Sparse quadratic classification rules via linear dimension reduction

Irina Gaynanova[a,*], Tianying Wang[a]

[a] *Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843*

Abstract

We consider the problem of high-dimensional classification between the two groups with unequal covariance matrices. Rather than estimating the full quadratic discriminant rule, we propose to perform simultaneous variable selection and linear dimension reduction on original data, with the subsequent application of quadratic discriminant analysis on the reduced space. In contrast to quadratic discriminant analysis, the proposed framework doesn't require estimation of precision matrices and scales linearly with the number of measurements, making it especially attractive for the use on high-dimensional datasets. We support the methodology with theoretical guarantees on variable selection consistency, and empirical comparison with competing approaches. We apply the method to gene expression data of breast cancer patients, and confirm the crucial importance of ESR1 gene in differentiating estrogen receptor status.

Keywords: Convex optimization, Discriminant analysis, High-dimensional statistics, Variable selection.

1. Introduction

We consider a binary classification problem: given n independent pairs $(X_1, Y_1), ..., (X_n, Y_n)$ from a random pair (X, Y) on $\mathbb{R}^p \times \{1, 2\}$, our goal is to both learn a rule that will assign one of two labels to a new data point $X \in \mathbb{R}^p$, and determine the subset of p variables that influences the rule. One of the popular classification tools is linear discriminant analysis, or LDA; see Chapter 11 in [36]. While it gives unsatisfactory results when applied to high-dimensional datasets [12], recent work suggests that additional regularization, variable selection in particular, leads to dramatic performance improvements. Earlier approaches perform variable selection and regularize the sample covariance matrix by treating it as diagonal [48, 53]. More recent methods directly estimate the discriminant directions by using convex optimization framework with sparsity-inducing penalties [5, 15, 35].

Despite these significant advances, a key underlying assumption of linear discriminant analysis is the equality of covariance matrices between the groups, viz. $\Sigma_1 = \Sigma_2$. This assumption is unlikely to be satisfied in practice, leading to suboptimal performance of the linear rule. When the measurements are normally distributed, viz. $X_i|Y_i = g \sim \mathcal{N}(\mu_g, \Sigma_g)$, $g \in \{1, 2\}$, with $\Sigma_1 \neq \Sigma_2$, the Bayes rule is quadratic, leading to quadratic discriminant analysis, or QDA. As with the linear case, quadratic discriminant analysis (QDA) performs poorly when p is large. This unsatisfactory performance is largely due to the estimation of precision matrices Σ_1^{-1} and Σ_2^{-1}, a task that is extremely challenging when $p \gg n$. In fact, even when $p = n/2$ and the assumption of equal covariance matrices is violated, the misclassification error rate of sample QDA is worse than the rates of regularized linear discriminant methods; see the supplement in [15].

Several extensions of sample QDA have been proposed. A common strategy is to jointly estimate Σ_1^{-1} and Σ_2^{-1}. Friedman [13], Ramey et al. [44] regularize sample covariance matrices by shrinkage. Wu et al. [55] impose equicorrelation structure on each covariance matrix by pooling both the diagonal and off-diagonal elements. Danaher et al. [11], Guo et al. [19], Price et al. [42], Simon and Tibshirani [46] use a penalized likelihood technique, where the penalty enforces similarity either between the covariance matrices Σ_g or the precision matrices Σ_g^{-1}. While these methods perform better than quadratic rules based on sample covariance matrices, they again rely on estimating two precision matrices. As such, additional assumptions on Σ_g^{-1} such as sparsity are usually enforced, and the estimation procedure scales quadratically with the number of measurements p. Moreover, the resulting classification rules still rely on all p variables, and therefore cannot be used for both classification and variable selection.

*Corresponding author

Email addresses: irinag@stat.tamu.edu (Irina Gaynanova), tianying@stat.tamu.edu (Tianying Wang)

Li and Shao [31] address the variable selection problem by enforcing sparsity in both the covariance matrices and the vector of mean differences via thresholding. The method comes with strong theoretical guarantees on classification consistency and promising empirical performance. Nevertheless, it again requires additional assumptions on Σ_g, and is computationally prohibitive for large p due to required matrices inversion together with a 3-dimensional search over tuning parameter values.

In summary, a significant progress in linear discriminant methods made it possible to apply them to large datasets and perform variable selection. In practice, however, the covariance matrices are often unequal, but the existing quadratic methods typically cannot perform variable selection, and are computationally prohibitive for large p. In this work we bridge the gap between the linear and the quadratic methods by developing a new classification rule that takes into account unequal covariance matrices without sacrificing either variable selection or computational speed.

Our key methodological contribution is a different approach for constructing a quadratic rule in high-dimensional settings compared to the ones taken in the literature. The existing methods rely on improved estimation of the full Bayes quadratic discriminant rule by exploring additional structural assumptions on Σ_g or Σ_g^{-1} [30, 31, 42, 46, 55]. In contrast, we modify Fisher's formulation of linear discriminant analysis for the case of unequal covariance matrices. The resulting method performs simultaneous variable selection and projection of original data on a lower-dimensional space, with the subsequent application of quadratic discriminant analysis. We call this approach discriminant analysis via projections, or DAP.

Unlike the existing quadratic methods, our rule is linear in p, which allows us to devise a very efficient optimization procedure to estimate simultaneously the projection directions and to perform variable selection. For $p = 500$, it takes around 1.5 seconds to implement our method, whereas the closest competing sparse quadratic method takes 30 minutes. This makes it possible to apply our approach in situations where other quadratic methods are computationally infeasible. Moreover, we connect the variables in our rule with the nonzero variables in the linear part of Bayes' quadratic rule, and prove the variable selection consistency of our method in high-dimensional settings. Empirical studies confirm that for large values of p, the proposed rule leads to competitive, and often smaller, misclassification error rates than the existing approaches. At the same time, our method consistently selects the sparsest models, thus achieving the best balance between model complexity and misclassification error rate. Finally, the application to gene expression data of breast cancer patients [7] confirms the crucial importance of ESR1 gene in differentiating estrogen receptor status; an insight that would be impossible to get with other approaches due to much higher complexity of corresponding classification rules.

The rest of this paper is organized as follows. In Section 2, we describe a new quadratic classification rule, discriminant analysis via projections. We connect the proposed approach to both linear and quadratic discriminant analysis, and derive an efficient optimization algorithm for sparse estimation. In Section 3, we provide theoretical guarantees on the variable selection consistency of our method in high-dimensional settings. In Section 4, we conduct empirical studies on both simulated and real data. In Section 5, we discuss possible extensions.

For a vector $v \in \mathbb{R}^p$, we let $\|v\|_1 = \sum_{i=1}^p |v_i|$, $\|v\|_2 = (\sum_{i=1}^p v_i^2)^{1/2}$, $\|v\|_\infty = \max_i |v_i|$. We use e_j to denote a unit norm vector with jth element being equal to one, and ι_p to denote the vector of ones of length p. For a matrix $M \in \mathbb{R}^{n\times p}$, we let $\|M\|_{\infty,2} = \max_{1\leq i\leq n}(\sum_{j=1}^p m_{ij}^2)^{1/2}$, $\|M\|_2 = \sup_{x:\|x\|_2=1}\|Mx\|_2$ and $|M|$ be the determinant of M. Given an index set A, we use M_A to denote the submatrix of M with columns indexed by A. For a square matrix M, we use M_{AA} to denote the submatrix of M with both rows and columns indexed by A. We use I to denote the identity matrix. We use $a_n \lesssim b_n$ to denote that there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for n sufficiently large. We also let $a \vee b = \max(a, b)$.

2. Discriminant analysis via projections

2.1. Review of Fisher's discriminant analysis

Consider n independent pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ from a random pair (X, Y) on $\mathbb{R}^p \times \{1, 2\}$. For $g \in \{1, 2\}$, let $\Sigma_g = \mathrm{cov}(X|Y = g)$, and assume $\Sigma_1 = \Sigma_2$. Fisher's discriminant analysis seeks a linear combination of p measurements that maximize between group variability with respect to within group variability [36, Chapter 11]:

$$\underset{v\in\mathbb{R}^p}{\text{maximize}} \left\{ \frac{v^\top(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^\top v}{v^\top W v} \right\}, \tag{1}$$

where $W = (n-2)^{-1}\sum_{g=1}^2 (n_g - 1)S_g$ is the pooled sample covariance matrix, S_g is the sample covariance matrix, n_g is the number of samples, and $\bar{x}_g$ is the sample mean for group g. Letting $\widehat{v}$ be a vector at which the maximum above is achieved, the resulting classification rule for a new observation with observed value $x \in \mathbb{R}^p$ is

$$h_{\widehat{v}}(x) = \underset{g\in\{1,2\}}{\text{argmin}} \left\{ (x^\top\widehat{v} - \bar{x}_g^\top\widehat{v})^\top(\widehat{v}^\top W\widehat{v})^{-1}(x^\top\widehat{v} - \bar{x}_g^\top\widehat{v}) - 2\ln(n_g/n) \right\}. \tag{2}$$

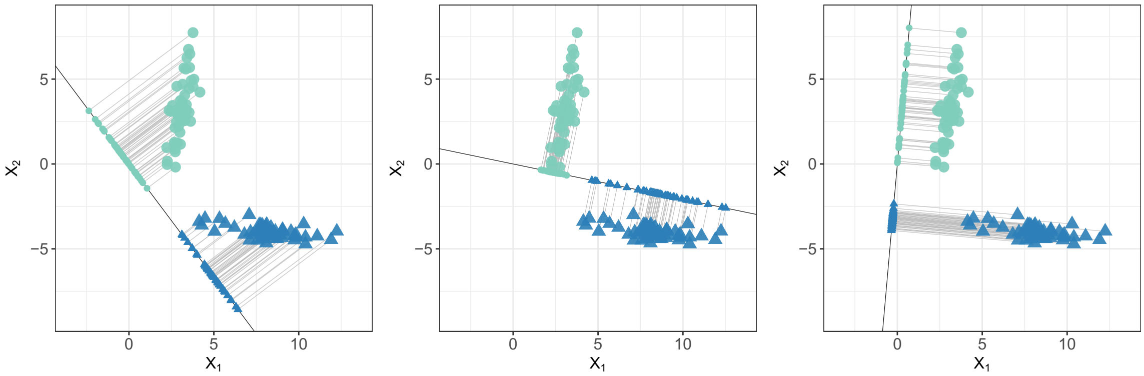

Figure 1: Two-group classification problem with $p = 2$ and unequal covariance matrices. Left: Projection using Fisher's discriminant vector. Middle: Projection using the covariance structure from the 1st group (circles). Right: Projection using the covariance structure from the 2nd group (triangles).

