-1} < 2$,

$$
\begin{aligned}
&|\widehat{\mathbf{Q}}\mathbf{x}_1^* - V^*\widehat{\mathbf{q}}|_\infty \\
&\quad \leq t_n^{-1}|\widehat{\mathbf{Q}}\mathbf{x}^* - V^*\widehat{\mathbf{q}}|_\infty + V^*|t_n^{-1} - 1||\widehat{\mathbf{q}}|_\infty \\
&\quad \leq 2\big(|\mathbf{Q}\mathbf{x}^* - V^*\mathbf{q}|_\infty + |\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty|\mathbf{x}^*|_1 + V^*|\widehat{\mathbf{q}} - \mathbf{q}|_\infty + V^*|t_n - 1||\widehat{\mathbf{q}}|_\infty\big) \\
&\quad \leq 2\big[\Gamma(\mathbf{x}^*)\bar{V}^* + c_0^{-1/2}\alpha_n\bar{V}^* + u_0^{-1/2}\Delta_n\bar{V}^* + |\widehat{\mathbf{q}}|_\infty c_0^{-1/2}u_0^{-1}\alpha_n\bar{V}^*\big] \\
&\quad \leq C_2\big(\lambda_0 k_0^{1/2} + s_0^{1/2}\Delta_n\big)\bar{V}^*.
\end{aligned}
$$

Here the third inequality follows from (20)–(21) and $V^* = \bar{V}^*(R^*)^{-1/2} \leq u_0^{-1/2}\bar{V}^*$. The last inequality is obtained as follows: from Lemma 9.2, we know that $|\widehat{\mathbf{q}}|_\infty \leq |\mathbf{q}|_\infty + |\widehat{\mathbf{q}} - \mathbf{q}|_\infty \leq 2C_0$ (see also the assumptions in the beginning of Section 5.2); we also use Lemma 9.3 and $\alpha_n\bar{V}^* \leq u_0^{-1/2}s_0^{1/2}\Delta_n$. By letting $C = 8C_2$, the choice of $\lambda = C\eta\max\{s_0^{1/2}\Delta_n, k_0^{1/2}\lambda_0\}\bar{V}^*$ for $\eta > 1$ ensures that

$$
|\widehat{\mathbf{Q}}\mathbf{x}_1^* - \widehat{\mathbf{q}}|_\infty \leq \lambda/4.
$$

Plugging this result into (23) gives

$$
\widehat{\mathbf{x}}^\top\widehat{\mathbf{Q}}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top\widehat{\mathbf{Q}}\mathbf{x}_1^* \geq -\frac{\lambda}{2}|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1. \tag{24}
$$

Combining (22) and (24) gives

$$
\lambda|\widehat{\mathbf{x}}|_1 - \frac{\lambda}{2}|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1 \leq \lambda|\mathbf{x}_1^*|_1. \tag{25}
$$

First, since $|\widehat{\mathbf{x}}|_1 = |\widehat{\mathbf{x}}_S|_1 + |\widehat{\mathbf{x}}_{S^c}|_1 \geq |\mathbf{x}_{1S}^*|_1 - |\widehat{\mathbf{x}}_S - \mathbf{x}_{1S}^*|_1 + |\widehat{\mathbf{x}}_{S^c}|_1$ and $|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1 = |\widehat{\mathbf{x}}_S - \mathbf{x}_{1S}^*|_1 + |\widehat{\mathbf{x}}_{S^c}|_1$, we immediately see from (25) that

$$
|(\widehat{\mathbf{x}} - \mathbf{x}_1^*)_{S^c}|_1 \leq 3|(\widehat{\mathbf{x}} - \mathbf{x}_1^*)_S|_1. \tag{26}
$$

Second, note that $|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1 \leq |\widehat{\mathbf{x}}|_1 + |\mathbf{x}_1^*|_1$. Plugging this into (25) gives

$$
|\widehat{\mathbf{x}}|_1 \leq 3|\mathbf{x}_1^*|_1 = 3t_n^{-1}|\mathbf{x}^*|_1 \leq 6c_0^{-1/2}\sqrt{|\mathbf{x}^*|_0}\bar{V}^*. \tag{27}
$$

Step 2. We use (26)–(27) to derive an upper bound for $(\widehat{\mathbf{x}})^\top\mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top\mathbf{Q}\mathbf{x}_1^*$.

