QUADRO: A Supervised Dimension Reduction Method via Rayleigh Quotient Optimization

Jianqing Fan, Tracy Ke, Han Liu, and Lucy Xia

Princeton University

November 22, 2013

Abstract

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 nonpolynomially many parameters, even though 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. Such projection f has applications in many statistical problems for analyzing high-dimensional binary-labeled data, including:

- *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 for 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 $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^d$, and $Y \in \{0, 1\}$ is the label. The *Rayleigh quotient* of f is defined as

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

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

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

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 $\boldsymbol{X}$.

The Rayleigh quotient, as a new criterion for finding a projection f, serves well for different purposes of statistical analysis. First, for dimension reduction, it is a meaningful criterion which takes care of both variance explanation and label explanation. In contrast, many popular dimension reduction methods such as principal component analysis only consider variance explanation. Second, as we shall see in Section 6, a monotone transform of the Rayleigh quotient approximates the classification error. As a result, starting from an f with a large Rayleigh quotient, we can also construct a classification rule with a small classification error. In addition, Rayleigh quotient maximization is a convex optimization for quadratic discriminant functions. Third, with appropriate regularization, this criterion can select features that are different from those selected by many existing dimension reduction and classification methods. Thus, it is a new feature selection tool for statistical studies.

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 other settings, the two criteria can be very different.

(c) In a "relaxed" setting where two classes follow elliptical distributions (including multivariate normal as a special case) with possibly non-equal 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.

From these observations, we conclude that the Rayleigh quotient maximization is indeed a new criterion for both dimension reduction and feature selection. In addition, if we use it for classification, the Rayleigh quotient also serves as a surrogate of the classification error. In the remaining of this section, we show (a) and (b). We will discuss (c) 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 defined as $I(A) = 1$ if an event A happens and $I(A) = 0$ otherwise. For each given c, we define the classification error of h_c to be $\mathrm{err}(h_c) \equiv \mathbb{P}(h_c(\boldsymbol{X}) \neq Y)$. The classification error of f is then defined by

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

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., Guo et al. (2005); Fan and Fan (2008); Cai and Liu (2011); Shao et al. (2011); Fan et al. (2012); Han et al. (2013)), quadratic discriminant analysis (QDA), support vector machine (SVM), logistic regression, boosting, etc.

We now compare $\mathrm{Rq}(f)$ and $\mathrm{Err}(f)$. Denote by $\pi = \mathbb{P}(Y = 0)$, $\boldsymbol{\mu}_1 = \mathbb{E}(\boldsymbol{X}|Y = 0)$, $\boldsymbol{\Sigma}_1 = \mathrm{cov}(\boldsymbol{X}|Y = 0)$, $\boldsymbol{\mu}_2 = \mathbb{E}(\boldsymbol{X}|Y = 1)$ and $\boldsymbol{\Sigma}_2 = \mathrm{cov}(\boldsymbol{X}|Y = 1)$. 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 $\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$,

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

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, i.e., $d = 2$, 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

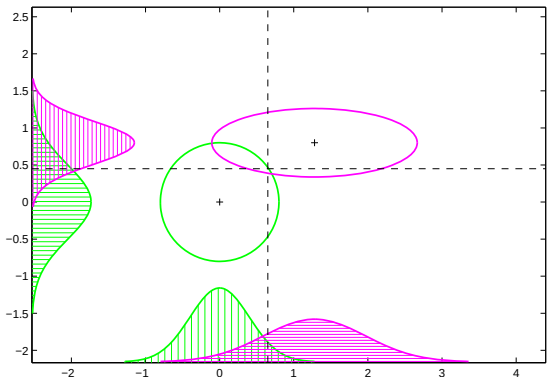


Figure 1: An example in $\mathbb{R}^2$. The green and purple represent class 1 and class 2 respectively. The ellipses are contours of distributions. Probability densities after being projected to X_1 and X_2 are also displayed. The dot lines correspond to optimal thresholds for classification using each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

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; 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.

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, i.e., $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{\Omega} \mathbf{x} - 2\boldsymbol{\delta}^\top \mathbf{x}$, 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 $d \gg n$).

Rayleigh quotient maximization can trace back to Fisher's linear discriminant analysis. However, 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 guess. 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., Fan and Fan (2008); Fan et al. (2012); Cai and Liu (2011)). 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, i.e., $\mathbf{a}_R = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, is no longer a solution to the Rayleigh quotient maximization. We propose a new method, named 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 $O_p\big(s\sqrt{\log(d)/n}\big)$, 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. 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 and real data examples. Section 8 concludes the paper. All proofs are relegated to Section 9.

Throughout this paper, 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 $\text{sign}(\cdot)$ be the sign function, where $\text{sign}(u) = 1$ when $u \ge 0$ and $\text{sign}(u) = -1$ 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

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

using (2), its Rayleigh quotient is

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

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 $\mathbb{E}(\boldsymbol{Z}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\boldsymbol{Z}) = \boldsymbol{\Sigma}$, by direct calculation,

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

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

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

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

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 $\boldsymbol{Z}$ that follows an elliptical distribution admits

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

where $\boldsymbol{U}$ 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 $\boldsymbol{U}$. 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 $\boldsymbol{Z}$.

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

$$\begin{aligned}\mathbb{E}[Q(\boldsymbol{Z})] &= \operatorname{tr}(\boldsymbol{\Omega\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\Omega\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{\delta},\\ \operatorname{var}[Q(\boldsymbol{Z})] &= 2(1+\gamma)\operatorname{tr}(\boldsymbol{\Omega\Sigma\Omega\Sigma}) + \gamma[\operatorname{tr}(\boldsymbol{\Omega\Sigma})]^2 + 4(\boldsymbol{\Omega\mu}-\boldsymbol{\delta})^\top\boldsymbol{\Sigma}(\boldsymbol{\Omega\mu}-\boldsymbol{\delta}),\end{aligned}$$

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

The proof is given in Section 9. The variance of $Q(\boldsymbol{Z})$ 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: $\boldsymbol{X}|(Y=0) \sim \mathcal{E}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, g_1)$ and $\boldsymbol{X}|(Y=1) \sim \mathcal{E}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, g_2)$. Without loss of generality, we assume the quantity γ is the same for both classes of conditional distributions. Let

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

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 $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ 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 $(\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2)$, we can solve the empirical version of (7) to obtain $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$, 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 $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ 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

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

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

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

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

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

Using zero as initial values, we iteratively update

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

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 $\boldsymbol{\Omega}$ and ν are fixed, we can write

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

where

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

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

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

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

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

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

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

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

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

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 $\max\{\rho|\mathbf{\Omega}^{(k)} - \mathbf{\Omega}^{(k-1)}|, \rho|\boldsymbol{\delta}^{(k)} - \boldsymbol{\delta}^{(k-1)}|, |\nu^{(k)} - \nu^{(k-1)}|/\rho\} \le \epsilon$ 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.

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

One possible estimator is the sample median

$$\widehat{\mu}_{Mj} = \text{median}\big(\{x_{1j}, \cdots, x_{nj}\}\big).$$

It can be shown that even under heavy-tailed distributions, $P\big(|\widehat{\mu}_{Mj} - \mu_j| > A\sqrt{\log(\delta^{-1})/n}\big) \le \delta$ 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. We denote the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ as $\sigma_1^2, \sigma_2^2, \cdots, \sigma_d^2$, and the off-diagonal elements as σ_{kj} for $k \neq j$. The estimator $\widehat{\boldsymbol{\mu}}_C = (\widehat{\mu}_{C,1}, \cdots, \widehat{\mu}_{C,d})^\top$ 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

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

where v is an upper bound of $\max\{\sigma_1^2, \cdots, \sigma_d^2\}$. 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 $P\big(|\widehat{\mu}_{Cj} - \mu_j| > \sqrt{\frac{2v\log(\delta^{-1})}{n(1 - 2\log(\delta^{-1})/n)}}\big) \le \delta$ 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 $h(y) = \log(1 + y + y^2/2)$ for $y \ge 0$, and $h(y) = -\log(1 - y + y^2/2)$ for $y < 0$ in practice. For the choice of v, any value larger than $\max\{\sigma_1^2, \cdots, \sigma_d^2\}$ would work in theory. Catoni (2012) introduced a Lepski's adaptation method to choose v. For simplicity, we take $v = 3\max\{\widetilde{\sigma}_1^2, \cdots, \widetilde{\sigma}_d^2\}$, 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 $\{x_{1j}^2, \cdots, x_{nj}^2\}$ and denote the estimator by $\widehat{\eta}_{Cj}$. This works as in our setting, $\mathbb{E}(X_j^4)$ is finite for each j; 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 ($\delta_0 < \min\{\sigma_1^2, \cdots, \sigma_d^2\}$). It is easy to see that when the fourth moments of X_j are uniformly upper bounded by a constant and $n \geq 4\log(d^2)$, $\max\{|\widehat{\sigma}_{Cj} - \sigma_j|, 1 \leq j \leq d\} = O_p(\sqrt{\log(d)/n})$.

Next, we consider estimating the correlation matrix $\mathbf{C}$. For this, we use the Kendall's tau correlation matrix proposed by Han and Liu (2012). The 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 $\widetilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$. 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 the Kendall's tau correlation coefficients using rank-based estimators:

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

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

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

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 that $\widehat{\mathbf{\Sigma}}$ 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:

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

Note that $|\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}|_\infty \leq |\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}|_\infty$ by definition. Therefore, $|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \leq |\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}|_\infty + |\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \leq 2|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty = O_p(\sqrt{\log(d)/n})$. 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 et al. (2013) provides more algorithmic details.

4.3 Estimation of kurtosis parameter

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

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

Motivated by this equality, we propose the estimator

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

where $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Omega}}$ are estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$, 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 need robust estimate to guarantee uniform convergence. In particular, we take $\widetilde{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_C$ and $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}_{\text{clime}}$ where $\widehat{\boldsymbol{\Omega}}_{\text{clime}}$ is the CLIME estimator proposed in Cai et al. (2011). We can also take the covariance estimator in Section 4.2, but 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 et al. (2011). Denote by $\boldsymbol{\Sigma}^{-1} = (\Omega_{jk})_{d\times d}$. From simple algebra,

$$|\widehat{\gamma} - \gamma| \le \max_{1\le j,k\le d} \left|\widetilde{\mu}_j \widetilde{\Omega}_{jk} \widetilde{\mu}_k - \mu_j \Omega_{jk} \mu_k\right| \le C \max\left\{|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}|_\infty,\ |\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1}|_\infty\right\}.$$

In Section 4.1, we have seen that $\|\widehat{\boldsymbol{\mu}}_C - \boldsymbol{\mu}\|_\infty = O_p(\sqrt{\log(d)/n})$. Moreover, Cai et al. (2011) showed that $|\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1}|_\infty = \|\boldsymbol{\Sigma}^{-1}\|_1 \cdot O_p(\sqrt{\log(d)/n})$ under mild conditions, where $\|\cdot\|_1$ is the matrix L_1-norm. Therefore, provided that $\|\boldsymbol{\Sigma}^{-1}\|_1 \le C$, 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 $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$. 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 $\lambda_1 \neq \lambda_2$. Moreover, we drop the symmetry constraint $\boldsymbol{\Omega} = \boldsymbol{\Omega}^\top$ 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 $(\boldsymbol{\Omega}, \boldsymbol{\delta})$ is equal to (up to a multiplicative constant)

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

The QUADRO estimates are

$$(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}}) = \underset{(\boldsymbol{\Omega},\boldsymbol{\delta}):\ \widehat{M}(\boldsymbol{\Omega},\boldsymbol{\delta})=1}{\operatorname{argmin}} \left\{\widehat{L}(\boldsymbol{\Omega}, \boldsymbol{\delta}) + \lambda|\boldsymbol{\Omega}|_1 + \lambda|\boldsymbol{\delta}|_1\right\}.$$

We shall compare the Rayleigh quotient of $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ 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 $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$. Here we adopt a class of solutions as the "oracle" instead of only $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$, because we want the results not tied to the sparsity assumption on $(\boldsymbol{\Omega}_0^*, \boldsymbol{\delta}_0^*)$ 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 et al. (2012), where they introduced a data-dependent intermediate solution to deal with such equality constraint. But 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

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

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:

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

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 9.2). 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 $S \subset \{1, 2, \cdots, d^2 + d\}$ and a nonnegative value $\bar{c}$, we define the *restricted eigenvalue*:

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

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 $S \subset \{1, 2, \cdots, d^2 + d\}$, there exist sets $U \subset \{1, \cdots, d\} \times \{1, \cdots, d\}$ and $V \subset \{1, \cdots, d\}$ 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

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

and it is easy to see that $U \subset U' \times U'$. The following result is proved in Section 9.