Hence, both the new observation $x \in \mathbb{R}^p$ and the data $X \in \mathbb{R}^{n\times p}$ are projected onto the line determined by $\widehat{v}$, and the classification is performed according to Mahalanobis distance to the class means in the projected space. Since both the objective function in (1) and the classification rule (2) are invariant to the scaling of discriminant vector $\widehat{v}$, when $n \gg p$ we can write $\widehat{v} = cW^{-1}(\bar{x}_1 - \bar{x}_2)$ for any constant $c \neq 0$. Moreover, Fisher's rule (2) coincides with the sample plug-in Bayes rule under the normality assumption, i.e., $X_i|Y_i = g \sim \mathcal{N}(\mu_g, \Sigma)$.

2.2. *Modification of Fisher's rule for the case of unequal covariance matrices*

Our proposal is based on the modification of criterion (1) to the case of unequal covariance matrices. Specifically, we consider two discriminant directions instead of one. For $g \in \{1, 2\}$, let

$$\widehat{v}_g = \operatorname*{argmax}_{v_g \in \mathbb{R}^p} \left\{ \frac{v_g^\top(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^\top v_g}{v_g^\top S_g v_g} \right\}. \tag{3}$$

Similar to Fisher's criterion, when $n_g \gg p$, the solutions to (3) can be expressed as $\widehat{v}_g = c_g S_g^{-1}(\bar{x}_1 - \bar{x}_2)$ for any $c_1 \neq 0, c_2 \neq 0$. Subsequently, given matrix $\widehat{V} = [\widehat{v}_1\, \widehat{v}_2]$, we modify rule (2) to take into account unequal covariance matrices as

$$h_{\widehat{V}}(x) = \operatorname*{argmin}_{g\in\{1,2\}} \left\{ (x - \bar{x}_g)^\top \widehat{V}(\widehat{V}^\top S_g \widehat{V})^{-1}\widehat{V}^\top(x - \bar{x}_g) + \ln|\widehat{V}^\top S_g \widehat{V}| - 2\ln(n_g/n) \right\}. \tag{4}$$

Remark 1. If $\widehat{v}_1$ and $\widehat{v}_2$ are linearly dependent, then $\widehat{V}$ has rank one, and $\widehat{V}^\top S_1 \widehat{V}$ and $\widehat{V}^\top S_2 \widehat{V}$ are both singular. In this case the subspace spanned by the columns of $\widehat{V}$ is the same as the subspace spanned by only one column, and we use $\widehat{V} = \widehat{v}_1$ in (4).

Rule (4) is equivalent to applying quadratic discriminant rule to $\widehat{V}^\top x$ instead of applying it directly to x. Unlike the equivalence between Fisher's rule and the linear discriminant rule, in Section 2.6 we show that rule (4) is generally not equivalent to quadratic discriminant analysis. Nevertheless, for a given $\widehat{V}$, formulation (4) allows to overcome possible rank degeneracy of S_g as well as perform variable selection. First, rule (4) requires inversion of 2×2 matrices $\widehat{V}^\top S_g \widehat{V}$, which are likely to be positive definite, in contrast to S_g. Secondly, since (4) effectively applies quadratic rule to $\widehat{V}^\top x$ instead of x, it only relies on those variables for which the corresponding rows of $\widehat{V}$ are nonzero. Hence, performing variable selection is equivalent to using row-sparse matrix $\widehat{V}$. Figure 1 shows that each $\widehat{v}_g$ from (3) can be viewed as a basis vector for the reduced space, and coincides with discriminant vector $\widehat{v}$ in Fisher's rule (1) if the pooled sample covariance matrix $W = S_1 = S_2$. Therefore, we call rule (4) the discriminant analysis via projections.

2.3. *Sparse estimation*

While rule (4) allows to overcome the potential singularity of sample covariance matrices, it still requires estimation of $O(p)$ parameters in $\widehat{V}$. Moreover, singularity of S_g leads to non-uniqueness of the solutions to (3) creating difficulties for the interpretation. Therefore, rule (4) may still have poor performance in the high-dimensional settings when $p \gg n$. At the same time, in the context of linear discriminant analysis the classification performance can be significantly improved by directly estimating the discriminant vector with sparsity regularization [5, 35].

Guided by this intuition, our goal is to obtain sparse estimates of $\psi_1 = c_1\Sigma_1^{-1}\delta$ and $\psi_2 = c_2\Sigma_2^{-1}\delta$ with $\delta = \mu_1 - \mu_2$, which are the population counterparts of $\widehat{v}_1$ and $\widehat{v}_2$ in (3). This approach leads to regularized row-sparse $\widehat{V}$ that can be used directly in rule (4). The direct estimation of ψ_g with sparse regularization has several advantages. First, the covariance matrices serve as nuisance parameters since $\psi_g \propto \Sigma_g^{-1}\delta$ are functions of covariance matrices, not the covariance matrices themselves. Second, as we discuss in more detail below, sparse penalization leads to unique well-defined solutions even when sample covariance matrices are singular. Finally, the sparsity in $\widehat{V}$ leads to simpler and more interpretable classification rule.

To produce sparse estimates of ψ_1 and ψ_2, we consider penalized empirical risk minimization framework:

$$\widehat{V} = [\widehat{v}_1\ \widehat{v}_2] = \underset{v_1, v_2 \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \widehat{L}_{\psi_1}(v_1) + \widehat{L}_{\psi_2}(v_2) + \lambda \operatorname{Pen}(V) \right\},$$

where $\widehat{L}_{\psi_1}(v_1)$, $\widehat{L}_{\psi_2}(v_2)$ are empirical loss functions associated with ψ_1, ψ_2, $\lambda > 0$ is the tuning parameter, and $\operatorname{Pen}(V)$ is the sparsity-inducing penalty.

Remark 2. Another possibility is to add sparse penalization directly within criterion (3). In linear discriminant analysis, this approach leads to significant improvement over sample plug-in rule [53]. However, it also leads to nonconvex optimization problem and potential difficulties in obtaining very sparse solutions [16]. Therefore, we do not pursue the direct penalization here.

First, we discuss our choice of penalty. As we are interested in simultaneous variable selection, that is row-sparsity of $\widehat{V}$, we propose to use group penalty. Specifically, we choose group-lasso, $\operatorname{Pen}(V) = \sum_{j=1}^{p}(v_{1j}^2 + v_{2j}^2)^{1/2}$, due to its convexity [56]. Other possibilities include nonconvex group penalties; we refer the reader to Huang et al. [23] for the review.

Next, we discuss our choice of empirical loss functions $\widehat{L}_{\psi_1}(v_1)$ and $\widehat{L}_{\psi_2}(v_2)$. Both criterion (3) and rule (4) are invariant to the scale of $\widehat{V}$, i.e., to the choice of constants c_1 and c_2. While the naive approach is to fix $c_1 = c_2 = 1$, we use $c_1 = \pi_2/(1 + \pi_2^2\delta^\top\Sigma_1^{-1}\delta)$, $c_2 = \pi_1/(1 + \pi_1^2\delta^\top\Sigma_2^{-1}\delta)$, which leads to a lower-bounded empirical loss function as well as significant computational savings. To be specific, we take advantage of the following equivalence due to the Sherman–Morrison formula.

Proposition 1. *For any $\rho \neq 0$, any non-singular matrix $M \in \mathbb{R}^{p\times p}$ and any vector $a \in \mathbb{R}^p$,*

$$(M + \rho^2 aa^\top)^{-1}\rho a = \rho M^{-1}a(1 + \rho^2 a^\top M^{-1}a)^{-1} \propto M^{-1}a.$$

Our choice of c_1 and c_2 leads to $\psi_1 = (\Sigma_1 + \pi_2^2\delta\delta^\top)^{-1}\pi_2\delta$ and $\psi_2 = (\Sigma_2 + \pi_1^2\delta\delta^\top)^{-1}\pi_1\delta$. Consider the following quadratic loss function associated with ψ_1

$$L_{\psi_1}(v_1) = (v_1 - \psi_1)^\top(\Sigma_1 + \pi_2^2\delta\delta^\top)(v_1 - \psi_1)/2 = v_1^\top\Sigma_1 v_1/2 + (\pi_2\delta^\top v_1 - 1)^2/2 + C,$$

where C is a constant independent of v_1. Consider the empirical version of this loss function

$$\widehat{L}_{\psi_1}(v_1) = v_1^\top S_1 v_1/2 + \left(n^{-1}n_2 d^\top v_1 - 1\right)^2/2 + C, \tag{5}$$

where $d = \bar{x}_1 - \bar{x}_2$. First, $\widehat{L}_{\psi_1}(v_1)$ is invariant under linear transformation of the data [45]. Second, $\widehat{L}_{\psi_1}(v_1)$ is always bounded from below by C, even when S_1 is singular. This ensures convergence of the block-coordinate descent algorithm without the need to regularize S_1, and in particular, is not the case for $c_1 = 1$.

Furthermore, let $X_1 \in \mathbb{R}^{n_1\times p}$ be the submatrix of X corresponding to the first group, and $X_2 \in \mathbb{R}^{n_2\times p}$ be the one corresponding to the second group. Let X be column-centered so that $\bar{x} = n^{-1}(n_1\bar{x}_1 + n_2\bar{x}_2) = 0$, and hence $d = n_2^{-1}n\bar{x}_1$. Then the loss (5) can be rewritten as

$$\begin{aligned}\widehat{L}_{\psi_1}(v_1) &= v_1^\top S_1 v_1/2 + \left(\bar{x}_1^\top v_1 - 1\right)^2/2 + C = n_1^{-1}v_1^\top X_1^\top X_1 v_1/2 - v_1^\top\bar{x}_1 + C \\ &= n_1^{-1}\|X_1 v_1 - \iota_{n_1}\|_2^2/2 + C.\end{aligned}$$

That is, the loss function can be expressed as the linear regression loss function. Similarly,

$$\widehat{L}_{\psi_2}(v_2) = n_2^{-1}\|X_2 v_2 + \iota_{n_2}\|_2^2/2 + C.$$

Therefore, our choice of c_1 and c_2 allows to re-express the problem of estimating ψ_1 and ψ_2 as a regression problem. This leads to the efficient optimization algorithm described in Section 2.4.

In summary, given the column-centered data matrix $X \in \mathbb{R}^{n\times p}$ with submatrices $X_1 \in \mathbb{R}^{n_1\times p}$, $X_2 \in \mathbb{R}^{n_2\times p}$ corresponding to two groups, we find $\widehat{V} = [\widehat{v}_1\ \widehat{v}_2] \in \mathbb{R}^{p\times 2}$ as the solution to

$$\underset{V=[v_1,v_2]\in\mathbb{R}^{p\times 2}}{\text{minimize}} \left\{ n_1^{-1}\|X_1 v_1 - \iota_{n_1}\|_2^2/2 + n_2^{-1}\|X_2 v_2 + \iota_{n_2}\|_2^2/2 + \lambda \sum_{j=1}^{p} (v_{1j}^2 + v_{2j}^2)^{1/2} \right\}. \tag{6}$$

If $\lambda = 0$, $\widehat{V}$ coincides with the solution to (3) up to the choice of scaling. If $\lambda > 0$, then $\widehat{V}$ is row-sparse leading to variable selection. Given $\widehat{V}$, we apply rule (4) for classification.