Note that

$$
\begin{aligned}
&\widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* \\
&\quad \geq \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - \big(|\widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}}| + |(\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^*|\big) \\
&\quad \geq \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - \big(|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty |\widehat{\mathbf{x}}|_1^2 + |\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty |\mathbf{x}_1^*|_1^2\big) \qquad (28) \\
&\quad \geq \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - 10 t_n^{-2} |\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty |\mathbf{x}^*|_1^2 \\
&\quad \geq \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - C_3 \beta_n V^*,
\end{aligned}
$$

where the last two inequalities are direct results of (27). Combining (22) and (28),

$$
\widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} + \lambda |\widehat{\mathbf{x}}|_1 \leq (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* + \lambda |\mathbf{x}_1^*|_1 + C_3 \beta_n V^*. \qquad (29)
$$

Similar to (23), we have

$$
\widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* = (\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \mathbf{Q}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) + 2(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top (\mathbf{Q}\mathbf{x}_1^* - V^* \widehat{\mathbf{q}}), \qquad (30)
$$

where

$$
\begin{aligned}
|\mathbf{Q}\mathbf{x}_1^* - V^* \widehat{\mathbf{q}}|_\infty &\leq t_n^{-1} \big(|\mathbf{Q}\mathbf{x}^* - V^* \mathbf{q}|_\infty + V^* |\widehat{\mathbf{q}} - \mathbf{q}|_\infty\big) + V^* |t_n^{-1} - 1| |\widehat{\mathbf{q}}|_\infty \\
&\leq 2\big[\Gamma(\mathbf{x}^*) \bar{V}^* + u_0^{-1/2} \Delta_n \bar{V}^* + |\widehat{\mathbf{q}}|_\infty c_0^{-1/2} u_0^{-1} \alpha_n \bar{V}^*\big] \\
&\leq \lambda/4.
\end{aligned}
$$

It follows that

$$
\widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* \geq (\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \mathbf{Q}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) - \frac{\lambda}{2} |\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1.
$$

Plugging this into (29), we obtain

$$
(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \mathbf{Q}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) + \lambda |\widehat{\mathbf{x}}|_1 - \frac{\lambda}{2} |\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1 \leq \lambda |\mathbf{x}_1^*|_1 + C_3 \beta_n V^*. \qquad (31)
$$

We can rewrite the second and third terms on the left-hand side of (31) as

$$
\lambda |\widehat{\mathbf{x}}_S|_1 - \frac{\lambda}{2} |\widehat{\mathbf{x}}_S - \mathbf{x}_{1S}^*|_1 + \frac{\lambda}{2} |\widehat{\mathbf{x}}_{S^c}|_1.
$$

Plugging this into (31) and by the triangular inequality $|\mathbf{x}_{1S}^*|_1 - |\widehat{\mathbf{x}}_S|_1 \leq |\widehat{\mathbf{x}}_S - \mathbf{x}_{1S}^*|_1$, we find that

$$
(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \mathbf{Q}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) + \frac{\lambda}{2} |\widehat{\mathbf{x}}_{S^c}|_1 \leq \frac{3\lambda}{2} |\widehat{\mathbf{x}}_S - \mathbf{x}_{1S}^*|_1 + C_3 \beta_n V^*.
$$

We drop the term $\frac{\lambda}{2} |\widehat{\mathbf{x}}_{S^c}|_1$ on the left-hand side and apply the Cauchy–Schwarz inequality to the term $|\widehat{\mathbf{x}}_S - \mathbf{x}_{1S}^*|_1$. This gives

$$
(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \mathbf{Q}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) \leq \frac{3\lambda}{2} \sqrt{|\mathbf{x}_1^*|_0} |\widehat{\mathbf{x}}_{1S} - \mathbf{x}_{1S}^*| + C_3 \beta_n V^*. \qquad (32)
$$