Proposition 5.1. *For any set $S \subset \{1, \cdots, d^2+d\}$, suppose U' and V are defined as above. Let $\widetilde{\boldsymbol{\Sigma}}_k$ be the submatrix of $\boldsymbol{\Sigma}_k$ 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 $\lambda_{\min}\big(\widetilde{\boldsymbol{\Sigma}}_k - v_1 \widetilde{\boldsymbol{\mu}}_k \widetilde{\boldsymbol{\mu}}_k^\top\big) \geq \frac{1}{2}\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_k) \geq \frac{v_2}{2}$ for $k = 1, 2$, then*

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

5.2 Oracle inequality on Rayleigh quotient

We shall assume that $\max\{|\boldsymbol{\Sigma}_k|_\infty, |\boldsymbol{\mu}_k|_\infty, k = 1, 2\} \leq 1$, and $|\widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k|_\infty \leq |\boldsymbol{\Sigma}_k|_\infty$, $|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty \leq |\boldsymbol{\mu}_k|_\infty$ for $k = 1, 2$, without loss of generality. For any $\lambda_0 \geq 0$, let $(\boldsymbol{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})$ be the associated oracle solution and S be the support of $\mathbf{x}^*_{\lambda_0} = [\mathrm{vec}(\boldsymbol{\Omega}^*_{\lambda_0})^\top, (\boldsymbol{\delta}^*_{\lambda_0})^\top]^\top$. Let $\Delta_n = \max\{|\widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k|_\infty, |\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty, k = 1, 2\}$. We have the following result for any given estimators, which proof can be found in Section 9.

Theorem 5.1. *Given $\lambda_0 \geq 0$, let S be the support of $\mathbf{x}^*_{\lambda_0}$, $s_0 = |S|$ and $k_0 = \max\{s_0, R(\boldsymbol{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})\}$. Suppose that $\Theta(S, 0) \geq c_0$, $\Theta(S, 3) \geq a_0$, and $R(\boldsymbol{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0}) \geq u_0$, for some positive constants a_0, c_0 and u_0. We assume $4s_0\Delta_n^2 \leq a_0 c_0$ 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 $\eta > 1$, by taking $\lambda = C\eta \max\{s_0^{1/2}\Delta_n, k_0^{1/2}\lambda_0\}[R(\boldsymbol{\Omega}^*_{\lambda_0}, \boldsymbol{\delta}^*_{\lambda_0})]^{-1/2}$,*

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

In Theorem 5.1, the rate of convergence has two parts. The term $s_0\Delta_n$ reflects how the stochastic errors of estimating $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ 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 compare with. In particular, if we compare $R(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}})$ with $R_{\max} \equiv R(\boldsymbol{\Omega}^*_0, \boldsymbol{\delta}^*_0)$, 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 Theorem 5.1 holds with $\lambda_0 = 0$, then for some positive constants A and C, when $\lambda > Cs_0^{1/2}R_{\max}^{-1/2}\Delta_n$, we have*

$$R(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\delta}}) \geq \big(1 - As_0\Delta_n\big)R_{\max}.$$

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

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

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, i.e. 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 $R_{\max}$.

6 Application to Classification

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

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

In this section, we first show that 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 nonGaussian distributions and non-equal 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

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

We reparametrize c as

$$c = tM_1(\boldsymbol{\Omega}, \boldsymbol{\delta}) + (1-t)M_2(\boldsymbol{\Omega}, \boldsymbol{\delta}). \tag{15}$$

Here $M_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \boldsymbol{\mu}_k^\top \boldsymbol{\Omega}\boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k^\top \boldsymbol{\delta} + \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k)$ is the mean of $Q(\boldsymbol{X})$ 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 $\boldsymbol{X}|(Y=0) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{X}|(Y=1) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. 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

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

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

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,*

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

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

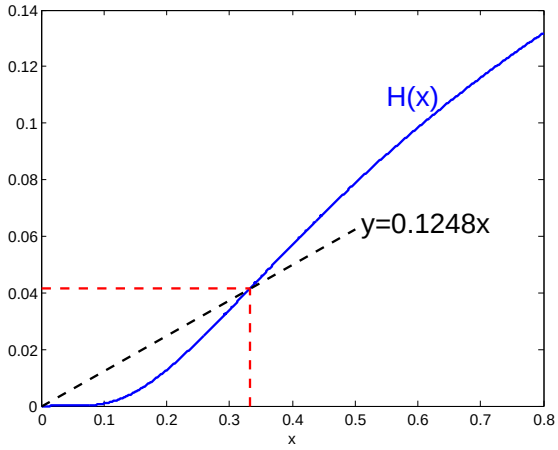


Figure 2: Function $H(x) = \bar{\Phi}(1/\sqrt{x})$.

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

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

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

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

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.

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

$$\begin{aligned} V_1 = V_1(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) &= \min\left\{(1-t)^2 R_1,\ \frac{1}{(1-t)^2 R_1}\right\}, \\ V_2 = V_2(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) &= \min\left\{t^2 R_2,\ \frac{1}{t^2 R_2}\right\}, \\ V = V(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) &= \max\{V_1/V_2,\ V_2/V_1\}. \end{aligned}$$

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

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

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

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

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

We remark that 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, i.e., $\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) \leq H\left(\frac{\pi R^{(t)}}{(1-t)^2}\right)$. For $x \in (0, 1/3)$, $H(x) \leq 0.1248x$. It follows that $\overline{\mathrm{Err}}(\boldsymbol{\Omega}, \boldsymbol{\delta}, t) \leq 0.1248 \cdot \frac{\pi R^{(t)}}{(1-t)^2}$.

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

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

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 both the Rayleigh quotient and classification error.

Table 1: Classification error and Rayleigh quotient for Gaussian examples. Means and standard deviations (in the parenthesis) of 100 replications are reported.

	Classification error			
	QUADRO	SLR	L-SLR	ROAD
Model 1	0.179	0.235	0.191	0.246
	(0.016)	(0.028)	(0.017)	(0.074)
Model 2	0.144	0.224	0.470	0.491
	(0.016)	(0.018)	(0.008)	(0.010)
Model 3	0.109	0.164	0.176	0.235
	(0.013)	(0.018)	(0.016)	(0.068)
	Rayleigh quotient			
	QUADRO	SLR	L-SLR	ROAD
Model 1	3.016	1.874	2.897	2.193
	(0.405)	(0.499)	(0.407)	(1.055)
Model 2	3.081	1.508	0	0
	(0.389)	(0.300)	(0)	(0)
Model 3	5.377	2.681	3.027	2.184
	(0.628)	(0.385)	(0.437)	(1.173)

7.1 Simulations under Gaussian models

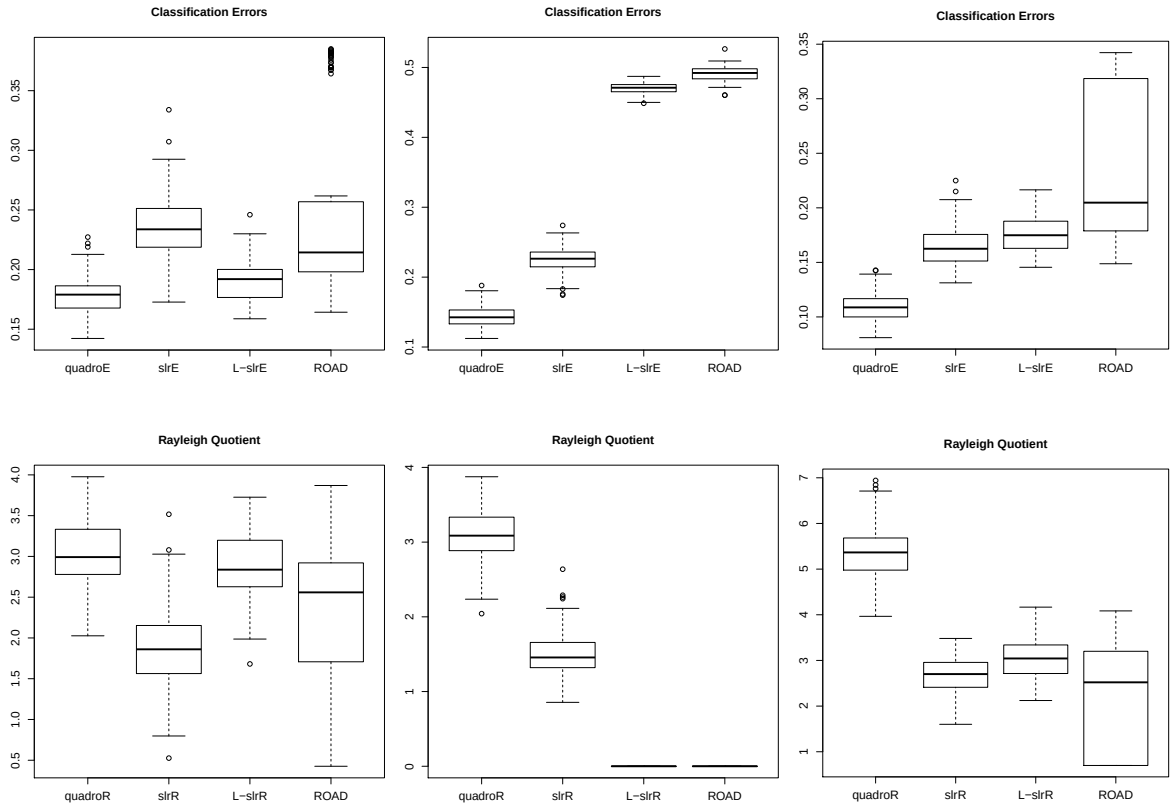
Let $n_1 = n_2 = 50$ and $d = 40$. For each given $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we generate 100 training datasets independently, each with n_1 data from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and n_2 data from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. 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 criteria (Rayleigh quotient or classification error) via $4,000$ newly generated testing data; we then choose the (λ_1, r) that optimize the criteria.