2.4. Optimization algorithm

In this section we derive a block-coordinate descent algorithm to solve (6). Consider the optimality conditions with respect to each block $v_j = (v_{1j}, v_{2j})^\top$:

$$n_1^{-1}X_{1j}^\top X_{1j} v_{1j} = n_1^{-1}X_{1j}^\top(\iota_{n_1} - \sum_{k\neq j} v_{1k}X_{1k}) - \lambda u_{1j}, \quad n_2^{-1}X_{2j}^\top X_{2j} v_{2j} = n_2^{-1}X_{2j}^\top(-\iota_{n_2} - \sum_{k\neq j} v_{2k}X_{2k}) - \lambda u_{2j};$$

see Chapter 5 in [3]. In the above, $u_j = (u_{1j}, u_{2j})^\top$ is the subgradient of $(v_{1j}^2 + v_{2j}^2)^{1/2}$ such that $u_j = v_j/\|v_j\|_2$ if $\|v_j\|_2 \neq 0$, and $u_j \in \{u : \|u\|_2 \leq 1\}$ if $\|v_j\|_2 = 0$.

In general, $n_1^{-1}X_{1j}^\top X_{1j} \neq n_2^{-1}X_{2j}^\top X_{2j}$, hence the block-update is not available in closed form and requires a line search [2]. However, guided by the computational considerations as well as the ideas of standardized group lasso [47], we pre-standardize X_1 and X_2 so that $n_1^{-1}\text{diag}(X_1^\top X_1) = n_2^{-1}\text{diag}(X_2^\top X_2) = \iota_p$, and then perform the back-scaling of $\widehat{v}_1, \widehat{v}_2$. This ensures that the penalization of different variables is independent of their relative scales. Finally, we are ready to present the algorithm.

Define the residual vectors r_1, r_2 as

$$r_{1j} = n_1^{-1}X_{1j}^\top\left(\iota_{n_1} - \sum_{l=1}^{p} v_{1l}X_{1l}\right), \quad r_{2j} = n_2^{-1}X_{2j}^\top\left(-\iota_{n_2} - \sum_{l=1}^{p} v_{2l}X_{2l}\right);$$

with $r_j = (r_{1j}, r_{2j})^\top$. From the optimality conditions, the equation for the jth block $v_j = (v_{1j}, v_{2j})^\top$ takes form

$$v_j = \left(1 - \lambda/\|v_j + r_j\|_2\right)_+ \left(v_j + r_j\right),$$

where $a_+ = \max(0, a)$. Starting with initial value $V^{(0)}$, the block-coordinate descent algorithm proceeds by iterating the updates of v_1, v_2 with updates of residuals r_1, r_2 until convergence. Due to the convexity of (6), the boundedness of the objective function from below, and the separability of the penalty with respect to block updates, the global optimum is finite and the algorithm is guaranteed to converge to the global optimum from any starting point [49].

2.5. Connection with sparse linear discriminant analysis

We show that sparse linear discriminant analysis can be viewed as a very special case of the proposed approach.

Proposition 2. *Consider the sparse discriminant analysis in Gaynanova et al. [15] that finds the discriminant vector $\widetilde{v}(\lambda)$ for a given value of tuning parameter $\lambda > 0$. Define $c = (n_1/n)^{1/2} + (n_2/n)^{1/2}$. Under the additional constraint $(n/n_1)^{1/2}v_1 = (n/n_2)^{1/2}v_2$, the solution to (6) satisfies*

$$(n/n_1)^{1/2}\widehat{v}_1(\lambda) = (n/n_2)^{1/2}\widehat{v}_2(\lambda) = c\,\widetilde{v}(\lambda/c)\,.$$

While in Proposition 2 we connect our approach with Gaynanova et al. [15] due to a more straightforward proof, in the two-group case the method of Gaynanova et al. [15] is equivalent to the method of Mai et al. [35]. Moreover, Mai and Zou [34] show equivalence for the two-group case between the methods of Mai et al. [35], Clemmensen et al. [9] and Wu et al. [54]. Therefore, when discriminant directions v_1 and v_2 are additionally restricted to be collinear as in Proposition 2, our proposed approach (6) reduces to this class of sparse linear discriminant analysis methods up to scaling.

2.6. Connection with quadratic discriminant analysis

Let Y be a group indicator, $\Pr(Y = 1) = \pi_1$ and $\Pr(Y = 2) = 1 - \pi_1 = \pi_2$, and consider $X|Y = g \sim \mathcal{N}(\mu_g, \Sigma_g)$ $(g = 1, 2)$. The Bayes rule assigns a new observation with observed value $x \in \mathbb{R}^p$ to group one if and only if

$$x^\top(\Sigma_2^{-1} - \Sigma_1^{-1})x - 2x^\top(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) + \ln\left(|\Sigma_2|/|\Sigma_1|\right) - \mu_1^\top\Sigma_1^{-1}\mu_1 + \mu_2^\top\Sigma_2^{-1}\mu_2 + 2\ln(\pi_1/\pi_2) > 0. \tag{7}$$

Consider centering x by the overall mean $\mathrm{E(X)} = \mu = \pi_1\mu_1 + \pi_2\mu_2$.

Proposition 3. *Let $\delta = \mu_1 - \mu_2$. The Bayes rule (7) can be written as*

$$\begin{aligned}(x-\mu)^\top(\Sigma_2^{-1} - \Sigma_1^{-1})(x-\mu) + \ln\left(|\Sigma_2|/|\Sigma_1|\right)\\ + 2(x-\mu)^\top(\pi_1\Sigma_2^{-1}\delta + \pi_2\Sigma_1^{-1}\delta) + \pi_1^2\delta^\top\Sigma_2^{-1}\delta - \pi_2^2\delta^\top\Sigma_1^{-1}\delta + 2\ln(\pi_1/\pi_2) > 0.\end{aligned} \tag{8}$$

Consider the population version of the proposed discriminant analysis via projections, that is applying Bayes rule to $\Psi^\top X$ with $\Psi^\top X|Y = g \sim \mathcal{N}(\Psi^\top\mu_g, \Psi^\top\Sigma_g\Psi)$ and $\Psi = [\psi_1,\ \psi_2] = [c_1\Sigma_1^{-1}\delta,\ c_2\Sigma_2^{-1}\delta]$, $c_1, c_2 \neq 0$.

Proposition 4. *Consider the population version of rule (4), that is substituting Ψ for $\widehat{V}$, Σ_g for S_g, μ_g for $\bar{x}_g$ and π_g for n_g/n. A new observation with value x is assigned to group one if and only if*

$$\begin{aligned}(x-\mu)^\top\Psi\left\{(\Psi^\top\Sigma_2\Psi)^{-1} - (\Psi^\top\Sigma_1\Psi)^{-1}\right\}\Psi^\top(x-\mu) + \ln\left(|\Psi^\top\Sigma_2\Psi|/|\Psi^\top\Sigma_1\Psi|\right)\\ + 2(x-\mu)^\top(\pi_1\Sigma_2^{-1}\delta + \pi_2\Sigma_1^{-1}\delta) + \pi_1^2\delta^\top\Sigma_2^{-1}\delta - \pi_2^2\delta^\top\Sigma_1^{-1}\delta + 2\ln(\pi_1/\pi_2) > 0.\end{aligned} \tag{9}$$

The only difference between the rules in Proposition 3 and 4 is on the first line, which involves the quadratic and the log terms. The linear terms and the remaining constant terms are identical. Therefore, rule (9) can be viewed as an approximation to rule (8). Further comparison between the two rules in terms of induced J-divergences between class-distributions is in Appendix B.

While rule (9) is not the same as the Bayes rule, and therefore will lead to inferior performance at the population level, in Section 4 we see this relationship to be reversed when the corresponding regularized sample versions are considered and p is large relative to the sample size n. The main advantage of rule (9) comes from the significant reduction in the number of parameters to be estimated. Specifically, matrix Ψ has $p \times 2$ elements leading to $O(p)$ parameters in rule (9). In contrast, the Bayes rule requires estimation of the $\Sigma_2^{-1} - \Sigma_1^{-1}$ leading to $O(p^2)$ parameters in total.

3. Variable selection consistency in high-dimensional settings

We establish the variable selection consistency of estimator in (6) under the following assumptions.

Assumption 1 (Normality)**.** *$X_i|Y_i = g \sim \mathcal{N}(\mu_g, \Sigma_g)$, $\Pr(Y_i = g) = \pi_g$ for $g = 1, 2$ with $0 < \pi_{\min} \leq \pi_1/\pi_2 \leq \pi_{\max} < 1$.*

Assumption 2 (Sparsity)**.** *Let $\delta = \mu_1 - \mu_2$, $A = \{i : (e_i^\top\Sigma_1^{-1}\delta)^2 + (e_i^\top\Sigma_2^{-1}\delta)^2 \neq 0\}$, $A^\complement = \{1, \dots, p\}/A$ and $\mathrm{card}(A) = s$. That is, A is the index set of nonzero variables in $\Sigma_1^{-1}\delta$ or $\Sigma_2^{-1}\delta$.*

Assumption 3 (Irrepresentability)**.** *There exist $\alpha \in (0, 1]$ such that*

$$\max_{\substack{u_1, u_2 \in \mathbb{R}^s\\ u_{1i}^2 + u_{2i}^2 \leq 1\ \forall i}} \|\Sigma_{1A^\complement A}\Sigma_{1AA}^{-1}u_1, \Sigma_{2A^\complement A}\Sigma_{2AA}^{-1}u_2\|_{\infty,2} \leq 1 - \alpha.$$

Assumption 4. *$0 < c \leq \lambda_{\min}(\Sigma_{gAA}) \leq \lambda_{\max}(\Sigma_{gAA}) \leq C$ and $e_j^\top\Sigma_g e_j \leq M$ for all $j \in \{1, \dots, p\}$ and $g \in \{1, 2\}$.*

Assumption 1 is standard in the context of discriminant analysis [17, 26, 35], and Assumptions 2–3 are typical in establishing variable selection consistency of penalized estimators in high-dimensional settings [1, 40, 50]. In light of Proposition 3, Assumption 2 can be interpreted as requiring the linear part of Bayes rule to be sparse, i.e., there are only s nonzero main effects. More specifically, the sparsity of both covariance matrices and mean differences as in Li and Shao [31] is sufficient for Assumption 2 to hold, but not necessary. We use Assumption 4 for the convenience of treating the parameters depending on Σ_g as constants and presenting the rates in Theorems 1 and 2 through only n, p and s. We refer the reader to the Online Supplement for the more general statements of Theorems 1 and 2 without the use of Assumption 4. To prove the variable selection consistency of estimator in (6), we use the primal-dual witness technique [50]. First, we prove that under the appropriate scaling of the sample sizes, and sufficiently large value of the tuning parameter λ, the variables in $A^\complement$ are set to zero with high probability. Let $\widehat{A} = \{i : \widehat{v}_{1i}^2 + \widehat{v}_{2i}^2 \neq 0\}$ denote the support of the solution to (6).