Since (26) holds, by the definition of $\Theta(S,3)$,

$$(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)\geq a_0|\widehat{\mathbf{x}}_S-\mathbf{x}_{1S}^*|^2.$$

We write temporarily $Y=(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)$ and $b=C_3\beta_nV^*$. Combining these with (32),

$$Y\leq\frac{3\lambda}{2\sqrt{a_0}}\sqrt{|\mathbf{x}_1^*|_0Y}+b.$$

Note that when $u^2\leq au+b$, we have $(u-\frac{a}{2})^2\leq b+\frac{a^2}{4}$, and hence $u^2\leq 2[\frac{a^2}{4}+(u-\frac{a}{2})^2]\leq a^2+2b$. As a result, the above inequality implies

$$(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)\leq\frac{9\lambda^2}{4a_0}|\mathbf{x}^*|_0+2C_3\beta_nV^*, \tag{33}$$

where we have used $|\mathbf{x}_1^*|_0=|\mathbf{x}^*|_0$. Furthermore, (30) yields that

$$\begin{aligned}\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}-(\mathbf{x}_1^*)^\top\mathbf{Q}\mathbf{x}_1^*&\leq(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)+\frac{\lambda}{2}|\widehat{\mathbf{x}}-\mathbf{x}_1^*|_1\\&\leq(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)+2\lambda|\mathbf{x}_1^*|_1 \qquad (34)\\&\leq(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)+4c_0^{-1/2}\bar{V}^*\lambda\sqrt{|\mathbf{x}^*|_0},\end{aligned}$$

where the second inequality is due to $|\widehat{\mathbf{x}}-\mathbf{x}_1^*|_1\leq|\widehat{\mathbf{x}}|_1+|\mathbf{x}_1^*|\leq 4|\mathbf{x}_1^*|_1$, and the last inequality is from (27). Recall that $\lambda=C\eta\max\{k_0^{1/2}\lambda_0,s_0^{1/2}\Delta_n\}\bar{V}^*$. As a result,

$$\lambda\sqrt{|\mathbf{x}^*|_0}=C\eta\max\{k_0^{1/2}s_0^{1/2}\lambda_0,s_0\Delta_n\}\bar{V}^*=C\eta T_n(\mathbf{x}^*)\bar{V}^*. \tag{35}$$

Combining (33), (34) and (35) gives

$$\begin{aligned}&\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}-(\mathbf{x}_1^*)^\top\mathbf{Q}\mathbf{x}_1^*\\&\leq\frac{9C^2}{4a_0}\eta^2[T_n(\mathbf{x}^*)]^2V^*+4Cc_0^{-1/2}\eta T_n(\mathbf{x}^*)V^*+2C_3\beta_nV^* \qquad (36)\\&\leq C_4\eta^2T_n(\mathbf{x}^*)V^*.\end{aligned}$$

Step 3. We use (36) to give a lower bound of $R(\widehat{\mathbf{x}})$.

Note that $R(\widehat{\mathbf{x}})=(\mathbf{q}^\top\widehat{\mathbf{x}})^2/(\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}})$. First, we look at the denominator $\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}$. From (21) and that $t_n>1/2$,

$$|t_n^{-2}-1|=t_n^{-1}(1+t_n^{-1})|t_n-1|\leq 6c_0^{-1/2}\alpha_n\bar{V}^*.$$

Combining with (36) and noting that $(\mathbf{x}_1^*)^\top\mathbf{Q}\mathbf{x}_1^*=t_n^{-2}(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*=t_n^{-2}V^*$, we have

$$\begin{aligned}\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}&\leq[t_n^{-2}+C_4\eta^2T_n(\mathbf{x}^*)](\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*\\&\leq[1+6c_0^{-1/2}\alpha_n\bar{V}^*+C_4\eta^2T_n(\mathbf{x}^*)](\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^* \qquad (37)\\&\leq[1+C_5\eta^2T_n(\mathbf{x}^*)](\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*.\end{aligned}$$