We compare QUADRO with three *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: ROAD (Fan et al., 2012) is a linear classification method for Gaussian distributed data. It is equivalent to a modified version of QUADRO by enforcing $\boldsymbol{\Omega}$ to be the zero matrix. Hence, ROAD only produces linear rules.

To make a fair comparison, the tuning parameters in SLR and L-SLR are selected in the same way as in QUADRO based on 4,000 testing data. ROAD is self-tuned by its package.

We consider three models:



(a) Model 1 (b) Model 2 (c) Model 3

Figure 3: Distributions of minimum classification errors (top panel) and maximum Rayleigh quotients (bottom panel) based on 100 replications for three different normal models. The tuning parameters for quadroE, slrE and L-slrE, are chosen to minimize the classification errors. They are also chosen to maximize Rayleigh quotients for quadroR, slrR, and L-slrR.

- *Model 1*: $\mathbf{\Sigma}_1$ is the identity matrix. $\mathbf{\Sigma}_2$ 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 $\boldsymbol{\mu}_2 = (0.7, \cdots, 0.7, 0, \cdots, 0)^\top$ with the first 10 elements of $\boldsymbol{\mu}_2$ being nonzero.
- *Model 2*: $\mathbf{\Sigma}_1$ 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. $\mathbf{\Sigma}_2 = (\mathbf{\Sigma}_1^{-1} + \mathbf{I})^{-1}$. We also set $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. In this model, neither $\mathbf{\Sigma}_1^{-1}$ nor $\mathbf{\Sigma}_2^{-1}$ is sparse, but $\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}$ is.
- *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*.

Table 1 reports the classification errors and Rayleigh quotients of all methods. Figures 3 contains the corresponding boxplots. In all three models, QUADRO outperforms other methods in terms of both classification error and Rayleigh quotient. In Model 2, since $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, no linear methods can do better in classification than the random guess. Therefore, both ROAD and L-SLR have poor performances in terms of classification error. In Models 1 and 3, $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. So, in the Bayes classification rule, both "linear" parts and "quadratic" parts play important roles. Comparing SLR and L-SLR, we see the former considers a broader class, while

Table 2: Classification error and Rayleigh quotient for elliptical examples. Means and standard deviations (in the parenthesis) of 100 replications are reported.

	Classification Error			
	QUADRO	QUADRO-0	SLR	L-SLR
Model 4	0.136 (0.015)	0.144 (0.015)	0.167 (0.019)	0.157 (0.017)
Model 5	0.161 (0.012)	0.173 (0.014)	0.184 (0.014)	0.184 (0.016)
Model 6	0.130 (0.017)	0.129 (0.016)	0.152 (0.018)	0.211 (0.018)
	Rayleigh quotient			
	QUADRO	QUADRO-0	SLR	L-SLR
Model 4	3.179 (0.413)	2.975 (0.410)	1.984 (0.511)	2.846 (0.450)
Model 5	2.415 (0.224)	2.191 (0.274)	1.625 (0.329)	2.166 (0.276)
Model 6	2.374 (0.194)	2.160 (0.213)	1.363 (0.246)	1.669 (0.223)

the latter is more robust, but neither of them perform uniformly well. However, QUADRO performs well in all cases.

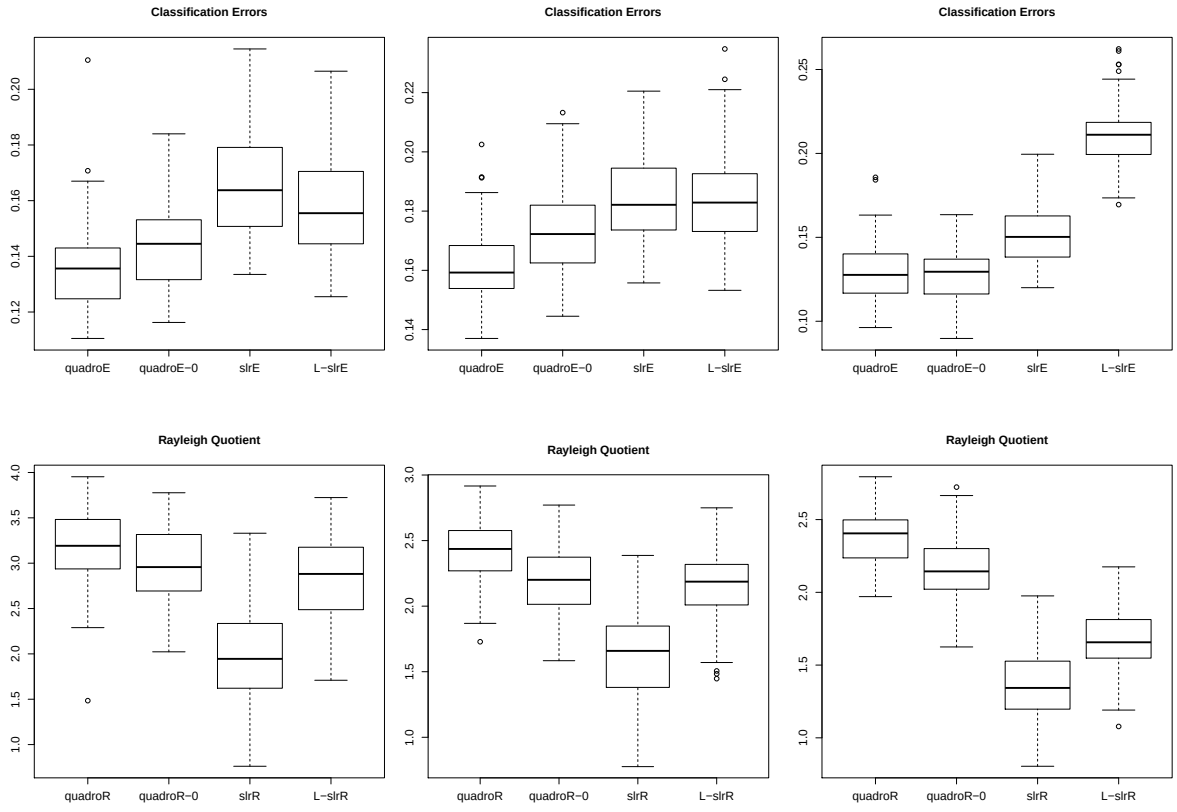
7.2 Simulations under elliptical models

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 that of SLR and L-SLR. ROAD is not designed for elliptical models, so we do not report its results here. 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:* The 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 a 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 $\Sigma_2(i,j) = 0.6^{|i-j|}$, i.e., $\boldsymbol{\Sigma}_2$ has an exponential off-diagonal decay.

Table 2 presents the average classification error and Rayleigh quotient over 100 replications. Figure 4 contains the boxplots. We see that in all models, QUADRO outperforms SLR and L-SLR. In addition, QUADRO is better than QUADRO-0, which illustrates the advantage of using the robust M-estimators for means and the rank-based estimators for covariance matrices.



(a) Model 4 (b) Model 5 (c) Model 6

Figure 4: Distributions of minimum classification errors (top panel) and maximum Rayleigh quotients (bottom panel) based on 100 replications across different elliptical distribution models. The tuning parameters for quadroE, slrE and L-slrE, are chosen to minimize the classification errors. They are also chosen to maximize Rayleigh quotients for quadroR, slrR, and L-slrR.

7.3 Real data analysis

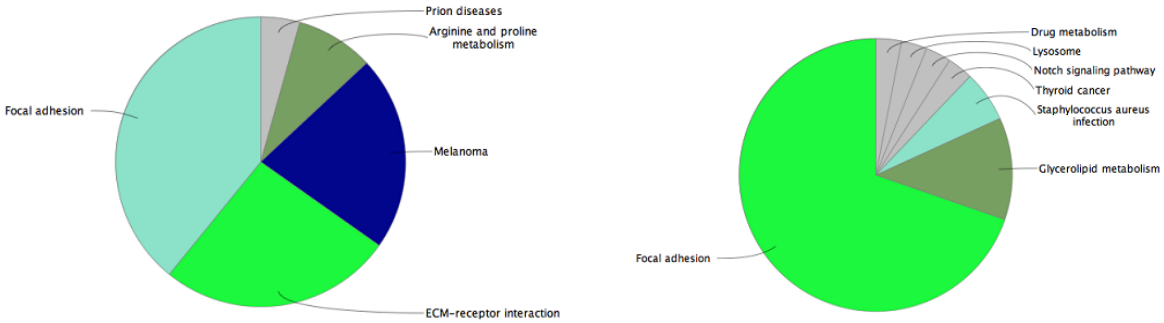
We apply QUADRO to a large-scale genomic dataset, GPL96, and compare the performance of QUADRO with SLR and L-SLR. The GPL96 data set contains 20,263 probes and 8,124 samples from 309 tissues. Among the tissues, breast tumor has 1,142 samples, which is the largest set. We merge the probes from the same gene by averaging them, and finally get 12,679 genes and 8,124 samples. We divide all samples into two groups: breast tumor or non-breast tumor.

First, we look at the classification errors. We replicate our experiment 100 times. Each time, with QUADRO, SLR and L-SLR we proceed with the following steps.

- Randomly choose a training set of 400 samples, 200 from breast tumor and 200 from non-breast 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.
- Use the rest 942 samples from breast tumor and another randomly chosen 942 samples from non-breast tumor as testing set, and calculate the testing error.

Table 3: Classification errors on GPL96 dataset, across methods QUADRO, SLR and L-SLR. Means and standard deviations (in the parenthesis) of 100 replications are reported.

Classification Errors		
QUADRO	SLR	L-SLR
0.014	0.025	0.025
(0.007)	(0.007)	(0.009)



(a) Quadro pathways (b) SLR pathways

Figure 5: Overall KEGG enrichment chart, using (a) QUADRO; (b) SLR.

The results are summarized in Table 3. We see that QUADRO outperforms SLR and L-SLR.

Next, we look at gene selection. 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 pathways that contain more than two genes; correspondingly, genes selected by SLR belong to 7 pathways. Using the tool ClueGo (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. The analysis is done by "DAVID Bioinformatics Resources" and the results are shown in Table 4. 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*.

Table 4: 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

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 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 of parameters even when the underlying distributions have the finite fourth moments. This allows us to establish oracle inequalities under relatively weaker conditions.

The Rayleigh optimization framework developed in this paper can also be extended to the multi-class case. Suppose the data are drawn independently from a joint distribution of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^d$ and Y takes values in $\{0, 1, \cdots, K-1\}$. The 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 $k = 0, 1, \cdots, K-1$. In this K-class situation,

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

Let $M_k(f) = \mathbb{E}[f(\boldsymbol{X})|Y=k]$ and $L_k(f) = \mathrm{var}[f(\boldsymbol{X})|Y=k]$. Similar to the two-class case,

maximizing $\mathrm{Rq}(f)$ is equivalent to solving the following optimization problem

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

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

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

To solve this problem, we first pick an order of $M_1(f), \cdots, M_K(f)$ 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 Proposition 2.1

We first present a lemma which is proved in Section A.