Theorem 1. *Let Assumptions 1–4 hold, the sample sizes satisfy* $\min_g n_g \gtrsim s \ln\{(p-s)\eta^{-1}\}$ *for some* $\eta \in (0,1)$*, and the tuning parameter satisfy* $\lambda \gtrsim [\ln\{(p-s)\eta^{-1}\}/n]^{1/2}$*. Then* $\Pr(\widehat{A} \subseteq A) \geq 1-\eta$*.*

Next, we show that under the additional assumption on the minimal signal strength defined as

$$\psi_{\min} = \min_{j \in A} \left\{ \pi_2^2 (e_j^\top \Sigma_1^{-1} \delta)^2 + \pi_1^2 (e_j^\top \Sigma_2^{-1} \delta)^2 \right\}^{1/2},$$

the true variables are nonzero with high probability leading to perfect recovery. In sparse linear models this assumption is often called the β-min condition [50]. According to Proposition 3, $\psi_{\min}$ can be interpreted as the smallest magnitude of the nonzero variables in the linear part of the Bayes quadratic discriminant rule.

Theorem 2. *Let the conditions of Theorem 1 hold and* $\psi_{\min} \gtrsim \lambda s^{1/2} (\max_g \delta_A^\top \Sigma_{gAA}^{-1} \delta_A \vee 1)$*. Then* $\Pr(\widehat{A} = A) \geq 1-\eta$*.*

Theorem 2 reveals the advantage of using the group penalty in joint sparse estimation of ψ_1 and ψ_2. If variable j is nonzero in both ψ_1 and ψ_2, then it is sufficient to have a large signal in only one of ψ_1 or ψ_2 for the minimal signal strength condition to hold. In contrast, separate estimation via the lasso penalty will lead to the requirement of sufficiently large signal in both ψ_1 and ψ_2 simultaneously.

4. Empirical studies

4.1. Simulated data

We compare the misclassification error rates and variable selection performance of the following methods: (i) Sample QDA, rule (7) with plug-in estimates $\bar{x}_1$, $\bar{x}_2$, S_1, S_2; (ii) Sparse QDA of Le and Hastie [30]; (iii) Sparse QDA of Li and Shao [31]; (iv) Sparse QDA via ridge fusion [42]; (v) Logistic regression with pairwise interactions and lasso penalty on the vector of coefficients; (vi) Regularized discriminant analysis [13]; (vii) Sparse LDA [15, 35]; (viii) Discriminant analysis via projections proposed in this paper, i.e., rule (4) with estimator from (6). Since the focus of the paper is on quadratic classification rules, we only use one linear discriminant analysis method for comparison. We expect that a choice of a different linear method, such as those of Witten and Tibshirani [53] or Niu et al. [39], will lead to similar conclusions. The details of all methods' implementation, together with tuning parameter selection criteria, are described in Appendix A.

We fix the sample sizes $n_1 = n_2 = 100$, the dimension $p \in \{100, 500\}$, and the group means $\mu_1 = 0_p$ and $\mu_2 = (\{1\}_5, \{-1\}_5, \{0\}_{p-10})$. We consider the following types of covariance structures:

1. Block-equicorrelation with block size $b \in \{10, 100\}$ and $\rho \in [0,1]$:

$$\Sigma_g = \begin{pmatrix} \rho I_b + (1-\rho)\iota_b \iota_b{}^\top & 0 \\ 0 & I_{p-b} \end{pmatrix}.$$

2. Block-autocorrelation with block size $b \in \{10, 100\}$ and $\rho \in [0,1]$:

$$\Sigma_g = \{\Sigma_g\}_{i,j}, \quad \{\Sigma_g\}_{i,j} = \begin{cases} \rho^{|i-j|} & (1 \leq i,j \leq b), \\ \mathbf{1}\{i=j\} & (\text{otherwise}). \end{cases}$$

3. Spiked with parameters $q_1, q_2 \in \mathbb{R}^p$: $\Sigma_g = 30 q_1 q_1^\top + 2 q_2 q_2^\top + I$.
   - (a) Block size $b = 10$: $q_1 = (\{1/\sqrt{5}\}_5, \{0\}_{p-5})$, $q_2 = (\{0\}_{p-5}, \{1/\sqrt{5}\}_5, \{0\}_{p-10})$.
   - (b) Block size $b = 100$: $q_1 = (1, \ldots, 100, \{0\}_{p-100})^\top$, $q_2 = (I - q_1 q_1^\top)(100, \ldots, 1, \{0\}_{p-100})^\top$. q_1 and q_2 are normalized so that $q_1^\top q_1 = 1$ and $q_2^\top q_2 = 1$.

These structures are common in assessing the performance of discriminant analysis methods [30, 35, 44]. We use eight combinations as described in Table 1, and fix the block sizes to make the Bayes error rate independent of p.

As expected, the sample QDA performs the worst, with misclassification error rates being larger than 40% consistently across all settings. Therefore, in Figure 2 we only present the rates for the other methods. First, we compare the proposed approach with sparse LDA. While in models 1, 2 and 8 they perform similarly, accounting for unequal covariance matrices results in drastic improvements on models 4–7. When comparing our approach to sparse QDA methods, the relative ranking often depends on p. For example, when $p = 100$, ridge fusion of Price et al. [42] is better than our proposal on models 2 and 8, but is significantly worse on the same models when

Table 1: List of considered models for Σ_1 and Σ_2.

Model	Σ_1	Σ_2
1	equicorrelation, $b = 100, \rho = 0.5$	equicorrelation, $b = 100\ \rho = 0.5$
2	autocorrelation, $b = 100\ \rho = 0.8$	equicorrelation, $b = 100, \rho = 0.5$
3	autocorrelation, $b = 10, \rho = 0.5$	equicorrelation, $b = 10, \rho = 0.8$
4	spiked, $b = 10$	spiked, $b = 10$ (q_1 and q_2 reversed)
5	spiked, $b = 100$	spiked, $b = 10$ (q_1 and q_2 reversed)
6	spiked, $b = 10$	equicorrelation, $b = 10, \rho = 0.8$
7	spiked, $b = 10$	equicorrelation, $b = 100, \rho = 0.3$
8	spiked, $b = 100$	equicorrelation, $b = 100, \rho = 0.3$

Table 2: Median time (seconds) over 10 replications to fully implement each classification method for one instance of model 8. DAP: Discriminant analysis via projections, proposed; SLDA: Sparse linear discriminant analysis; RDA: Regularized discriminant analysis; SLOG: Sparse logistic regression with interactions; SQDA_LH: Sparse QDA of Le and Hastie [30]; SQDA_RF: Sparse QDA via ridge fusion; SQDA_LS: Sparse QDA of Li and Shao [31].

p	DAP	SLDA	RDA	SLOG	SQDA_LH	SQDA_RF	SQDA_LS
100	0.6	0.4	3.1	2.7	139.5	868.5	52.6
300	1.0	1.4	5.0	28.8	2071.9	11681.4	481.5
500	1.4	1.7	5.0	117.1	7282.2	45161.7	1791.4

$p = 500$. Similarly, sparse QDA of Le and Hastie [30] is significantly better than our proposal on models 6 and 8 when $p = 100$, but significantly worse on the same models when $p = 500$. This confirms that the proposed rule is well-suited to high-dimensional settings. Among the sparse QDA approaches, we find that the method of Li and Shao [31] is most consistent across dimensions. In particular, it leads to better error rates on models 4 and 5 (2% difference in median error rates). Nevertheless, it still leads to significantly worse error rates on models 1, 2, 6 and 8. Finally, the proposed approach performs better than regularized discriminant analysis in all cases but model 2, $p = 100$, and performs as well or better than the sparse logistic regression in all scenarios.

Overall, we found that no method is universally the best in terms of error rates since the relative ranking depends on the particular model and the underlying dimension. This is consistent with previous research. In the words of Wu et al. [55], "it is difficult to imagine that there could be a universally optimal discriminant analysis method for high-dimensional data. Almost every method can enjoy some advantages under certain circumstances." Nevertheless, three methods stand out as the best across all models and dimensions: our proposal and sparse QDA methods of Le and Hastie [30] and Li and Shao [31]. Moreover, our proposal achieves comparable, and in certain scenarios significantly better, error rates than the best other methods in all the cases with $p = 500$ except model 2.

In summary, Figure 3 shows that the proposed discriminant analysis via projections significantly improved over sparse LDA method, and results in competitive, and often better, misclassification error rates than existing QDA proposals. The real advantages of our approach, however, become certain when comparing variable selection performance and computational speed. Figure 3 reveals that the proposed method consistently uses the sparsest model (less than 50 variables for most scenarios). In comparison, the methods of Le and Hastie [30] and Price et al. [42] always use all p variables, and are such much less interpretable.

We further compare the execution time of each method on a Linux machine with Intel Xeon X5560 @2.80 GHz. We define execution time as the full time for method's implementation: tuning parameter selection plus model fitting plus classification. We use one instance of model 8 with $p \in \{100, 300, 500\}$, and R package `microbenchmark` [37] with 10 evaluations of each expression. Table 2 shows that the execution times increase dramatically with p for logistic regression with interactions and sparse QDA methods, whereas the times are quite consistent across dimensions for sparse LDA, RDA and our approach. Logistic regression is noticeably faster than sparse QDA methods mainly due to the difference in tuning parameter selection criterion: it uses BIC instead of cross-validation. Using cross-validation for logistic regression makes it too computationally demanding for the range of p we considered. Sparse LDA and the proposed method are the fastest, confirming that they are well-suited for the use on high-dimensional datasets in practice.

4.2. Benchmark datasets

We compare the proposed discriminant analysis via projections with competitors on three benchmark datasets described in Table 3. These datasets are commonly used to assess classification performance [32, 39, 44], and are publicly available from the R package `datamicroarray` [43].