Second, we look at the numerator $\mathbf{q}^\top\widehat{\mathbf{x}}$. Since $\widehat{\mathbf{q}}^\top\widehat{\mathbf{x}} = 1$, by (27),

$$|\mathbf{q}^\top\widehat{\mathbf{x}} - 1| \le |\widehat{\mathbf{q}} - \mathbf{q}|_\infty |\widehat{\mathbf{x}}|_1 \le 6c_0^{-1/2}\alpha_n \bar{V}^* \le C_6 T_n(\mathbf{x}^*). \tag{38}$$

Combining (37) and (38) gives

$$\begin{aligned} R(\widehat{\mathbf{x}}) &= \frac{(\mathbf{q}^\top\widehat{\mathbf{x}})^2}{\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}} \ge \frac{[1 - C_6 T_n(\mathbf{x}^*)]^2}{1 + C_5\eta^2 T_n(\mathbf{x}^*)} \frac{1}{(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*} \\ &\ge \big[1 - A\eta^2 T_n(\mathbf{x}^*)\big] \frac{(\mathbf{q}^\top\mathbf{x}^*)^2}{(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*} \\ &= \big[1 - A\eta^2 T_n(\mathbf{x}^*)\big] R(\mathbf{x}^*), \end{aligned} \tag{39}$$

where $A = A(a_0, c_0, u_0)$ is a positive constant.

9.2. *Proof of Proposition 6.1.* Denote by $\mathbb{P}(i|j)$ the probability that a new sample from class j is misclassified to class i, for $i, j \in \{1, 2\}$ and $i \ne j$. The classification error of h is

$$\operatorname{err}(h) = \pi\mathbb{P}(2|1) + (1 - \pi)\mathbb{P}(1|2).$$

Write $M_k = M_k(\boldsymbol{\Omega}, \boldsymbol{\delta})$ and $L_k = L_k(\boldsymbol{\Omega}, \boldsymbol{\delta})$ for short. It suffices to show that

$$\begin{aligned} \mathbb{P}(2|1) &= \bar{\Phi}\Big(\frac{(1-t)M}{\sqrt{L_1}}\Big) + \frac{O(q) + o(d)}{L_1^{3/2}}, \\ \mathbb{P}(1|2) &= \bar{\Phi}\Big(\frac{tM}{\sqrt{L_2}}\Big) + \frac{O(q) + o(d)}{L_2^{3/2}}. \end{aligned}$$

We only consider $\mathbb{P}(2|1)$. The analysis of $\mathbb{P}(1|2)$ is similar. Suppose $\mathbf{X}|\text{class } 1 \overset{(d)}{=} \mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Define

$$\mathbf{Y} = \boldsymbol{\Sigma}_1^{-1/2}(\mathbf{Z} - \boldsymbol{\mu}_1),$$

so that $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathbf{Z} = \boldsymbol{\Sigma}_1^{1/2}\mathbf{Y} + \boldsymbol{\mu}_1$. Note that

$$\begin{aligned} Q(\mathbf{Z}) &= (\boldsymbol{\Sigma}_1^{1/2}\mathbf{Y} + \boldsymbol{\mu}_1)^\top \boldsymbol{\Omega}(\boldsymbol{\Sigma}_1^{1/2}\mathbf{Y} + \boldsymbol{\mu}_1) - 2(\boldsymbol{\Sigma}_1^{1/2}\mathbf{Y} + \boldsymbol{\mu}_1)^\top \boldsymbol{\delta} \\ &= \mathbf{Y}^\top \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}_1^{1/2}\mathbf{Y} + 2\mathbf{Y}^\top \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}) + \boldsymbol{\mu}_1^\top \boldsymbol{\Omega}\boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_1^\top \boldsymbol{\delta}. \end{aligned} \tag{40}$$