Lemma 9.1. *If $\boldsymbol{U}$ follows a uniform distribution on $\mathcal{S}^{d-1}$, for any $d\times d$ diagonal matrix $\mathbf{S}$ and any vector $\boldsymbol{\beta}\in\mathbb{R}^d$, we have*

- $\mathbb{E}(\boldsymbol{U}^\top\mathbf{S}\boldsymbol{U}) = \frac{\mathrm{tr}(\mathbf{S})}{d}$, $\mathbb{E}[(\boldsymbol{U}^\top\mathbf{S}\boldsymbol{U})^2] = \frac{2\,\mathrm{tr}(\mathbf{S}^2)+[\mathrm{tr}(\mathbf{S})]^2}{d^2+2d}$;
- $\mathbb{E}(\boldsymbol{U}^\top\boldsymbol{\beta}) = 0$, $\mathbb{E}[(\boldsymbol{U}^\top\boldsymbol{\beta})^2] = \frac{\|\boldsymbol{\beta}\|^2}{d}$;
- $\mathbb{E}(\boldsymbol{U}^\top\mathbf{S}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\beta}) = 0$.

Now, we show the claim of Proposition 2.1. Let $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Z}-\boldsymbol{\mu})$, then $\boldsymbol{Y} = \xi\boldsymbol{U}$ where $\boldsymbol{U}$ follows a uniform distribution on $\mathcal{S}^{d-1}$ and is independent of ξ. The quadratic form $Q(\boldsymbol{Z})$ can be rewritten as

$$\begin{aligned} Q(\boldsymbol{Z}) &= \boldsymbol{Z}^\top\boldsymbol{\Omega}\boldsymbol{Z} - 2\boldsymbol{\delta}^\top\boldsymbol{Z} \\ &= \boldsymbol{Y}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Y} + 2\boldsymbol{Y}^\top\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu}-\boldsymbol{\delta}) + \boldsymbol{\mu}^\top\boldsymbol{\Omega}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{\delta} \\ &= \bar{Q}(\boldsymbol{Y}) + c, \end{aligned}$$

where $c = \boldsymbol{\mu}^\top\boldsymbol{\Omega}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{\delta}$. Therefore, $\mathbb{E}[Q(\boldsymbol{Z})] = \mathbb{E}[\bar{Q}(\boldsymbol{Y})] + c$ and $\mathrm{var}[Q(\boldsymbol{Z})] = \mathrm{var}[\bar{Q}(\boldsymbol{Y})]$.

Furthermore, we let

$$\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{1/2} = \mathbf{K}\mathbf{S}\mathbf{K}^\top$$

be the eigenvalue decomposition of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{1/2}$, where $\mathbf{K}$ is an orthogonal matrix and $\mathbf{S}$ is a diagonal matrix. We also define

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

Notice that $\mathbf{Y}^\top \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Omega} \boldsymbol{\Sigma}^{1/2} \mathbf{Y} = \xi^2 \boldsymbol{U}^\top \mathbf{K} \mathbf{S} \mathbf{K}^\top \boldsymbol{U} = \xi^2 \boldsymbol{U}_1^\top \mathbf{S} \boldsymbol{U}_1$, where $\boldsymbol{U}_1 = \mathbf{K}^\top \boldsymbol{U}$. Since $\mathbf{K}$ is an orthogonal matrix, $\boldsymbol{U}_1$ follows the same distribution as $\boldsymbol{U}$ and is also independent of ξ. Moreover, we can write $\boldsymbol{Y}^\top \boldsymbol{\Sigma}^{1/2} (\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\delta}) = \xi \boldsymbol{U}_1^\top \boldsymbol{\beta}$. To save notation, we still use $\boldsymbol{U}$ to represent $\boldsymbol{U}_1$. It follows that

$$\bar{Q}(\boldsymbol{Y}) = \xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U} + 2\xi \boldsymbol{U}^\top \boldsymbol{\beta}.$$

Let's calculate $\mathbb{E}[\bar{Q}(\boldsymbol{Y})]$ first.

$$\begin{aligned}
\mathbb{E}[\bar{Q}(\boldsymbol{Y})] &= \mathbb{E}(\xi^2)\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) + 2\mathbb{E}(\xi)\mathbb{E}[\boldsymbol{U}^\top \boldsymbol{\beta}] \\
&= \frac{\mathbb{E}(\xi^2)}{d} \operatorname{tr}(\mathbf{S}) = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}).
\end{aligned}$$

The first equality is due to the fact that ξ and $\boldsymbol{U}$ are independent; the second equality is from Lemma 9.1; and the last inequality is because $\mathbb{E}(\xi^2) = d$ and $\operatorname{tr}(\mathbf{S}) = \operatorname{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{1/2}) = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma})$. It follows that

$$\mathbb{E}[Q(\boldsymbol{Z})] = \mathbb{E}[\bar{Q}(\boldsymbol{Y})] + c = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{\Omega} \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \boldsymbol{\delta}.$$

Next, we calculate $\operatorname{var}[\bar{Q}(\mathbf{Y})]$. It follows that

$$\begin{aligned}
\operatorname{var}[\bar{Q}(\mathbf{Y})] &= \operatorname{var}(\xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U} + 2\xi \boldsymbol{U}^\top \boldsymbol{\beta}) \\
&= \operatorname{var}(\xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) + 4\operatorname{var}(\xi \boldsymbol{U}^\top \boldsymbol{\beta}) + 4\operatorname{cov}(\xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}, \xi \boldsymbol{U}^\top \boldsymbol{\beta}).
\end{aligned}$$

Let's look at them term by term. First,

$$\begin{aligned}
\operatorname{var}(\xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) &= \mathbb{E}[\xi^4 (\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U})^2] - \mathbb{E}^2(\xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) \\
&= \mathbb{E}(\xi^4)\mathbb{E}[(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U})^2] - \mathbb{E}^2(\xi^2)\mathbb{E}^2(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) \\
&= \mathbb{E}(\xi^4) \frac{2\operatorname{tr}(\mathbf{S}^2) + \operatorname{tr}^2(\mathbf{S})}{2d + d^2} - \mathbb{E}^2(\xi^2) \frac{\operatorname{tr}^2(\mathbf{S})}{d^2} \\
&= 2(\gamma + 1)\operatorname{tr}(\mathbf{S}^2) + \gamma \operatorname{tr}^2(\mathbf{S}) \\
&= 2(\gamma + 1)\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma}) + \gamma[\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma})]^2.
\end{aligned}$$

The third equality comes from Lemma 9.1; the last equality follows from the fact that $\operatorname{tr}(\mathbf{S}^2) = \operatorname{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{1/2}) = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma})$. Second,

$$\begin{aligned}
\operatorname{var}(\xi \boldsymbol{U}^\top \boldsymbol{\beta}) &= \mathbb{E}(\xi^2 (\boldsymbol{U}^\top \boldsymbol{\beta})^2) - \mathbb{E}^2(\xi \boldsymbol{U}^\top \boldsymbol{\beta}) \\
&= \mathbb{E}(\xi^2)\mathbb{E}[(\boldsymbol{U}^\top \boldsymbol{\beta})^2] - \mathbb{E}^2(\xi)\mathbb{E}^2(\boldsymbol{U}^\top \boldsymbol{\beta}) \\
&= \mathbb{E}(\xi^2) \frac{\|\boldsymbol{\beta}\|^2}{d} = (\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\delta})^\top \boldsymbol{\Sigma} (\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\delta}).
\end{aligned}$$

In the last equality, we have used $\mathbb{E}(\xi^2) = d$, $\boldsymbol{\beta} = \mathbf{K}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\delta})$ and $\mathbf{K}^\top \mathbf{K} = \mathbf{I}_d$. Last,

$$\begin{aligned}
\operatorname{cov}(\xi^2 \boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}, \xi \boldsymbol{U}^\top \boldsymbol{\beta}) &= \mathbb{E}(\xi^3)\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U} \boldsymbol{U}^\top \boldsymbol{\beta}) - \mathbb{E}(\xi^2)\mathbb{E}(\mathbf{U}^\top \mathbf{S} \mathbf{U})\mathbb{E}(\xi \mathbf{U}^\top \boldsymbol{\beta}) \\
&= \mathbb{E}(\xi^3)\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U} \boldsymbol{U}^\top \boldsymbol{\beta}) = 0
\end{aligned}$$

Combining the above gives

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

□

9.2 Proof of Proposition 5.1

For any $d \times 1$ vector $\mathbf{v}$ and $d \times d$ matrix $\mathbf{A}$, we denote by $\mathrm{Supp}(\mathbf{v})$ the support of $\mathbf{v}$, which is contained in $\{1, \cdots, d\}$, and by $\mathrm{Supp}(\mathbf{A})$ the support of $\mathbf{A}$, which is contained in $\{1, \cdots, d\} \times \{1, \cdots, d\}$. Let $\theta = \sqrt{1 + v_1(1+\gamma)/2} > 1$ and

$$c = (1+\kappa) \min\big\{(1+\gamma){v_2}^2,\ 4(1-1/\theta^2)v_2\big\}.$$

The claim then becomes $\Theta(S, 0) \geq c$, or in other words,

$$\mathbf{x}^\top \mathbf{Q}\mathbf{x} \geq c|\mathbf{x}|^2, \qquad \text{when } \mathrm{Supp}(\mathbf{x}) \subset S.$$

First, using (13) and Lemma 9.2, we find that for each $\mathbf{x}$, there exits unique $(\boldsymbol{\Omega}, \boldsymbol{\delta})$ such that $\mathbf{x} = \mathbf{x}(\boldsymbol{\Omega}, \boldsymbol{\delta})$ and $\mathbf{x}^\top \mathbf{Q}\mathbf{x} = L(\boldsymbol{\Omega}, \boldsymbol{\delta})$. Second, by definition of U' and V, $\mathrm{Supp}(\mathbf{x}) \subset S$ implies that $\mathrm{Supp}(\boldsymbol{\Omega}) \subset U' \times U'$ and $\mathrm{Supp}(\boldsymbol{\delta}) \subset V$. Therefore, it suffices to show

$$L(\boldsymbol{\Omega}, \boldsymbol{\delta}) \geq c(|\boldsymbol{\Omega}|^2 + |\boldsymbol{\delta}|^2), \qquad \text{when } \mathrm{Supp}(\boldsymbol{\Omega}) \subset U' \times U' \text{ and } \mathrm{Supp}(\boldsymbol{\delta}) \subset V. \tag{20}$$

Now, we show (20). From (5) and that $\gamma \geq 0$,

$$L_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) \geq 2(1+\gamma)\,\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}\boldsymbol{\Sigma}_k) + 4(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_k(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta}), \qquad k = 1, 2.$$

Let $\widetilde{\boldsymbol{\Omega}}$ be the submatrix of $\boldsymbol{\Omega}$ by restricting rows and columns to the set $U' \cup V$, and $\widetilde{\boldsymbol{\delta}}$ be the subvector of $\boldsymbol{\delta}$ by restricting the elements to the set $U' \cup V$. It is easy to see that when $\mathrm{Supp}(\boldsymbol{\Omega}) \subset U' \times U'$ and $\mathrm{Supp}(\boldsymbol{\delta}) \subset V$,