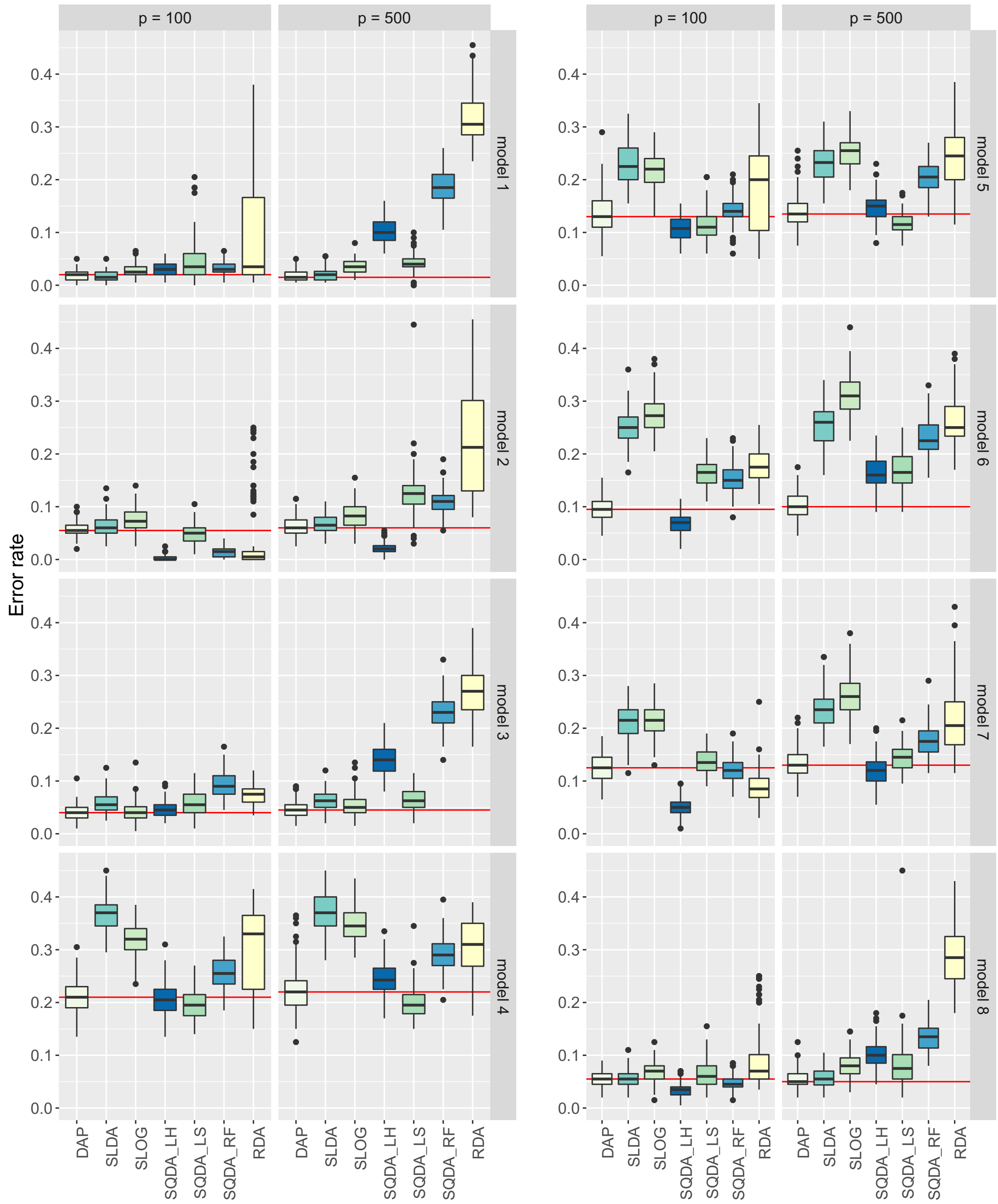

Figure 2: Misclassification error rates over 100 replications, the horizontal lines show the median errors of the proposed DAP, discriminant analysis via projections. SLDA: Sparse linear discriminant analysis; SLOG: Sparse logistic regression with interactions; SQDA_LH: Sparse QDA of Le and Hastie [30]; SQDA_LS: Sparse QDA of Li and Shao [31]; SQDA_RF: Sparse QDA via ridge fusion; RDA: Regularized discriminant analysis.

Table 3: Description of benchmark datasets used for methods comparison

Dataset	# samples in group 1	# samples in group 2	# gene expressions
chin [7]	75 (ER-positive)	43 (ER-negative)	22,215
gravier [18]	111 (good, no event)	57 (poor)	2,905
chowdary [8]	62 (breast tissue)	42 (colon tissue)	22,283

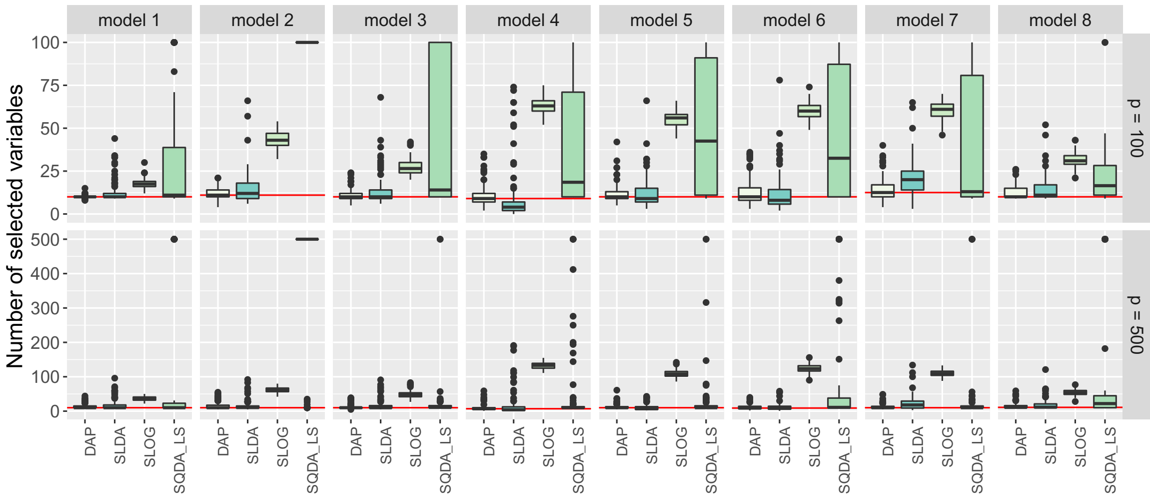

Figure 3: Number of selected variables over 100 replications, the horizontal lines indicate the median model sizes of proposed DAP, discriminant analysis via projections. RDA, SQDA_RF and SQDA_LH use all p variables, not shown. SLDA: Sparse linear discriminant analysis; SLOG: Sparse logistic regression with interactions; SQDA_LH: Sparse QDA of Le and Hastie [30]; SQDA_LS: Sparse QDA of Li and Shao [31]; SQDA_RF: Sparse QDA via ridge fusion; RDA: Regularized discriminant analysis.

We randomly split each dataset 100 times preserving the class proportions, and use 80% for training and 20% for testing. To reduce the computational cost associated with sparse quadratic discriminant analysis, we reduce the number of variables at each split by selecting the top $p = 1000$ variables with largest absolute value of the two-sample t-statistic on the training data, similar approach has been taken in Cai and Liu [5]. For fair comparison, we use the same set of 1000 variables for each of the methods. We do not consider sample quadratic discriminant analysis given its uniformly poor performance in Section 4.1. We also do not consider sparse logistic regression with interactions or ridge fusion due to computational issues when $p = 1000$ and their inferiority to other approaches in Section 4.1.

The results are shown in Figure 4. For *chin* and *chowdary*, similar misclassification error rates are reported in Niu et al. [39]. For the *chin* dataset, the error rates are the worst for linear discriminant analysis confirming the importance of taking into account unequal covariance matrices, and are the same for other methods. At the same time, the proposed DAP rule selects significantly smaller model than the competitors (median model size is one). For the *chowdary* dataset, the best performing method is RDA [13], however the relative difference is only one misclassification on the test data. The smallest model again corresponds to proposed DAP. For the *gravier* dataset, the best performing methods are ours and sparse QDA of Li and Shao [31]. Surprisingly, however, the method of Li and Shao [31] results in no variable selection on these datasets, the model size is 1000 over almost all replications (not shown). We suspect that the poor variable selection performance may be due to the crudeness of bisection procedure for selecting the tuning parameters. In summary, the proposed approach, discriminant analysis via projections, consistently selects the smallest model, often using less than 20 variables to achieve the same or better error rates than alternative methods. We conclude that it exhibits excellent prediction accuracy with the smallest model complexity.

We further analyze the *chin* dataset using variable selection results of our approach. Figure 4 reveals that the median model size is 1. This means that in most of the replications it is sufficient to look at the expression level of only one gene to achieve the same misclassification error rate as the other methods. We investigate whether the same gene is selected at each replication, and find that estrogen receptor 1 gene ESR1 is selected in 97 out of 100 cases. Our finding confirms previous studies on a strong link between ESR1 gene and estrogen receptor protein expression in breast cancer patients [21, 24, 28]. We refer the reader to Holst [20] for a review on the importance of ESR1 gene amplification in breast cancer. The gene with the second highest frequency of selection, 26 out of 100 cases, is LPIN1, which is also found to be differentially expressed in ER positive and negative patients in previous studies [6]. The relatively low selection frequency of LPIN1 is due to the median model size one, which leads to only ESR1 being selected and no other gene. While the strong link between ER protein expression status and ESR1 gene is not surprising, unlike the previous studies we did not focus on the ESR1 gene in advance. We consider all 22 thousand genes, and let our method determine that ESR1 is crucial for ER status of breast cancer. We want to emphasize that this insight is not possible with other approaches we tried. Regularized

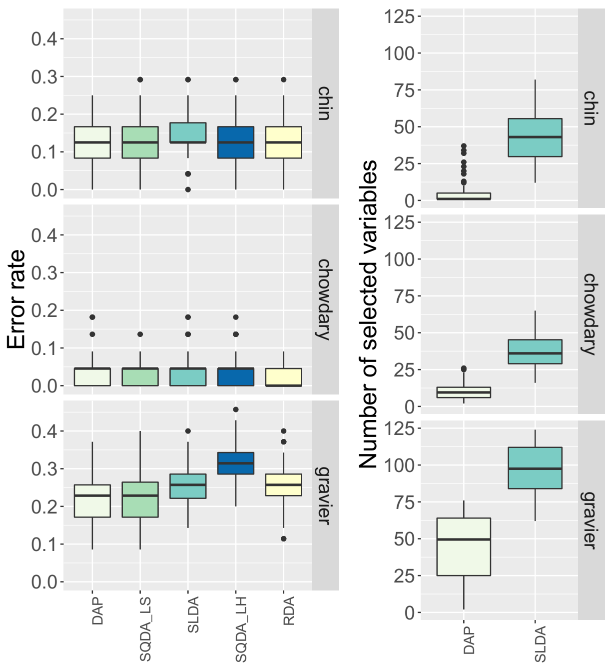

Figure 4: Left: Misclassification error rates over 100 splits. Right: Number of variables used in corresponding classification rules. DAP consistently selects the smallest model. SQDA_LS, SQDA_LH and RDA always use all $p = 1000$ variables, not shown. DAP: Discriminant analysis via projections, proposed method; SQDA_LS: Sparse QDA of Li and Shao [31]; SQDA_LH: Sparse QDA of Le and Hastie [30]; SLDA: Sparse linear discriminant analysis; RDA: Regularized discriminant analysis.

discriminant analysis of Friedman [13] and sparse QDA by Le and Hastie [30] use all 1000 variables, hence cannot be directly used for identifying important genes. Sparse LDA selects a smaller number of genes, but it has worse misclassification error rate and the median model size is still 45 variables, significantly larger than the number of variables used by our approach.