Recall that $\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}_1^{1/2} = \mathbf{K}_1\mathbf{S}_1\mathbf{K}_1^\top$ is the eigen-decomposition by excluding the 0 eigenvalues. Since $\boldsymbol{\Sigma}_1$ has full rank and the rank of $\boldsymbol{\Omega}$ is q, the rank of $\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}_1^{1/2}$ is q. Therefore, $\mathbf{S}_1$ is a $q \times q$ diagonal matrix, and $\mathbf{K}_1$ is a $d \times q$ matrix satisfying $\mathbf{K}_1^\top\mathbf{K}_1 = \mathbf{I}_q$. Let $\widetilde{\mathbf{K}}_1$ be any $d \times (d - q)$ matrix such that $\mathbf{K} = [\mathbf{K}_1, \widetilde{\mathbf{K}}_1]$ is a $d \times d$ orthogonal matrix. Since $\mathbf{I}_d = \mathbf{K}\mathbf{K}^\top = \mathbf{K}_1\mathbf{K}_1^\top + \widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_1^\top$, we have

$$\mathbf{Y}^\top \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}) = \mathbf{Y}^\top \mathbf{K}_1\mathbf{K}_1^\top \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}) + \mathbf{Y}^\top \widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_1^\top \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}).$$

We recall that $\boldsymbol{\beta}_1 = \mathbf{K}_1^\top \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\beta})$. Let $\widetilde{\boldsymbol{\beta}}_1 = \widetilde{\mathbf{K}}_1^\top \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})$, $\mathbf{W} = \mathbf{K}_1^\top \mathbf{Y}$, $\widetilde{\mathbf{W}} = \widetilde{\mathbf{K}}_1^\top \mathbf{Y}$ and $c_1 = \boldsymbol{\mu}_1^\top \boldsymbol{\Omega}\boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_1^\top \boldsymbol{\delta}$. It follows from (40) that

$$\begin{aligned} Q(\mathbf{Z}) &= \mathbf{Y}^\top \mathbf{K}_1 \mathbf{S}_1 \mathbf{K}_1^\top \mathbf{Y} + 2\mathbf{Y}^\top \mathbf{K}_1 \boldsymbol{\beta}_1 + 2\mathbf{Y}^\top \widetilde{\mathbf{K}}_1 \widetilde{\boldsymbol{\beta}}_1 + c_1 \\ &= \mathbf{W}^\top \mathbf{S}_1 \mathbf{W} + 2\mathbf{W}^\top \boldsymbol{\beta}_1 + 2\widetilde{\mathbf{W}}^\top \widetilde{\boldsymbol{\beta}}_1 + c_1 \\ &\equiv \bar{Q}_1(\mathbf{W}) + \bar{F}_1(\widetilde{\mathbf{W}}) + c_1, \end{aligned}$$

where $\bar{Q}_1(\mathbf{w}) = \mathbf{w}^\top \mathbf{S}_1 \mathbf{w} + 2\mathbf{w}^\top \boldsymbol{\beta}_1$ and $\bar{F}_1(\mathbf{w}) = 2\mathbf{w}^\top \widetilde{\boldsymbol{\beta}}_1$. Therefore,

$$\mathbb{P}(2|1) = \mathbb{P}\big(Q(\mathbf{Z}) > c\big) = \mathbb{P}\big(\bar{Q}_1(\mathbf{W}) + \bar{F}_1(\widetilde{\mathbf{W}}) > c - c_1\big).$$

We write for convenience $\mathbf{W} = (W_1, \ldots, W_q)^\top$, $\widetilde{\mathbf{W}} = (W_{q+1}, \ldots, W_d)^\top$, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1q})^\top$ and $\widetilde{\boldsymbol{\beta}}_1 = (\beta_{1(q+1)}, \ldots, \beta_{1d})^\top$, and notice that $W_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ for $1 \le i \le d$. Moreover,

$$\bar{Q}_1(\mathbf{W}) + \bar{F}_1(\widetilde{\mathbf{W}}) = \sum_{i=1}^{q} (s_i W_i^2 + 2W_i \beta_{1i}) + \sum_{i=q+1}^{d} 2W_i \beta_{1i} \equiv \sum_{i=1}^{d} \xi_i, \tag{41}$$

where $\xi_i = s_i W_i^2 I\{1 \le i \le q\} + 2W_i \beta_{1i}$, for $1 \le i \le d$. The right-hand side of (41) is a sum of independent variables, so we can apply the Edgeworth expansion to its distribution function, as described in detail below.