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}\boldsymbol{\Sigma}_k) = \mathrm{tr}(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Sigma}}_k\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Sigma}}_k),\\
&(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_k(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta}) = (\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\delta}})^\top \widetilde{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\delta}}),
\end{aligned}$$

where we recall that $\widetilde{\boldsymbol{\Sigma}}_k$ is the submatrix of $\boldsymbol{\Sigma}_k$ by restrcting rows and columns to the set $U' \cup V$, $\widetilde{\boldsymbol{\mu}}_k$ to the set $U' \cup V$. It follows that

$$\begin{aligned}
& L_k(\boldsymbol{\Omega}, \boldsymbol{\delta})\\
\geq\ & 2(1+\gamma)\,\mathrm{tr}(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Sigma}}_k\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Sigma}}_k) + 4(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\delta}})^\top \widetilde{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\delta}})\\
=\ & 2(1+\gamma)\,\mathrm{tr}(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Sigma}}_k\widetilde{\boldsymbol{\Omega}}(\widetilde{\boldsymbol{\Sigma}}_k - v_1\widetilde{\boldsymbol{\mu}}_k\widetilde{\boldsymbol{\mu}}_k^\top)) + 4(1-1/\theta^2)\widetilde{\boldsymbol{\delta}}^\top \widetilde{\boldsymbol{\Sigma}}_k\widetilde{\boldsymbol{\delta}}\\
& +4(\theta\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}}_k - \theta^{-1}\widetilde{\boldsymbol{\delta}})^\top \widetilde{\boldsymbol{\Sigma}}_k(\theta\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\mu}}_k - \theta^{-1}\widetilde{\boldsymbol{\delta}})\\
\geq\ & 2(1+\gamma)\,\mathrm{tr}(\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Sigma}}_k\widetilde{\boldsymbol{\Omega}}(\widetilde{\boldsymbol{\Sigma}}_k - v_1\widetilde{\boldsymbol{\mu}}_k\widetilde{\boldsymbol{\mu}}_k^\top)) + 4(1-1/\theta^2)\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_k)|\widetilde{\boldsymbol{\delta}}|^2.
\end{aligned} \tag{21}$$

Denote by I_1 the first term in (21). We aim to derive a lower bound for I_1. It is well known that $\mathrm{tr}(\mathbf{A}^\top\mathbf{B}\mathbf{C}\mathbf{D}^\top) = \mathrm{vec}(\mathbf{A})^\top(\mathbf{D} \otimes \mathbf{B})\,\mathrm{vec}(\mathbf{C})$, where $\mathrm{vec}(\mathbf{A})$ be the vectorization of $\mathbf{A}$ by stacking all the columns, $\mathbf{D} \otimes \mathbf{B}$ is the Kronecker product of $\mathbf{D}$ and $\mathbf{B}$. Using this formula and that $\boldsymbol{\Sigma}_k$ is symmetric, we find that

$$\begin{aligned}
I_1 &= 2(1+\gamma)\,\mathrm{vec}(\widetilde{\boldsymbol{\Omega}})^\top[(\widetilde{\boldsymbol{\Sigma}}_k - v_1\widetilde{\boldsymbol{\mu}}_k\widetilde{\boldsymbol{\mu}}_k^\top) \otimes \widetilde{\boldsymbol{\Sigma}}_k]\,\mathrm{vec}(\widetilde{\boldsymbol{\Omega}})\\
&\geq 2(1+\gamma)|\widetilde{\boldsymbol{\Omega}}|^2\lambda_{\min}((\widetilde{\boldsymbol{\Sigma}}_k - v_1\widetilde{\boldsymbol{\mu}}_k\widetilde{\boldsymbol{\mu}}_k^\top) \otimes \widetilde{\boldsymbol{\Sigma}}_k)\\
&\geq (1+\gamma)\lambda_{\min}^2(\widetilde{\boldsymbol{\Sigma}}_k)|\widetilde{\boldsymbol{\Omega}}|^2.
\end{aligned}$$

The last inequality is from the property that $\lambda_{\min}(\mathbf{A}\otimes\mathbf{B}) = \lambda_{\min}(\mathbf{A})\lambda_{\min}(\mathbf{B})$ when $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite, and also the assumption that $\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_k - v_1\widetilde{\boldsymbol{\mu}}_k\widetilde{\boldsymbol{\mu}}_k) \geq \frac{1}{2}\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_k)$. Plugging I_1 into (21), we have

$$\begin{aligned} & L_1(\mathbf{\Omega},\boldsymbol{\delta}) + \kappa L_2(\mathbf{\Omega},\boldsymbol{\delta}) \\ \geq\ & (1+\gamma)[\lambda^2_{\min}(\widetilde{\mathbf{\Sigma}}_1) + \kappa\lambda^2_{\min}(\widetilde{\mathbf{\Sigma}}_2)]|\widetilde{\mathbf{\Omega}}|^2 + 4(1-1/\theta^2)[\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_1) + \kappa\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_2)]|\widetilde{\boldsymbol{\delta}}|^2 \\ \geq\ & (1+\gamma)(1+\kappa){v_2}^2|\widetilde{\mathbf{\Omega}}|^2 + 4(1-1/\theta^2)(1+\kappa)v_2|\widetilde{\boldsymbol{\delta}}|^2 \\ \geq\ & c(|\widetilde{\mathbf{\Omega}}|^2 + |\widetilde{\boldsymbol{\delta}}|^2) = c(|\mathbf{\Omega}|^2 + |\boldsymbol{\delta}|^2). \end{aligned}$$

This proves (20). □

9.3 Proof of Theorem 5.1

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

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

We introduce the following lemma which is proved in Section A.

Lemma 9.2. *$M(\mathbf{\Omega},\boldsymbol{\delta}) = \mathbf{q}^\top\mathbf{x}$ and $L(\mathbf{\Omega},\boldsymbol{\delta}) = \mathbf{x}^\top\mathbf{Q}\mathbf{x}$.*

Let $\mathbf{x}^*_{\lambda_0} = [\mathrm{vec}(\mathbf{\Omega}^*_{\lambda_0})^\top, (\boldsymbol{\delta}^*_{\lambda_0})^\top]^\top$ and $\widehat{\mathbf{x}} = [\mathrm{vec}(\widehat{\mathbf{\Omega}})^\top, \widehat{\boldsymbol{\delta}}^\top]^\top$. Using Lemma 9.2,

$$\begin{aligned} \mathbf{x}^*_{\lambda_0} &= \min_{\mathbf{x}\in\mathbb{R}^d:\mathbf{q}^\top\mathbf{x}=1}\{\mathbf{x}^\top\mathbf{Q}\mathbf{x} + \lambda_0|\mathbf{x}|_1\}, \\ \widehat{\mathbf{x}} &= \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d:\widehat{\mathbf{q}}^\top\mathbf{x}=1}\{\mathbf{x}^\top\widehat{\mathbf{Q}}\mathbf{x} + \lambda|\mathbf{x}|_1\}, \end{aligned}$$

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

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

In addition, we have the following lemma, which is proved in Section A.

Lemma 9.3. *$\max\{|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty, |\widehat{\mathbf{q}} - \mathbf{q}|_\infty\} \leq C_0\max\{|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty, |\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty, k = 1,2\}$ 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 $\lambda = C\eta\max\{s_0^{1/2}\Delta_n,\ k_0^{1/2}\lambda_0\}[R(\mathbf{x}^*_{\lambda_0})]^{-1/2}$,*

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

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 Cauchy-Schwartz inequality,

$$|\mathbf{x}^*|_1 \leq \sqrt{|\mathbf{x}^*|_0}\ |\mathbf{x}^*| \leq c_0^{-1/2}\sqrt{|\mathbf{x}^*|_0}\bar{V}^* \tag{22}$$

It follows that

$$|\widehat{\mathbf{q}}^\top \mathbf{x}^* - \mathbf{q}^\top \mathbf{x}^*| \leq |\widehat{\mathbf{q}} - \mathbf{q}|_\infty |\mathbf{x}^*|_1 \leq c_0^{-1/2}\Delta_n\sqrt{|\mathbf{x}^*|_0}\bar{V}^* = c_0^{-1/2}\alpha_n\bar{V}^*. \tag{23}$$

Let $t_n = \widehat{\mathbf{q}}^\top \mathbf{x}^*$. Then (23) 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 $\widehat{\mathbf{x}}$,

$$\widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} + \lambda|\widehat{\mathbf{x}}|_1 \leq (\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* + \lambda|\mathbf{x}_1^*|_1. \tag{24}$$

By direct calculation,

$$\begin{aligned}
\widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* &= (\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) + 2(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* \\
&= (\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) + 2(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top (\widehat{\mathbf{Q}}\mathbf{x}_1^* - V^*\widehat{\mathbf{q}}) \\
&\geq 2(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top (\widehat{\mathbf{Q}}\mathbf{x}_1^* - V^*\widehat{\mathbf{q}})
\end{aligned} \tag{25}$$

where the second equality is because $\widehat{\mathbf{q}}^\top \widehat{\mathbf{x}} = \widehat{\mathbf{q}}^\top \mathbf{x}_1^* = 1$. We aim to bound $|\widehat{\mathbf{Q}}\mathbf{x}_1^* - V^*\widehat{\mathbf{q}}|_\infty$. The following lemma is proved in Section A.

Lemma 9.4. *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 $t_n^{-1} < 2$,

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

Here the third inequality follows from (22)-(23) and $V^* = \bar{V}^*(R^*)^{-1/2} \leq u_0^{-1/2}\bar{V}^*$. The last inequality is obtained as follows: from Lemma 9.3, 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 9.4 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 (25) gives

$$\widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* \geq -\frac{\lambda}{2}|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1. \tag{26}$$

Combining (24) and (26) gives

$$\lambda|\widehat{\mathbf{x}}|_1 - \frac{\lambda}{2}|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1 \leq \lambda|\mathbf{x}_1^*|_1. \tag{27}$$

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 (27) that

$$|(\widehat{\mathbf{x}} - \mathbf{x}_1^*)_{S^c}|_1 \leq 3|(\widehat{\mathbf{x}} - \mathbf{x}_1^*)_S|_1. \tag{28}$$

Second, note that $|\widehat{\mathbf{x}} - \mathbf{x}_1^*|_1 \leq |\widehat{\mathbf{x}}|_1 + |\mathbf{x}_1^*|_1$. Plugging it into (27) gives

$$|\widehat{\mathbf{x}}|_1 \leq 3|\mathbf{x}_1^*|_1 = 3t_n^{-1}|\mathbf{x}^*|_1 \leq 6c_0^{-1/2}\sqrt{|\mathbf{x}^*|_0}\ \bar{V}^*. \tag{29}$$

Step 2: We use (28)-(29) 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

$$\begin{aligned}
& \widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* \\
\geq\ & \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - \big(|\widehat{\mathbf{x}}^\top \widehat{\mathbf{Q}}\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}}| + |(\mathbf{x}_1^*)^\top \widehat{\mathbf{Q}}\mathbf{x}_1^* - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^*|\big) \\
\geq\ & \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - \big(|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty |\widehat{\mathbf{x}}|_1^2 + |\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty |\mathbf{x}_1^*|_1^2\big) \\
\geq\ & \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - 10t_n^{-2}|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty |\mathbf{x}^*|_1^2 \\
\geq\ & \widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* - C_3\beta_n V^*,
\end{aligned} \tag{30}$$

where the last two inequalities are direct results of (29). Combining (24) and (30),