5. Discussion

In this work we propose a new rule for high-dimensional classification in the case of unequal covariance matrices. While the proposed approach in general differs from the Bayes rule on the population level, we show that the nonzero variables in our rule correspond to nonzero variables in the linear part of the Bayes quadratic rule. This connection combined with computational efficiency of our approach suggests that one can potentially use our method as a variable screening tool. Indeed, the empirical studies in Section 4.1 indicate that the performance of full quadratic methods deteriorates significantly with increase in p, however for small p they are computationally feasible and may lead to better error rates. We have not explored the screening properties of our approach in this work, but leave it for future investigation.

We focus on the two-group classification setting, however extending the methodology to the multi-group setting will likely lead to even further computational gains. One of the main challenges in the multi-group case is the likely rank degeneracy of the matrix of discriminant vectors when the number of groups is large. Performing simultaneous low-rank and sparse estimation of the matrix of discriminant vectors in the multi-group case is an interesting direction for future research.

Acknowledgments

We thank the Editor, the Associate Editor and two anonymous referees for the valuable comments that helped improve this manuscript. IG was supported by NSF grant DMS-1712943. TW was supported by the National Cancer Institute grant U01-CA057030. All the plots are generated using ggplot2 [52].

Appendix

Appendix A. Implementation details

In this section we describe implementation details for the methods considered in Section 4.1. We use the R package `JGL` [10] to implement sparse QDA of Le and Hastie [30]; R package `MGSDA` [14] to implement sparse LDA [15, 35]; R package `grpreg` [4] to implement logistic regression with pairwise interactions and lasso penalty on the vector of coefficients; R package `RidgeFusion` [41] to implement ridge fusion for joint estimation of precision matrices [42]; R package `sparsediscrim` to implement regularized discriminant analysis [13]. We found no available R code for sparse QDA of Li and Shao [31], and implemented the method ourselves. We use the R package `DAP` [51] to implement the proposed discriminant analysis via projections.

For logistic regression, we select the tuning parameter using BIC option in the `grpreg`. For ridge fusion, we use the default selection in `RidgeFusion` with 5 folds. For Li and Shao [31], we use the proposed bisection procedure with the maximal interval length of 0.05. For all other methods, we use 5-fold cross-validation to minimize misclassification error rate.

Appendix B. Bounds on misclassification error rates through J-divergence

Let $P_e = \pi_1 P_{e1} + \pi_2 P_{e2}$ be the Bayes error rate, where for $g \in \{1, 2\}$, P_{eg} is the probability of incorrectly assigning a new observation x into class g, and π_g are prior class probabilities. To our knowledge, the exact form of P_e is not available for discriminant analysis unless $\Sigma_1 = \Sigma_2$. However, Kadota and Shepp [25] show that it satisfies

$$2^{-1}\min(\pi_1,\pi_2)\exp(-J/8) \le P_e \le \sqrt{\pi_1\pi_2}(J/4)^{-1/4}, \tag{A.1}$$

where J is the divergence between class distributions [33]. Let $X_1 \sim F_1$ and $X_2 \sim F_2$, then $J = J(X_1, X_2) = KL(X_1 \parallel X_2) + KL(X_2 \parallel X_1)$, where $KL(X_1 \parallel X_2)$ is the Kullback–Leibler divergence between probability distributions F_1 and F_2

While the bound (A.1) is loose, it shows that in general larger values of J-divergence lead to smaller misclassification error rates. For the Bayes QDA rule, inequality (A.1) holds with J-divergence between original class distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$. The population version of the proposed DAP rule is also a Bayes rule, but applied to $\mathcal{N}(\Psi^\top\mu_1, \Psi^\top\Sigma_1)$ and $\mathcal{N}(\Psi^\top\mu_2, \Psi^\top\Sigma_2)$. Hence, (A.1) can be used to bound the error of proposed approach by using J-divergence between the projected class distributions. Finally, the projection based on LDA rule leads to class distributions $\mathcal{N}(V^\top\mu_1, V^\top\Sigma_1)$ and $\mathcal{N}(V^\top\mu_2, V^\top\Sigma_2)$ for $V = (\pi_1\Sigma_1 + \pi_2\Sigma_2)^{-1}\delta$, however (A.1) cannot be applied to LDA since it still uses a linear rule on projected data rather than a quadratic rule. Since we can use J-divergence to characterize the relative difference before class distributions before and after projections, we can bound misclassification error rates for both QDA and proposed DAP. We further obtain an explicit form of J-divergence for the original class distributions (QDA), the class distributions induced by proposed projections approach (DAP) and the class distributions after applying LDA-based projection as in Figure 1 in case $\pi_1 = \pi_2 = 1/2$.

Proposition 5. *Let $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$, $X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ and $\delta = \mu_1 - \mu_2$. Let $\Psi = [\psi_1\ \psi_2]$ with $\psi_1 = \Sigma_1^{-1}\delta$ and $\psi_2 = \Sigma_2^{-1}\delta$, $V = 2(\Sigma_1 + \Sigma_2)^{-1}\delta$. Then*

$$\begin{aligned}
J_{Bayes} &= J(X_1, X_2) = \frac{1}{2}\delta^\top(\Sigma_1^{-1} + \Sigma_2^{-1})\delta + \frac{1}{2}\operatorname{tr}(\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2) - p,\\
J_{DAP} &= J(X_1^\top\Psi, X_2^\top\Psi) = \frac{1}{2}\delta^\top(\Sigma_1^{-1} + \Sigma_2^{-1})\delta + \frac{1}{2}\frac{A_3}{A_1A_2}(A_1 + A_2) - 2,\\
J_{LDA} &= J(X_1^\top V, X_2^\top V) = \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}(\delta\delta^\top + \Sigma_1)(\Sigma_1+\Sigma_2)^{-1}\delta}{(\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta} + \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}(\delta\delta^\top + \Sigma_2)(\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta} - 1,
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \delta^\top\Sigma_1^{-1}\delta\,\delta^\top\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\delta - (\delta^\top\Sigma_2^{-1}\delta)^2, \quad A_2 = \delta^\top\Sigma_2^{-1}\delta\,\delta^\top\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}\delta - (\delta^\top\Sigma_1^{-1}\delta)^2,\\
A_3 &= \delta^\top\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}\delta\,\delta^\top\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\delta - \delta^\top\Sigma_1^{-1}\delta\,\delta^\top\Sigma_2^{-1}\delta.
\end{aligned}$$

Proposition 5 reveals that J_{Bayes} and J_{DAP} in general differ in covariance terms, confirming that the rules are not always identical and on a population level DAP does lead to the loss of discriminatory power. Proposition 5 allows to calculate the differences in J divergences exactly for given values of δ, Σ_1 and Σ_2, and as a result assess their effect on error bounds in (A.1). On the other hand, the first term in J_{DAP} is identical to the first term in J_{Bayes}, which captures the mean differences. This is not the case for LDA-induced projection, thus supporting that DAP performs better than LDA when $\Sigma_1 \neq \Sigma_2$.

Proof for Proposition 5. From Kullback [27]

$$KL(X_1\|X_2) = \frac{1}{2}\Big\{\delta^\top\Sigma_2^{-1}\delta + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1 - I_p)\Big\} + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|},$$

hence

$$J_{\mathrm{Bayes}} = J(X_1, X_2) = \frac{1}{2}\delta^\top(\Sigma_1^{-1} + \Sigma_2^{-1})\delta + \frac{1}{2}\mathrm{tr}(\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2) - p.$$

Since for $g \in \{1,2\}$, $\Psi^\top X_g \sim \mathcal{N}(\Psi^\top\mu_g, \Psi^\top\Sigma_g\Psi)$, it follows

$$\begin{aligned} J_{DAP} \quad = \quad & \frac{1}{2}\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})\delta(\Psi^\top\Sigma_1\Psi)^{-1}\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})^\top\delta + \frac{1}{2}\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})\delta(\Psi^\top\Sigma_2\Psi)^{-1}\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})^\top\delta \\ & + \frac{1}{2}\mathrm{tr}\{(\Psi^\top\Sigma_2\Psi)^{-1}(\Psi^\top\Sigma_1\Psi)\} + \frac{1}{2}\mathrm{tr}\{(\Psi^\top\Sigma_1\Psi)^{-1}(\Psi^\top\Sigma_2\Psi)\} - 2. \end{aligned}$$

First, we simplify the term $\mathrm{tr}\{(\Psi^\top\Sigma_2\Psi)^{-1}(\Psi^\top\Sigma_1\Psi)\}$. Since

$$(\Psi^\top\Sigma_2\Psi)^{-1} = \frac{1}{A_2}\begin{pmatrix} \psi_2^\top\Sigma_2\psi_2 & -\psi_1^\top\Sigma_2\psi_2 \\ -\psi_2^\top\Sigma_2\psi_1 & \psi_1^\top\Sigma_2\psi_1 \end{pmatrix} = \frac{1}{A_2}\begin{pmatrix} \delta^\top\Sigma_2^{-1}\delta & -\delta^\top\Sigma_1^{-1}\delta \\ -\delta^\top\Sigma_1^{-1}\delta & \delta^\top\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}\delta \end{pmatrix},$$

we have

$$(\Psi^\top\Sigma_2\Psi)^{-1}(\Psi^\top\Sigma_1\Psi) = \frac{1}{A_2}\begin{pmatrix} 0 & -A_1 \\ A_2 & A_3 \end{pmatrix},$$

leading to $\mathrm{tr}\{(\Psi^\top\Sigma_2\Psi)^{-1}(\Psi^\top\Sigma_1\Psi)\} = A_3/A_2$. Thus,

$$\frac{1}{2}\mathrm{tr}\{(\Psi^\top\Sigma_2\Psi)^{-1}(\Psi^\top\Sigma_1\Psi)\} + \frac{1}{2}\mathrm{tr}\{(\Psi^\top\Sigma_1\Psi)^{-1}(\Psi^\top\Sigma_2\Psi)\} = \frac{1}{2}\Big(\frac{A_3}{A_2} + \frac{A_3}{A_1}\Big) = \frac{1}{2}\frac{A_3}{A_1A_2}(A_1 + A_2).$$

Now we simplify the term $\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})\delta(\Psi^\top\Sigma_2\Psi)^{-1}\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})^\top\delta$. Using the expression for $(\Psi^\top\Sigma_2\Psi)^{-1}$ from above

$$\begin{aligned} \delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})\delta(\Psi^\top\Sigma_2\Psi)^{-1}\delta^\top(\Sigma_1^{-1},\Sigma_2^{-1})^\top\delta &= \frac{1}{A_2}(\delta^\top\Sigma_1^{-1}\delta,\ \delta^\top\Sigma_2^{-1}\delta)\begin{pmatrix} \delta^\top\Sigma_2^{-1}\delta & -\delta^\top\Sigma_1^{-1}\delta \\ -\delta^\top\Sigma_1^{-1}\delta & \delta^\top\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}\delta \end{pmatrix}\begin{pmatrix} \delta^\top\Sigma_1^{-1}\delta \\ \delta^\top\Sigma_2^{-1}\delta \end{pmatrix} \\ &= \frac{1}{A_2}A_2\delta^\top\Sigma_2^{-1}\delta = \delta^\top\Sigma_2^{-1}\delta. \end{aligned}$$