Note that $\mathbb{E}(W_i^2) = 1$, $\mathbb{E}(W_i^4) = 3$, $\mathbb{E}(W_i^6) = 15$ and $\mathbb{E}(W_i^{2j+1}) = 0$ for nonnegative integers j. By direct calculation,

$$\begin{aligned} \eta_1 &\equiv \sum_{i=1}^{d} \mathbb{E}(\xi_i) = \sum_{i=1}^{q} s_i = \operatorname{tr}(\mathbf{S}_1) = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_1), \\ \eta_2 &\equiv \sum_{i=1}^{d} \operatorname{var}(\xi_i) = \sum_{i=1}^{q} (2s_i^2 + 4\beta_{1i}^2) + \sum_{i=q+1}^{d} 4\beta_{1i}^2 = 2\operatorname{tr}(\mathbf{S}_1^2) + 4|\boldsymbol{\beta}_1|^2 + 4|\widetilde{\boldsymbol{\beta}}_1|^2 \\ &= 2\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_1 \boldsymbol{\Omega}\boldsymbol{\Sigma}_1) + 4(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_1 (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}), \\ \eta_3 &\equiv \sum_{i=1}^{d} \mathbb{E}\big[\xi_i - \mathbb{E}(\xi_i)\big]^3 = \sum_{i=1}^{d} (8s_i^3 + 24\beta_{1i}^2 s_i) \\ &= 8\operatorname{tr}(\mathbf{S}_1^3) + 24\boldsymbol{\beta}_1^\top \mathbf{S}_1 \boldsymbol{\beta}_1 = 8\operatorname{tr}\big[(\boldsymbol{\Omega}\boldsymbol{\Sigma}_1)^3\big] + 24(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_1 \boldsymbol{\Omega}\boldsymbol{\Sigma}_1 (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}). \end{aligned}$$

Notice that $\mathbb{E}(|\xi_i - \mathbb{E}(\xi_i)|^3) < \infty$, as $\max\{|s_i|, |\beta_{1i}|, 1 \le i \le d\} \le C_0$ by assumption. Using results from Chapter XVI of Feller (1966), we know

$$\begin{aligned} \mathbb{P}(2|1) &= \mathbb{P}\left(\sum_{i=1}^{d} \xi_i > c - c_1\right) \\ &= \mathbb{P}\left(\frac{\sum_{i=1}^{d} \xi_i - \mathbb{E}(\sum_{i=1}^{d} \xi_i)}{\sqrt{\sum_{i=1}^{d} \operatorname{var}(\xi_i)}} > \frac{c - c_1 - \mathbb{E}(\sum_{i=1}^{d} \xi_i)}{\sqrt{\sum_{i=1}^{d} \operatorname{var}(\xi_i)}}\right) \end{aligned}$$

$$= \bar{\Phi}\left(\frac{c - c_1 - \eta_1}{\sqrt{\eta_2}}\right) + \frac{\eta_3(1 - ((c_1 - c + \eta_1)^2/\eta_2))}{6\eta_2^{3/2}}\phi\left(\frac{c_1 - c + \eta_1}{\sqrt{\eta_2}}\right) + o\left(\frac{d}{\eta_2^{3/2}}\right),$$

where ϕ is the probability density function of the standard normal distribution. It is observed that $\eta_2 = L_1(\boldsymbol{\Omega}, \boldsymbol{\delta})$ and $c_1 + \eta_1 = M_1(\boldsymbol{\Omega}, \boldsymbol{\delta})$. Also, $c = tM_1(\boldsymbol{\Omega}, \boldsymbol{\delta}) + (1-t)M_2(\boldsymbol{\Omega}, \boldsymbol{\delta})$. As a result,

$$\frac{c - c_1 - \eta_1}{\sqrt{\eta_2}} = \frac{[tM_1 + (1-t)M_2] - M_1}{\sqrt{L_1}} = \frac{(1-t)(M_2 - M_1)}{\sqrt{L_1}} = (1-t)\frac{M}{\sqrt{L_1}}.$$

Plugging this into the expression of $\mathbb{P}(2|1)$, the first term is $\bar{\Phi}((1-t)\frac{M}{\sqrt{L_1}})$. Moreover, since the function $(1-u^2)\phi(u)$ is uniformly bounded, the second term is $O(\frac{\eta_3}{\eta_2^{3/2}})$. Here $\eta_2 = L_1$, and $\eta_3 = O(q)$ as s_i's and β_{1i}'s are abounded in magnitude. Combining the above gives

$$\mathbb{P}(2|1) = \bar{\Phi}\left(\frac{(1-t)M}{\sqrt{L_1}}\right) + \frac{O(q) + o(d)}{L_1^{3/2}}.$$

The proof is now complete.