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

Similarly to (25), we have

$$\widehat{\mathbf{x}}^\top \mathbf{Q}\widehat{\mathbf{x}} - (\mathbf{x}_1^*)^\top \mathbf{Q}\mathbf{x}_1^* = (\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top \mathbf{Q}(\widehat{\mathbf{x}} - \mathbf{x}_1^*) + 2(\widehat{\mathbf{x}} - \mathbf{x}_1^*)^\top (\mathbf{Q}\mathbf{x}_1^* - V^*\widehat{\mathbf{q}}), \tag{32}$$

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 it into (31), we obtain

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

We can rewrite the second and third terms on the left hand side of (33) 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 it into (33), 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

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

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

$$(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)\leq\frac{3\lambda}{2}\sqrt{|\mathbf{x}_1^*|_0}\,|\widehat{\mathbf{x}}_{1S}-\mathbf{x}_{1S}^*|+C_3\beta_nV^*. \tag{34}$$

Since (28) holds, by 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 to (34),

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

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

$$(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)\leq\frac{9\lambda^2}{4a_0}|\mathbf{x}^*|_0+2C_3\beta_nV^*, \tag{35}$$

where we have used $|\mathbf{x}_1^*|_0=|\mathbf{x}^*|_0$. Furthermore, (32) yields that

$$\begin{aligned}\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}-(\mathbf{x}_1^*)^\top\mathbf{Q}\mathbf{x}_1^*&\leq(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)+\frac{\lambda}{2}|\widehat{\mathbf{x}}-\mathbf{x}_1^*|_1\\&\leq(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)+2\lambda|\mathbf{x}_1^*|_1\\&\leq(\widehat{\mathbf{x}}-\mathbf{x}_1^*)^\top\mathbf{Q}(\widehat{\mathbf{x}}-\mathbf{x}_1^*)+4c_0^{-1/2}\bar{V}^*\lambda\sqrt{|\mathbf{x}^*|_0}.\end{aligned} \tag{36}$$

where the second inequality is because $|\widehat{\mathbf{x}}-\mathbf{x}_1^*|_1\leq|\widehat{\mathbf{x}}|_1+|\mathbf{x}_1^*|\leq 4|\mathbf{x}_1^*|_1$, and the last inequality is from (29). Recall that $\lambda=C\eta\max\{k_0^{1/2}\lambda_0,s_0^{1/2}\Delta_n\}\bar{V}^*$. As a result,

$$\lambda\sqrt{|\mathbf{x}^*|_0}=C\eta\max\{k_0^{1/2}s_0^{1/2}\lambda_0,s_0\Delta_n\}\bar{V}^*=C\eta T_n(\mathbf{x}^*)\bar{V}^*. \tag{37}$$

Combining (35), (36) and (37) gives

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

$$\leq C_4\eta^2 T_n(\mathbf{x}^*)V^*. \tag{38}$$

Step 3: We use (38) 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 (23) 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 it to (38) 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

$$\begin{aligned}\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}} &\leq \big[t_n^{-2} + C_4\eta^2 T_n(\mathbf{x}^*)\big](\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^* \\ &\leq \big[1 + 6c_0^{-1/2}\alpha_n\bar{V}^* + C_4\eta^2 T_n(\mathbf{x}^*)\big](\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^* \\ &\leq [1 + C_5\eta^2 T_n(\mathbf{x}^*)](\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*. \end{aligned} \tag{39}$$

Second, we look at the numerator $\mathbf{q}^\top\widehat{\mathbf{x}}$. Since $\widehat{\mathbf{q}}^\top\widehat{\mathbf{x}} = 1$, by (29),

$$|\mathbf{q}^\top\widehat{\mathbf{x}} - 1| \leq |\widehat{\mathbf{q}} - \mathbf{q}|_\infty|\widehat{\mathbf{x}}|_1 \leq 6c_0^{-1/2}\alpha_n\bar{V}^* \leq C_6 T_n(\mathbf{x}^*). \tag{40}$$

Combining (39) and (40) gives

$$\begin{aligned} R(\widehat{\mathbf{x}}) = \frac{(\mathbf{q}^\top\widehat{\mathbf{x}})^2}{\widehat{\mathbf{x}}^\top\mathbf{Q}\widehat{\mathbf{x}}} &\geq \frac{\big[1 - C_6 T_n(\mathbf{x}^*)\big]^2}{1 + C_5\eta^2 T_n(\mathbf{x}^*)}\frac{1}{(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*} \\ &\geq \big[1 - A\eta^2 T_n(\mathbf{x}^*)\big]\frac{(\mathbf{q}^\top\mathbf{x}^*)^2}{(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^*} \\ &= \big[1 - A\eta^2 T_n(\mathbf{x}^*)\big]R(\mathbf{x}^*), \end{aligned} \tag{41}$$

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

9.4 Proof of Proposition 6.1

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

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

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

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

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

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

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

$$Q(\boldsymbol{Z}) = (\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{Y} + \boldsymbol{\mu}_1)^\top\boldsymbol{\Omega}(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{Y} + \boldsymbol{\mu}_1) - 2(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{Y} + \boldsymbol{\mu}_1)^\top\boldsymbol{\delta}$$

$$= \boldsymbol{Y}^\top \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Omega} \boldsymbol{\Sigma}_1^{1/2} \mathbf{Y} + 2\boldsymbol{Y}^\top \boldsymbol{\Sigma}_1^{1/2} (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}) + \boldsymbol{\mu}_1^\top \boldsymbol{\Omega} \boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_1^\top \boldsymbol{\delta}. \tag{42}$$

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

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

We recall that $\boldsymbol{\beta}_1 = \mathbf{K}_1^\top \boldsymbol{\Sigma}_1^{1/2} (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\beta})$. Let $\widetilde{\boldsymbol{\beta}}_1 = \widetilde{\mathbf{K}}_1^\top \boldsymbol{\Sigma}_1^{1/2} (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})$, $\boldsymbol{W} = \mathbf{K}_1^\top \boldsymbol{Y}$, $\widetilde{\boldsymbol{W}} = \widetilde{\mathbf{K}}_1^\top \boldsymbol{Y}$ and $c_1 = \boldsymbol{\mu}_1^\top \boldsymbol{\Omega} \boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_1^\top \boldsymbol{\delta}$. It follows from (42) that

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

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,

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

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

$$\bar{Q}_1(\boldsymbol{W}) + \bar{F}_1(\widetilde{\boldsymbol{W}}) = \sum_{i=1}^{q} (s_i W_i^2 + 2W_i \beta_{1i}) + \sum_{i=q+1}^{d} 2W_i \beta_{1i} \equiv \sum_{i=1}^{d} \xi_i, \tag{43}$$

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 (43) 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,

$$\begin{aligned} \eta_1 &\equiv \sum_{i=1}^{d} \mathbb{E}(\xi_i) = \sum_{i=1}^{q} s_i = \operatorname{tr}(\mathbf{S}_1) = \operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_1), \\ \eta_2 &\equiv \sum_{i=1}^{d} \operatorname{var}(\xi_i) = \sum_{i=1}^{q} (2s_i^2 + 4\beta_{1i}^2) + \sum_{i=q+1}^{d} 4\beta_{1i}^2 \\ &= 2\operatorname{tr}(\mathbf{S}_1^2) + 4|\boldsymbol{\beta}_1|^2 + 4|\widetilde{\boldsymbol{\beta}}_1|^2 \\ &= 2\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_1 \boldsymbol{\Omega}\boldsymbol{\Sigma}_1) + 4(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_1 (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}), \\ \eta_3 &\equiv \sum_{i=1}^{d} \mathbb{E}[\xi_i - \mathbb{E}(\xi_i)]^3 = \sum_{i=1}^{d} (8s_i^3 + 24\beta_{1i}^2 s_i) \\ &= 8\operatorname{tr}(\mathbf{S}_1^3) + 24\boldsymbol{\beta}_1^\top \mathbf{S}_1 \boldsymbol{\beta}_1 \\ &= 8\operatorname{tr}[(\boldsymbol{\Omega}\boldsymbol{\Sigma}_1)^3] + 24(\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_1 \boldsymbol{\Omega}\boldsymbol{\Sigma}_1 (\boldsymbol{\Omega}\boldsymbol{\mu}_1 - \boldsymbol{\delta}). \end{aligned} \tag{44}$$

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. Therefore, using results in Chapter XVI of Feller (1966), we know

$$\begin{aligned}
\mathbb{P}(2|1) &= \mathbb{P}\big(\sum_{i=1}^{d} \xi_i > c - c_1\big) \\
&= \mathbb{P}\left(\frac{\sum_{i=1}^d \xi_i - \mathbb{E}(\sum_{i=1}^d \xi_i)}{\sqrt{\sum_{i=1}^d \mathrm{var}(\xi_i)}} > \frac{c - c_1 - \mathbb{E}(\sum_{i=1}^d \xi_i)}{\sqrt{\sum_{i=1}^d \mathrm{var}(\xi_i)}}\right) \\
&= \bar{\Phi}\left(\frac{c - c_1 - \eta_1}{\sqrt{\eta_2}}\right) + \frac{\eta_3(1 - (\frac{(c_1 - c + \eta_1)^2}{\eta_2}))}{6\eta_2^{3/2}} \phi\left(\frac{c_1 - c + \eta_1}{\sqrt{\eta_2}}\right) + o\left(\frac{d}{\eta_2^{3/2}}\right), \qquad (45)
\end{aligned}$$

where ϕ is the probability density function of the standard normal distribution. It is observed that $\eta_2 = L_1(\mathbf{\Omega}, \boldsymbol{\delta})$ and $c_1 + \eta_1 = M_1(\mathbf{\Omega}, \boldsymbol{\delta})$. Also, $c = tM_1(\mathbf{\Omega}, \boldsymbol{\delta}) + (1-t)M_2(\mathbf{\Omega}, \boldsymbol{\delta})$. 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 it into (45), 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. □

9.5 Proof of Proposition 6.2

Given $(\mathbf{\Omega}, \boldsymbol{\delta}, t)$, recall that $R_k = R_k(\mathbf{\Omega}, \boldsymbol{\delta})$, for $k = 1, 2$. Let $x_1 = \big[(1-t)^2 R_1\big]^{-1}$, $x_2 = \big[t^2 R_2\big]^{-1}$, and $x = \pi x_1 + (1-\pi)x_2$. By direct calculation,

$$\overline{\mathrm{Err}}(\mathbf{\Omega}, \boldsymbol{\delta}, t) = \pi H(x_1) + (1-\pi)H(x_2), \qquad H\left(\frac{\pi}{(1-t)^2 R_{\kappa(t)}}\right) = H(x). \qquad (46)$$

Since H is twice continuously differentiable, from the Taylor expansion,

$$\begin{aligned}
H(x_1) &= H(x) + H'(x)(x_1 - x) + \frac{1}{2}H''(z_1)(x_1 - x)^2, \\
H(x_2) &= H(x) + H'(x)(x_2 - x) + \frac{1}{2}H''(z_2)(x_2 - x)^2,
\end{aligned}$$

where z_1 is a number between x_1 and x, and z_2 is a number between x_2 and x. Noticing that $\pi(x_1 - x) + (1-\pi)(x_2 - x) = 0$, we further obtain

$$\begin{aligned}
\pi H(x_1) + (1-\pi)H(x_2) &= H(x) + \frac{\pi}{2}H''(z_1)(x_1 - x)^2 + \frac{1-\pi}{2}H''(z_2)(x_2 - x)^2 \\
&= H(x) + \frac{\pi(1-\pi)}{2}[(1-\pi)H''(z_1) + \pi H''(z_2)](x_1 - x_2)^2.
\end{aligned}$$