Using similar approach with $(\Psi^\top\Sigma_2\Psi)^{-1}$ and combining the above leads to

$$J_{DAP} = \frac{1}{2}\delta^\top(\Sigma_1^{-1} + \Sigma_2^{-1})\delta + \frac{1}{2}\frac{A_3}{A_1A_2}(A_1 + A_2) - 2.$$

In contrast, for the LDA-based projection, we have

$$\begin{aligned} KL(V^\top X_1\|V^\top X_2) \quad = \quad & \frac{1}{2}\Big\{\delta^\top V(V^\top\Sigma_2V)^{-1}V^\top\delta + \ln\frac{|V^\top\Sigma_2V|}{|V^\top\Sigma_1V|}\Big\} + \frac{1}{2}\mathrm{tr}\Big\{(V^\top\Sigma_2V)^{-1}(V^\top\Sigma_1V) - 1\Big\} \\ = \quad & \frac{1}{2}\frac{[\delta^\top(\Sigma_1+\Sigma_2)^{-1}\delta]^2}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta} + \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta} \\ & + \frac{1}{2}\ln\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta} - \frac{1}{2}. \end{aligned}$$

By the definition of J-divergence, we have

$$\begin{aligned}
J_{LDA} &= \frac{1}{2}\left[\delta^\top V\{(V^\top\Sigma_2 V)^{-1} + (V^\top\Sigma_1 V)^{-1}\}V^\top\delta\right] + \frac{1}{2}\mathrm{tr}\left\{(V^\top\Sigma_2 V)^{-1}(V^\top\Sigma_1 V) + (V^\top\Sigma_1 V)^{-1}(V^\top\Sigma_2 V) - 2\right\} \\
&= \frac{1}{2}\frac{[\delta^\top(\Sigma_1+\Sigma_2)^{-1}\delta]^2}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta} + \frac{1}{2}\frac{[\delta^\top(\Sigma_1+\Sigma_2)^{-1}\delta]^2}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta} \\
&\quad + \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta} + \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta} - 1 \\
&= \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}[\delta\delta^\top+\Sigma_1](\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\delta} + \frac{1}{2}\frac{\delta^\top(\Sigma_1+\Sigma_2)^{-1}[\delta\delta^\top+\Sigma_2](\Sigma_1+\Sigma_2)^{-1}\delta}{\delta^\top(\Sigma_1+\Sigma_2)^{-1}\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\delta} - 1.
\end{aligned}$$

□

Appendix C. Proofs of propositions

Proof of Proposition 2. From Gaynanova et al. [15], $\widetilde{v}(\lambda) = \mathrm{argmin}_v\, L_1(v,\lambda)$, where

$$L_1(v,\lambda) = v^\top(n_1 S_1 + n_2 S_2)\,v/(2n) + n_1 n_2 d^\top v v^\top d/(2n^2) - n_1^{1/2} n_2^{1/2} d^\top v/n + \lambda\|v\|_1.$$

From (6), $\{\widehat{v_1}(\lambda), \widehat{v_2}(\lambda)\} = \mathrm{argmin}_{v_1,v_2}\, L_2(v_1,v_2,\lambda)$, where

$$L_2(v_1,v_2,\lambda) = (v_1^\top S_1 v_1 + v_2^\top S_2 v_2)/2 + \left(n_2 n^{-1} d^\top v_1 - 1\right)^2/2 + \left(n_1 n^{-1} d^\top v_2 - 1\right)^2/2 + \lambda\sum_{j=1}^p (v_{1j}^2 + v_{2j}^2)^{1/2}.$$

Under the constraint $(n/n_1)^{1/2}v_1 = (n/n_2)^{1/2}v_2 = v$, this leads to $\widehat{v}(\lambda) = \mathrm{argmin}_v\, L_2(v,\lambda)$, where using $c = (n_1/n)^{1/2} + (n_2/n)^{1/2}$,

$$L_2(v,\lambda) = v^\top(n_1 S_1 + n_2 S_2)\,v/(2n) + n_1 n_2 d^\top v v^\top d/(2n^2) - n_1^{1/2} n_2^{1/2} c d^\top v/n + \lambda\|v\|_1.$$

Furthermore,

$$L_1(v/c,\lambda/c) = c^{-2}\left\{v^\top(n_1 S_1 + n_2 S_2)\,v/(2n) + n_1 n_2 d^\top v v^\top d/(2n^2) - n_1^{1/2} n_2^{1/2} c d^\top v/n + \lambda\|v\|_1\right\} = c^{-2} L_2(v,\lambda).$$

Since for any $c > 0$, $\mathrm{argmin}_x f(x/c) = c\{\mathrm{argmin}_x f(x)\}$, it follows that $c\widetilde{v}(\lambda/c) = \widehat{v}(\lambda)$. □

Proof of Proposition 3. Since $\ln(|\Sigma_2|/|\Sigma_1|)$ and $2\ln(\pi_1/\pi_2)$ are present in both rules, it remains to show the equivalence of the quadratic term, the linear term and the remaining constants. Substituting $x = x - \mu + \mu$ in the Bayes rule (7) leads to

$$\begin{aligned}
x^\top(\Sigma_2^{-1} - \Sigma_1^{-1})x &= (x-\mu)^\top(\Sigma_2^{-1} - \Sigma_1^{-1})(x-\mu) + 2(x-\mu)^\top(\Sigma_2^{-1} - \Sigma_1^{-1})\mu + \mu^\top(\Sigma_2^{-1} - \Sigma_1^{-1})\mu, \\
-2x^\top(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) &= -2(x-\mu)^\top(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) - 2\mu^\top(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1).
\end{aligned}$$

From the above, the quadratic term in $x-\mu$ is the same as stated in the Proposition, hence it remains to consider the linear terms and the constants.

Consider the linear terms in $x-\mu$ from the above. Recall that $\delta = \mu_1 - \mu_2$, therefore

$$\begin{aligned}
2(x-\mu)^\top(\Sigma_2^{-1} - \Sigma_1^{-1})\mu - 2(x-\mu)^\top(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) &= 2(x-\mu)^\top\{\Sigma_2^{-1}(\mu-\mu_2) - \Sigma_1^{-1}(\mu-\mu_1)\} \\
&= 2(x-\mu)^\top(\pi_1\Sigma_2^{-1}\delta + \pi_2\Sigma_1^{-1}\delta),
\end{aligned}$$

which is the same as the linear term in the statement of the proposition.

Finally, we complete the proof by showing the equivalence of the remaining constants:

$$\begin{aligned}
&\mu^\top(\Sigma_2^{-1} - \Sigma_1^{-1})\mu - 2\mu^\top(\Sigma_2^{-1}\mu_2 - \Sigma_2^{-1}\mu_1) - \mu_1^\top\Sigma_1^{-1}\mu_1 + \mu_2^\top\Sigma_2^{-1}\mu_2 \\
&= (\mu^\top\Sigma_2^{-1}\mu - 2\mu^\top\Sigma_2^{-1}\mu_2 + \mu_2^\top\Sigma_2^{-1}\mu_2) - (\mu^\top\Sigma_1^{-1}\mu - 2\mu^\top\Sigma_1^{-1}\mu_1 + \mu_1^\top\Sigma_1^{-1}\mu_1) \\
&= \pi_1^2\delta^\top\Sigma_2^{-1}\delta - \pi_2^2\delta^\top\Sigma_1^{-1}\delta.
\end{aligned}$$

□

Proof of Proposition 4. Since $\Psi^\top X|Y = g \sim \mathcal{N}(\Psi^\top \mu_g, \Psi^\top \Sigma_g \Psi)$, from Proposition 3 the Bayes rule applied to $\Psi^\top x$ has the form

$$\begin{aligned}(x-\mu)^\top \Psi\Big\{(\Psi^\top\Sigma_2\Psi)^{-1} - (\Psi^\top\Sigma_1\Psi)^{-1}\Big\}\Psi^\top(x-\mu) + \ln\Big(|\Psi^\top\Sigma_2\Psi|/|\Psi^\top\Sigma_1\Psi|\Big) \\ + 2(x-\mu)^\top\Big\{\pi_1\Psi(\Psi^\top\Sigma_2\Psi)^{-1}\Psi^\top\delta + \pi_2\Psi(\Psi^\top\Sigma_1\Psi)^{-1}\Psi^\top\delta\Big\} \\ + \pi_1^2\delta^\top\Psi(\Psi^\top\Sigma_2\Psi)^{-1}\Psi^\top\delta - \pi_2^2\delta^\top\Psi(\Psi^\top\Sigma_1\Psi)^{-1}\Psi^\top\delta + 2\ln(\pi_1/\pi_2) > 0.\end{aligned} \tag{A.1}$$

Since

$$(\Psi^\top\Sigma_1\Psi)^{-1} = \frac{1}{\psi_1^\top\Sigma_1\psi_1\psi_2^\top\Sigma_1\psi_2 - (\psi_1^\top\Sigma_1\psi_2)^2}\begin{pmatrix} \psi_2^\top\Sigma_1\psi_2 & -\psi_1^\top\Sigma_1\psi_2 \\ -\psi_2^\top\Sigma_1\psi_1 & \psi_1^\top\Sigma_1\psi_1 \end{pmatrix}.$$

it follows that

$$\Psi(\Psi^\top\Sigma_1\Psi)^{-1}\Psi^\top = \frac{\psi_1\psi_2^\top\Sigma_1\psi_2\psi_1^\top - \psi_2\psi_2^\top\Sigma_1\psi_1\psi_1^\top + \psi_2\psi_1^\top\Sigma_1\psi_1\psi_2^\top - \psi_1\psi_1^\top\Sigma_1\psi_2\psi_2^\top}{\psi_1^\top\Sigma_1\psi_1\psi_2^\top\Sigma_1\psi_2 - (\psi_1^\top\Sigma_1\psi_2)^2}.$$

Recall that $\psi_1 = c_1\Sigma_1^{-1}\delta$, and substituting $\delta = c_1^{-1}\Sigma_1\psi_1$ into the above equation leads to

$$\Psi(\Psi^\top\Sigma_1\Psi)^{-1}\Psi\delta = \frac{c_1^{-1}\psi_1\Big\{\psi_2^\top\Sigma_1\psi_2\psi_1^\top\Sigma_1\psi_1 - (\psi_1^\top\Sigma_1\psi_2)^2\Big\}}{\psi_1^\top\Sigma_1\psi_1\psi_2^\top\Sigma_1\psi_2 - (\psi_1^\top\Sigma_1\psi_2)^2} = c_1^{-1}\psi_1 = \Sigma_1^{-1}\delta.$$

Similarly, $\Psi(\Psi^\top\Sigma_2\Psi)^{-1}\Psi^\top\delta = \Sigma_2^{-1}\delta$. Substituting these into (A.1) completes the proof. □