SUPPLEMENTARY MATERIAL

Supplement to "QUADRO: A supervised dimension reduction method via Rayleigh quotient optimization" (DOI: 10.1214/14-AOS1307SUPP; .pdf). Owing to space constraints, numerical tables for simulation and some of the technical proofs are relegated to a supplementary document. It contains proofs of Propositions 2.1, 5.1 and 6.2.

REFERENCES

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BINDEA, G., MLECNIK, B., HACKL, H., CHAROENTONG, P., TOSOLINI, M., KIRILOVSKY, A., FRIDMAN, W.-H., PAGÈS, F., TRAJANOSKI, Z. and GALON, J. (2009). ClueGO: A cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25** 1091–1093.

CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865

COOK, R. D. and WEISBERG, S. (1991). Comment on "Sliced inverse regression for dimension reduction." *J. Amer. Statist. Assoc.* **86** 328–332.

COUDRET, R., LIQUET, B. and SARACCO, J. (2014). Comparison of sliced inverse regression approaches for underdetermined cases. *J. SFdS* **155** 72–96. MR3211755

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

FAN, J., FENG, Y. and TONG, X. (2012). A road to classification in high dimensional space. *J. Roy. Statist. Soc. B* **74** 745–771. MR2965958

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

FAN, J., KE, Z. T., LIU, H. and XIA, L. (2015). Supplement to "QUADRO: A supervised dimension reduction method via Rayleigh quotient optimization." DOI:10.1214/14-AOS1307SUPP.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications. Vol. II*. Wiley, New York.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.

FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84** 165–175. MR0999675

GUO, Y., HASTIE, T. and TIBSHIRANI, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics* **1** 1–18.

HAN, F. and LIU, H. (2012). Transelliptical component analysis. *Adv. Neural Inf. Process. Syst.* **25** 368–376.

HAN, F., ZHAO, T. and LIU, H. (2013). CODA: High dimensional copula discriminant analysis. *J. Mach. Learn. Res.* **14** 629–671. MR3033343

JIANG, B. and LIU, J. S. (2013). Sliced inverse regression with variable selection and interaction detection. Preprint. Available at arXiv:1304.4056.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

KENT, J. T. (1991). Discussion of Li (1991). *J. Amer. Statist. Assoc.* **86** 336–337.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, K.-C. (2000). High dimensional data analysis via the SIR/PHD approach. Lecture notes, Dept. Statistics, UCLA, Los Angeles, CA. Available at http://www.stat.ucla.edu/~kcli/sir-PHD.pdf.

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409

LI, L. and YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131. MR2422826

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

LUPARELLO, C. (2013). Aspects of collagen changes in breast cancer. *J. Carcinogene Mutagene S*13:007. DOI:10.4172/2157-2518.S13-007.

MARUYAMA, Y. and SEO, T. (2003). Estimation of moment parameter in elliptical distributions. *J. Japan Statist. Soc.* **33** 215–229. MR2039896

SHAO, J., WANG, Y., DENG, X. and WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39** 1241–1265. MR2816353

WEI, Z. and LI, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23** 1537–1544.

WITTEN, D. M. and TIBSHIRANI, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 753–772. MR2867457

WU, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *J. Comput. Graph. Statist.* **17** 590–610. MR2528238

ZHAO, T., ROEDER, K. and LIU, H. (2013). Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. Unpublished manuscript.

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZHONG, W., ZENG, P., MA, P., LIU, J. S. and ZHU, Y. (2005). RSIR: Regularized sliced inverse regression for motif discovery. *Bioinformatics* **21** 4169–4175.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

J. FAN
H. LIU
L. XIA
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu
hanliu@princeton.edu
lxia@princeton.edu

Z. KE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: zke@galton.uchicago.edu