Here, the second equality is because $x_1 - x = (1-\pi)(x_1 - x_2)$ and $x_2 - x = \pi(x_2 - x_1)$. Let $A = \sup_{z\in[\min\{x_1,x_2\},\max\{x_1,x_2\}]} |H''(z)|$. It follows that

$$\big|\pi H(x_1) + (1-\pi)H(x_2) - H(x)\big| \le \frac{\pi(1-\pi)}{2} \cdot A(x_1 - x_2)^2. \tag{47}$$

Now, we bound $A(x_1 - x_2)^2$. Write for short $x_{\min} = \min\{x_1, x_2\}$ and $x_{\max} = \max\{x_1, x_2\}$. It is easy to see that

$$A(x_1 - x_2)^2 \le \sup_{z\in[x_{\min},x_{\max}]} |z^2 H''(z)| \cdot \left(\frac{x_1 - x_2}{x_{\min}}\right)^2. \tag{48}$$

By direct calculation, the function $z^2 H''(z)$ has the expression

$$z^2 H''(z) = \frac{1}{4\sqrt{2\pi}} \left(\frac{1}{z} - 3\right) e^{-\frac{1}{2z}} \frac{1}{z^{1/2}}.$$

On one hand, as $z \to \infty$, $|z^2 H''(z)| \le C z^{-1/2}$; on the other hand, as $z \to 0$, $|z^2 H''(z)| \le C e^{-1/(2z)} z^{-3/2} \le C z^{1/2}$. Therefore, we have $|z^2 H''(z)| \le C \min\{\sqrt{z}, 1/\sqrt{z}\}$. Plugging it into (48), we have

$$A(x_1 - x_2)^2 \le C\big[\max\big\{x_1 \wedge (1/x_1),\ x_2 \wedge (1/x_2)\big\}\big]^{1/2} \cdot \left(\frac{x_1 - x_2}{x_{\min}}\right)^2.$$

Note that $x_1 \wedge (1/x_1) = V_1$, $x_2 \wedge (1/x_2) = V_2$ and $|x_1 - x_2|/x_{\min} = |V - 1|$. It follows that

$$A(x_1 - x_2)^2 \le C\big[\max\{V_1, V_2\}\big]^{1/2} \cdot |V - 1|^2. \tag{49}$$

We combine (46), (47) and (49), and note that $\pi(1-\pi) \le 1/4$. The first claim then follows.

When $t = 1/2$, we find $V - 1 = \Delta R / R_0$. The second claim follows immediately. □

A Supplementary Proofs

A.1 Proof of Lemma 9.1

Consider $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d)$. Then $\boldsymbol{Y} \overset{(d)}{=} R\boldsymbol{U}$, where $R^2 \sim \chi^2_d$ and it is independent of $\boldsymbol{U}$. Since $\mathbb{E}(R^2) = d$ and $\mathrm{var}(R^2) = 2d$, it follows that

$$\begin{aligned}
\mathbb{E}(\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y}) &= \mathbb{E}(R^2)\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) = d\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}),\\
\mathbb{E}[(\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y})^2] &= \mathbb{E}(R^4)\mathbb{E}[(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U})^2] = (d^2 + 2d)\mathbb{E}[(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U})^2],\\
\mathbb{E}(\boldsymbol{Y}^\top \boldsymbol{\beta}) &= \mathbb{E}(R)\mathbb{E}(\boldsymbol{U}^\top \boldsymbol{\beta}),\\
\mathbb{E}[(\boldsymbol{Y}^\top \boldsymbol{\beta})^2] &= \mathbb{E}(R^2)\mathbb{E}[(\boldsymbol{U}^\top \boldsymbol{\beta})^2] = d\mathbb{E}[(\boldsymbol{U}^\top \boldsymbol{\beta})^2],\\
\mathbb{E}(\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y} \boldsymbol{Y}^\top \boldsymbol{\beta}) &= \mathbb{E}(R^3)\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U} \boldsymbol{U}^\top \boldsymbol{\beta}).
\end{aligned}$$

First, note that $\boldsymbol{Y}^\top \boldsymbol{\beta} \sim N(0, \|\boldsymbol{\beta}\|^2)$. So $\mathbb{E}(\boldsymbol{Y}^\top \boldsymbol{\beta}) = 0$ and $\mathbb{E}[(\boldsymbol{Y}^\top \boldsymbol{\beta})^2] = \|\boldsymbol{\beta}\|^2$. We immediately have

$$\mathbb{E}(\boldsymbol{U}^\top \boldsymbol{\beta}) = 0, \qquad \mathbb{E}[(\boldsymbol{U}^\top \boldsymbol{\beta})^2] = \|\boldsymbol{\beta}\|^2/d.$$

Second, write $\mathbf{S} = \mathrm{diag}(s_1, \cdots, s_d)$ and $\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y} = \sum_{i=1}^d s_i Y_i^2$, where $Y_i \overset{iid}{\sim} N(0,1)$. Therefore,

$$\mathbb{E}(\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y}) = \sum_{i=1}^d s_i \mathbb{E}(Y_i^2) = \sum_{i=1}^d s_i = \mathrm{tr}(\mathbf{S}),$$
$$\mathrm{var}(\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y}) = \sum_{i=1}^d s_i^2 \mathrm{var}(Y_i^2) = \sum_{i=1}^d 2 s_i^2 = 2\,\mathrm{tr}(\mathbf{S}^2).$$

We immediately have

$$\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U}) = \mathrm{tr}(\mathbf{S})/d, \qquad \mathbb{E}[(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U})^2] = \frac{[\mathrm{tr}(\mathbf{S})]^2 + 2\,\mathrm{tr}(\mathbf{S}^2)}{d^2 + 2d}.$$

Last, note that

$$\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y} \boldsymbol{Y}^\top \boldsymbol{\beta} = (\sum_{i=1}^d s_i Y_i^2)(\sum_{j=1}^d \beta_j Y_j) = \sum_{i=1}^d s_i \beta_i Y_i^3 + \sum_{i \neq j} s_i \beta_j Y_i^2 Y_j.$$

So $\mathbb{E}(\boldsymbol{Y}^\top \mathbf{S} \boldsymbol{Y} \boldsymbol{Y}^\top \boldsymbol{\beta}) = 0$. Since $\mathbb{E}(R^3) \neq 0$, we immediately have $\mathbb{E}(\boldsymbol{U}^\top \mathbf{S} \boldsymbol{U} \boldsymbol{U}^\top \boldsymbol{\beta}) = 0$. □

A.2 Proof of Lemma 9.2

Recall that

$$\begin{aligned} M(\boldsymbol{\Omega}, \boldsymbol{\delta}) &= -\boldsymbol{\mu}_1^\top \boldsymbol{\Omega} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^\top \boldsymbol{\Omega} \boldsymbol{\mu}_2 + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\delta} - \mathrm{tr}(\boldsymbol{\Omega}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)) \\ &= \mathrm{tr}\left[\boldsymbol{\Omega}(\boldsymbol{\Sigma}_2 + \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top - \boldsymbol{\Sigma}_1 - \boldsymbol{\mu}_1 \boldsymbol{\mu}_1^\top)\right] + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\delta}. \end{aligned}$$

It is well known that for any matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathrm{tr}(\mathbf{A}^\top \mathbf{B}) = \mathrm{vec}(\mathbf{A})^\top \mathrm{vec}(\mathbf{B})$. So we have

$$\begin{aligned} M(\boldsymbol{\Omega}, \boldsymbol{\delta}) &= \mathrm{vec}(\boldsymbol{\Omega})^\top \mathrm{vec}(\boldsymbol{\Sigma}_2 + \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top - \boldsymbol{\Sigma}_1 - \boldsymbol{\mu}_1 \boldsymbol{\mu}_1^\top) + 2\boldsymbol{\delta}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \mathbf{x}^\top \mathbf{q}. \end{aligned}$$

Moreover, for $k = 1, 2$,

$$\begin{aligned} L_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) &= 2(1+\gamma)\,\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}\boldsymbol{\Sigma}_k) + \gamma[\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k)]^2 + 4(\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}_k (\boldsymbol{\Omega}\boldsymbol{\mu}_k - \boldsymbol{\delta}) \\ &= 2(1+\gamma)\,\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}\boldsymbol{\Sigma}_k) + \gamma[\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k)]^2 + 4\boldsymbol{\mu}_k^\top \boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}\boldsymbol{\mu}_k - 8\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\Omega}\boldsymbol{\mu}_k + 4\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\delta} \\ &= 2\,\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Sigma}_k\boldsymbol{\Omega}((1+\gamma)\boldsymbol{\Sigma}_k + 2\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top)\right] - 8\,\mathrm{tr}(\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\Omega}\boldsymbol{\mu}_k) + 4\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\delta} + \gamma[\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_k)]^2. \end{aligned}$$

From linear algebra, $\mathrm{tr}(\mathbf{A}^\top \mathbf{B} \mathbf{C} \mathbf{D}^\top) = \mathrm{vec}(\mathbf{A})^\top (\mathbf{D} \otimes \mathbf{B}) \mathrm{vec}(\mathbf{C})$. It follows that

$$\begin{aligned} L_k(\boldsymbol{\Omega}, \boldsymbol{\delta}) &= 2\,\mathrm{vec}(\boldsymbol{\Omega})^\top \left[((1+\gamma)\boldsymbol{\Sigma}_k + 2\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top) \otimes \boldsymbol{\Sigma}_k\right] \mathrm{vec}(\boldsymbol{\Omega}) - 8\boldsymbol{\delta}^\top (\boldsymbol{\mu}_k^\top \otimes \boldsymbol{\Sigma}_k) \mathrm{vec}(\boldsymbol{\Omega}) \\ &\quad + 4\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\delta} + \gamma\,\mathrm{vec}(\boldsymbol{\Omega})^\top \mathrm{vec}(\boldsymbol{\Sigma}_k)\,\mathrm{vec}(\boldsymbol{\Sigma}_k)^\top \mathrm{vec}(\boldsymbol{\Omega}) \\ &= 2\begin{bmatrix}\mathrm{vec}(\boldsymbol{\Omega})^\top & \boldsymbol{\delta}^\top\end{bmatrix} \begin{bmatrix}((1+\gamma)\boldsymbol{\Sigma}_k + 2\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top) \otimes \boldsymbol{\Sigma}_k & -2(\boldsymbol{\mu}_k \otimes \boldsymbol{\Sigma}_k) \\ -2(\boldsymbol{\mu}_k^\top \otimes \boldsymbol{\Sigma}_k) & 2\boldsymbol{\Sigma}_k\end{bmatrix} \begin{bmatrix}\mathrm{vec}(\boldsymbol{\Omega}) \\ \boldsymbol{\delta}\end{bmatrix} \\ &\quad + \gamma\,\mathrm{vec}(\boldsymbol{\Omega})^\top \mathrm{vec}(\boldsymbol{\Sigma}_k)\,\mathrm{vec}(\boldsymbol{\Sigma}_k)^\top \mathrm{vec}(\boldsymbol{\Omega}) \\ &= \mathbf{x}^\top \mathbf{Q}_k \mathbf{x}. \end{aligned}$$

Note that $\mathbf{Q} = \mathbf{Q}_1 + \kappa \mathbf{Q}_2$ and $L = L_1 + \kappa L_2$. This immediately implies $L(\boldsymbol{\Omega}, \boldsymbol{\delta}) = \mathbf{x}^\top \mathbf{Q} \mathbf{x}$. □

A.3 Proof of Lemma 9.3

Let $\Delta_n = \max\{|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty, |\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty, k = 1, 2\}$ to save notation. We recall that $|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty \le |\boldsymbol{\mu}_k|_\infty \le 1$ and $|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty \le |\mathbf{\Sigma}_k|_\infty$, for $k = 1, 2$, as assumed in the beginning of Section 5.2.

Consider $|\widehat{\mathbf{q}} - \mathbf{q}|_\infty$ first. Note that $|\operatorname{vec}(\mathbf{A}) - \operatorname{vec}(\mathbf{B})|_\infty = |\mathbf{A} - \mathbf{B}|_\infty$ for any matrices $\mathbf{A}$ and $\mathbf{B}$. It follows that

$$\begin{aligned} |\widehat{\mathbf{q}} - \mathbf{q}|_\infty &\le \sum_{k=1}^{2} \big(|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty + |\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_k^\top - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top|_\infty + 2|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k|_\infty\big) \\ &\le C\Delta_n + \sum_{k=1}^{2} |\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_k^\top - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top|_\infty. \end{aligned}$$

Write $\boldsymbol{\mu}_k = \boldsymbol{\mu}$ for short. We have $|\widehat{\mu}(i)\widehat{\mu}(j) - \mu(i)\mu(j)| \le |\widehat{\mu}(j)||\widehat{\mu}(i) - \mu(i)| + |\mu(i)||\widehat{\mu}(j) - \mu(j)| \le (|\widehat{\boldsymbol{\mu}}|_\infty + |\boldsymbol{\mu}|_\infty)|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_\infty$. Since $|\widehat{\boldsymbol{\mu}}|_\infty \le |\boldsymbol{\mu}|_\infty + |\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_\infty \le 2|\boldsymbol{\mu}|_\infty \le 2$, it follows that $|\widehat{\mu}(i)\widehat{\mu}(j) - \mu(i)\mu(j)| \le 3|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_\infty \le 3\Delta_n$. As a result,

$$|\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_k^\top - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top|_\infty \le 3\Delta_n, \qquad k = 1, 2. \tag{50}$$

We immediately have $|\widehat{\mathbf{q}} - \mathbf{q}|_\infty \le C\Delta_n$.

Next, consider $|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty$. It is easy to see that

$$\begin{aligned} |\widehat{\mathbf{Q}}_k - \mathbf{Q}_k|_\infty \le\ & 2(1+\gamma)|\widehat{\mathbf{\Sigma}}_k \otimes \widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k \otimes \mathbf{\Sigma}_k|_\infty + 4|(\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_k^\top) \otimes \widehat{\mathbf{\Sigma}}_k - (\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top) \otimes \mathbf{\Sigma}_k|_\infty \\ &+ \gamma|\operatorname{vec}(\widehat{\mathbf{\Sigma}}_k)\operatorname{vec}(\widehat{\mathbf{\Sigma}}_k)^\top - \operatorname{vec}(\mathbf{\Sigma}_k)\operatorname{vec}(\mathbf{\Sigma}_k)^\top|_\infty \\ &+ 4|\widehat{\boldsymbol{\mu}}_k \otimes \widehat{\mathbf{\Sigma}}_k - \boldsymbol{\mu}_k \otimes \mathbf{\Sigma}_k|_\infty + 4|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty. \end{aligned}$$

Here $|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty \le \Delta_n$, and using a similar argument as in (50), we can show that $|\operatorname{vec}(\widehat{\mathbf{\Sigma}}_k)\operatorname{vec}(\widehat{\mathbf{\Sigma}}_k)^\top - \operatorname{vec}(\mathbf{\Sigma}_k)\operatorname{vec}(\mathbf{\Sigma}_k)^\top|_\infty \le C|\operatorname{vec}(\widehat{\mathbf{\Sigma}}_k) - \operatorname{vec}(\mathbf{\Sigma}_k)|_\infty \le C\Delta_n$. To bound the other terms, it suffices to show that

$$|\mathbf{A} \otimes \mathbf{B} - \mathbf{A}' \otimes \mathbf{B}'|_\infty \le C(|\mathbf{A} - \mathbf{A}'|_\infty + |\mathbf{B} - \mathbf{B}'|_\infty), \tag{51}$$

when $|\mathbf{A} - \mathbf{A}'|_\infty \le |\mathbf{A}|_\infty \le C$ and $|\mathbf{B} - \mathbf{B}'|_\infty \le |\mathbf{B}|_\infty \le C$. To see this, note that $|A(i,j)B(k,l) - A'(i,j)B'(k,l)| \le |B'(k,l)||A(i,j) - A'(i,j)| + |A(i,j)||B(k,l) - B'(k,l)| \le (|\mathbf{A}|_\infty + |\mathbf{B}'|_\infty)(|\mathbf{A} - \mathbf{A}'|_\infty + |\mathbf{B} - \mathbf{B}'|_\infty) \le C(|\mathbf{A} - \mathbf{A}'|_\infty + |\mathbf{B} - \mathbf{B}'|_\infty)$. This proves (51). It follows immediately that $|\widehat{\mathbf{Q}} - \mathbf{Q}|_\infty \le C\Delta_n$. □

A.4 Proof of Lemma 9.4

Write $\mathbf{x}^* = \mathbf{x}^*_{\lambda_0}$ for short. From KKT conditions, there exists a dual variable θ such that

$$\mathbf{q}^\top\mathbf{x}^* = 1, \qquad 2\mathbf{Q}\mathbf{x}^* + \lambda_0\widetilde{\operatorname{sign}}(\mathbf{x}^*) = \theta\mathbf{q}.$$

Here $\widetilde{\operatorname{sign}}(\mathbf{x}^*)$ is the vector whose j-th coordinate is -1, 1 or any value between $[-1, 1]$ when $x_j^* < 0$, $x_j^* > 0$ and $x_j^* = 0$, respectively. We multiply both sides of the second equation by $(\mathbf{x}^*)^\top$, and note that $(\mathbf{x}^*)^\top\widetilde{\operatorname{sign}}(\mathbf{x}^*) = |\mathbf{x}^*|_1$. It follows that $\theta = 2(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^* + \lambda_0|\mathbf{x}^*|_1$ and

$$\begin{aligned} \mathbf{Q}\mathbf{x}^* &= \Big[(\mathbf{x}^*)^\top\mathbf{Q}\mathbf{x}^* + \frac{\lambda_0}{2}|\mathbf{x}^*|_1\Big]\mathbf{q} - \frac{\lambda_0}{2}\widetilde{\operatorname{sign}}(\mathbf{x}^*) \\ &= V^*\mathbf{q} + \frac{\lambda_0}{2}\big[|\mathbf{x}^*|_1\mathbf{q} - \widetilde{\operatorname{sign}}(\mathbf{x}^*)\big], \end{aligned} \tag{52}$$

where $V^* = (\mathbf{x}^*)^\top \mathbf{Q}\mathbf{x}^* = [R(\mathbf{x}^*)]^{-1}$. Let $\bar{V}^* = (V^*)^{1/2}$. Then $\Gamma(\mathbf{x}) = \frac{1}{\bar{V}^*}|\mathbf{Q}\mathbf{x}^* - V^*\mathbf{q}|_\infty$, and (52) implies

$$\Gamma(\mathbf{x}^*) \leq \frac{\lambda_0}{2\bar{V}^*}(|\mathbf{x}^*|_1|\mathbf{q}|_\infty + 1) = \frac{|\mathbf{q}|_\infty}{2\bar{V}^*}\lambda_0|\mathbf{x}^*|_1 + \frac{\lambda_0}{2}[R(\mathbf{x}^*)]^{1/2}. \tag{53}$$

From the Cauchy-Schwartz inequality,

$$|\mathbf{x}^*|_1 \leq |\mathbf{x}^*|_0^{1/2}|\mathbf{x}^*| \leq |\mathbf{x}^*|_0^{1/2}\frac{1}{\sqrt{c_0}}\sqrt{(\mathbf{x}^*)^\top \mathbf{Q}\mathbf{x}^*} = \frac{\bar{V}^*}{\sqrt{c_0}}|\mathbf{x}^*|_0^{1/2},$$

where we have used the fact that $(\mathbf{x}^*)^\top \mathbf{Q}\mathbf{x}^* \geq \lambda_{\min}(\mathbf{Q}_{SS})|\mathbf{x}^*|^2 \geq c_0|\mathbf{x}^*|^2$. Plugging it into (53) gives

$$\begin{aligned}\Gamma(\mathbf{x}^*) &\leq \frac{|\mathbf{q}|_\infty}{2\sqrt{c_0}}\lambda_0|\mathbf{x}^*|_0^{1/2} + \frac{\lambda_0}{2}[R(\mathbf{x}^*)]^{1/2} \\ &\leq C_1\lambda_0\big[\max\{|\mathbf{x}^*|_0, R(\mathbf{x}^*)\}\big]^{1/2},\end{aligned}$$

where C_1 is a constant that only depends on c_0 (noting that $|\mathbf{q}|_\infty \leq C\max\{|\boldsymbol{\mu}_k|_\infty, |\boldsymbol{\Sigma}_k|_\infty, k = 1, 2\} \leq C$).

References

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.

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained L1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the Americal Statistical Assocition*, 106, 594–607.

Catoni, O. (2012), "Challenging the empirical mean and empirical variance: a deviation study," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Institut Henri Poincaré, vol. 48, pp. 1148–1185.

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

Fan, J., Feng, Y., and Tong, X. (2012), "A ROAD to classification in high dimensional space," *J. Roy. Statist. Soc. B*, 74.

Feller, W. (1966), *An Introduction to Probability Theory and Its Applications*, vol. II, John Wiley and Sons, Inc. New York, London.

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," *Advances in Neural Information Processing Systems*, 25, 368–376.

Han, F., Zhao, T., and Liu, H. (2013), "CODA: High Dimensional Copula Discriminant Analysis," *Journal of Machine Learning Research*, 14, 629–671.

Kendall, M. (1938), "A New Measure Of Rank Correlation," *Biometrika*, 30, 81–93.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-dimensional Semiparametric Gaussian Copula Graphical Models," *The Annals of Statistics*, 40, 2293–2326.

Luparello, C. (2013), "Aspects of Collagen Changes in Breast Cancer," *J Carcinogene Mutagene S*, 13.

Maruyama, Y. and Seo, T. (2003), "Estimation of moment parameter in elliptical distributions," *Journal of Japanese Statistical Society*, 33, 215–229.

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.

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

Zhao, T., Roeder, K., and Liu, H. (2013), "Positive Semidefinite Rank-based Correlation Matrix Estimation with Application to Semiparametric Graph Estimation," *manuscript*.