Appendix D. Proofs of the main theorems

We will use the following quantities throughout the proofs:

$$\gamma = 1 + \max\Big(\pi_1\pi_2^{-1}\|\Sigma_{1AA}^{-1/2}\Sigma_{2AA}\Sigma_{1AA}^{-1/2}\|_2, \pi_2\pi_1^{-1}\|\Sigma_{2AA}^{-1/2}\Sigma_{1AA}\Sigma_{2AA}^{-1/2}\|_2\Big), \tag{B.1}$$

$$\begin{aligned}\Sigma_{gA^{\complement}A^{\complement}:A} &= \Sigma_{gA^{\complement}A^{\complement}} - \Sigma_{gA^{\complement}A}\Sigma_{gAA}^{-1}\Sigma_{gA^{\complement}A} \quad (g=1,2),\\ \Sigma_{d_1} &= \Sigma_{1A^{\complement}A^{\complement}:A} + \pi_1\pi_2^{-1}\Big(\Sigma_{2A^{\complement}A^{\complement}} + \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{2AA}\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}} - \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{2AA^{\complement}} - \Sigma_{2A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}}\Big),\\ \Sigma_{d_2} &= \Sigma_{2A^{\complement}A^{\complement}:A} + \pi_2\pi_1^{-1}\Big(\Sigma_{1A^{\complement}A^{\complement}} + \Sigma_{2A^{\complement}A}\Sigma_{2AA}^{-1}\Sigma_{1AA}\Sigma_{2AA}^{-1}\Sigma_{2AA^{\complement}} - \Sigma_{2A^{\complement}A}\Sigma_{2AA}^{-1}\Sigma_{1AA^{\complement}} - \Sigma_{1A^{\complement}A}\Sigma_{2AA}^{-1}\Sigma_{2AA^{\complement}}\Big).\end{aligned} \tag{B.2}$$

The quantities in (B.2) can be viewed as conditional variance terms, their origin is made precise in Lemma 2. Let $\sigma^2_{gjj:A} = e_j^\top\Sigma_{gA^{\complement}A^{\complement}:A}e_j$ and $\sigma^2_{jdg} = e_j^\top\Sigma_{d_g}e_j$ be the diagonal elements of corresponding matrices. Under Assumption 4, $\sigma_{gjj:A}$, σ_{jdg} and γ can be treated as constants.

We define the oracle $(\tilde v_{1A}, \tilde v_{2A})$ as the solution to

$$\underset{v_1,v_2\in\mathbb{R}^s}{\text{minimize}}\Big\{n_1^{-1}\|X_{1A}v_1 - \iota_{n_1}\|_2^2/2 + n_2^{-1}\|X_{2A}v_2 + \iota_{n_2}\|_2^2/2 + \lambda\sum_{j=1}^s(v_{1j}^2 + v_{2j}^2)^{1/2}\Big\}, \tag{B.3}$$

and let $\tilde u_A = (\tilde u_{1A}, \tilde u_{2A})$ be the subgradient of $\sum_{j=1}^s(v_{1j}^2+v_{2j}^2)^{1/2}$ evaluated at $(\tilde v_{1A}, \tilde v_{2A})$ such that $\tilde u_{Aj} = \tilde v_{Aj}/\|\tilde v_{Aj}\|_2$ if $\|\tilde v_{Aj}\|_2 \neq 0$, and $\tilde u_{Aj} \in \{u : \|u\|_2 \leq 1\}$ if $\|\tilde v_{Aj}\|_2 = 0$.

Theorem 3 (Equivalent to Theorem 1). *Let Assumptions 1–3 hold. Let the sample sizes satisfy*

$$\min(n_1,n_2) \gtrsim \max_{g=1,2}\|\Sigma_{gAA}^{-1}\|_2 \max_{g=1,2;\ j\in A^{\complement}}(\sigma^2_{gjj:A} \vee \sigma^2_{jdg})s\ln\{(p-s)\eta^{-1}\},$$

for some $\eta \in (0,1)$, and the tuning parameter satisfy

$$\lambda \gtrsim \max_{g=1,2;\ j\in A^{\complement}}(\sigma^2_{gjj:A} \vee \sigma^2_{jdg})\Big[n^{-1}\ln\{(p-s)\eta^{-1}\}\Big]^{1/2}.$$

Then $\Pr(\widehat{A} \subseteq A) \geq 1 - \eta$.

Proof. Using the results of Section 2.3,

$$[\widehat{v}_1\ \widehat{v}_2] = \operatorname*{argmin}_{v_1\in\mathbb{R}^p, v_2\in\mathbb{R}^p} \Big\{\widehat{L}_{\psi_1}(v_1) + \widehat{L}_{\psi_2}(v_2) + \lambda \sum_{j=1}^{p}(v_{1j}^2 + v_{2j}^2)^{1/2}\Big\},$$

$$2\{\widehat{L}_{\psi_1}(v_1) + \widehat{L}_{\psi_2}(v_2)\} = v_1^\top S_1 v_1 + v_2^\top S_2 v_2 + \left(n^{-1}n_2 d^\top v_1 - 1\right)^2 + \left(n^{-1}n_1 d^\top v_2 - 1\right)^2.$$

Let $\rho_1 = n_1/n$ and $\rho_2 = n_2/n$. The optimality conditions stated in Chapter 5 of [3] lead to

$$\begin{aligned}
(S_{1AA} + \rho_2^2 d_A d_A^\top)\widehat{v}_{1A} + (S_{1AA^\complement} + \rho_2^2 d_A d_{A^\complement}^\top)\widehat{v}_{1A^\complement} - \rho_2 d_A &= -\lambda u_{1A},\\
(S_{2AA} + \rho_1^2 d_A d_A^\top)\widehat{v}_{2A} + (S_{2AA^\complement} + \rho_1^2 d_A d_{A^\complement}^\top)\widehat{v}_{2A^\complement} - \rho_1 d_A &= -\lambda u_{2A},\\
(S_{1A^\complement A} + \rho_2^2 d_{A^\complement} d_A^\top)\widehat{v}_{1A} + (S_{1A^\complement A^\complement} + \rho_2^2 d_{A^\complement} d_{A^\complement}^\top)\widehat{v}_{1A^\complement} - \rho_2 d_{A^\complement} &= -\lambda u_{1A^\complement},\\
(S_{2A^\complement A} + \rho_1^2 d_{A^\complement} d_A^\top)\widehat{v}_{2A} + (S_{2A^\complement A^\complement} + \rho_1^2 d_{A^\complement} d_{A^\complement}^\top)\widehat{v}_{2A^\complement} - \rho_2 d_{A^\complement} &= -\lambda u_{2A^\complement},
\end{aligned}$$

where u is defined in (**??**). Consider $\widehat{v}_1 = (\tilde{v}_{1A}, 0_{p-s})$, $\widehat{v}_2 = (\tilde{v}_{2A}, 0_{p-s})$, where $\tilde{v}_{1A}$, $\tilde{v}_{2A}$ are the solutions to the oracle problem (B.3). From the above optimality conditions, it is sufficient to have

$$\left\|(S_{1A^\complement A} + \rho_2^2 d_{A^\complement} d_A^\top)\tilde{v}_{1A} - \rho_2 d_{A^\complement}\ ,\ (S_{2A^\complement A} + \rho_1^2 d_{A^\complement} d_A^\top)\tilde{v}_{2A} - \rho_1 d_{A^\complement}\right\|_{\infty,2} < \lambda$$

for $\widehat{V} = [\widehat{v}_1\ \widehat{v}_2]$ to be the solution to (6), which leads to $\widehat{A} \subseteq A$. We next show that the above inequality holds with high probability under the stated conditions.

Using the form of $\widetilde{v}_{1A}$ (Theorem 5) and Sherman–Morrison identity, we find

$$\begin{aligned}
(S_{1A^\complement A} + \rho_2^2 d_{A^\complement} d_A^\top)\tilde{v}_{1A} - \rho_2 d_{A^\complement} &= S_{1A^\complement A}\rho_2 S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} + \rho_2^2 d_{A^\complement} d_A \rho_2^\top S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&\quad - \lambda S_{1A^\complement A}\left(S_{1AA} + \rho_2^2 d_A d_A^\top\right)^{-1} \tilde{u}_{1A} - \lambda \rho_2^2 d_{A^\complement} d_A^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top\right)^{-1} \tilde{u}_{1A} - \rho_2 d_{A^\complement}\\
&= \rho_2\left(S_{1A^\complement A} S_{1AA}^{-1} d_A - d_{A^\complement}\right)(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} - \lambda S_{1A^\complement A} S_{1AA}^{-1}\tilde{u}_{1A}\\
&\quad + \lambda\rho_2^2 S_{1A^\complement A} S_{1AA}^{-1} d_A d_A^\top S_{1AA}^{-1}\tilde{u}_{1A}(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&\quad - \lambda\rho_2^2 d_{A^\complement} d_A^\top S_{1AA}^{-1}\tilde{u}_{1A}(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&= \rho_2\left(S_{1A^\complement A} S_{1AA}^{-1} d_A - d_{A^\complement}\right)(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} - \lambda S_{1A^\complement A} S_{1AA}^{-1}\tilde{u}_{1A}\\
&\quad + \rho_2^2\lambda(S_{1A^\complement A} S_{1AA}^{-1} d_A - d_{A^\complement}) d_A^\top S_{1AA}^{-1}\tilde{u}_{1A}(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}.
\end{aligned}$$

Using normality, there exists $U_1 \in \mathbb{R}^{p\times(n_1-1)}$ with columns $u_{1,i} \sim \mathcal{N}(0, \Sigma_1)$ such that $(n_1 - 1)S_1 = U_1 U_1^\top$. Similar to [17], let $E_{d1} = d_{A^\complement} - \Sigma_{1A^\complement A}\Sigma_{1AA}^{-1} d_A$, $E_{U1} = U_{1A^\complement} - \Sigma_{1A^\complement A}\Sigma_{1AA}^{-1} U_{1A}$. Then

$$\begin{aligned}
S_{1A^\complement A} S_{1AA}^{-1} &= (n_1 - 1)^{-1} U_{1A^\complement} U_{1A}^\top S_{1AA}^{-1}\\
&= (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} + (n_1 - 1)^{-1}\Sigma_{1A^\complement A}\Sigma_{1AA}^{-1} U_{1A} U_{1A}^\top S_{1,AA}^{-1}\\
&= \Sigma_{1A^\complement A}\Sigma_{1AA}^{-1} + (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1},
\end{aligned}$$

and $S_{1A^\complement A} S_{1AA}^{-1} d_A - d_{A^\complement} = (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} d_A - E_{d1}$. Combining the above two displays gives

$$\begin{aligned}
(S_{1A^\complement A} + \rho_2^2 d_{A^\complement} d_A^\top)\tilde{v}_{1A} - \rho_2 d_{A^\complement} &= -\lambda\Sigma_{1A^\complement A}\Sigma_{1AA}^{-1}\tilde{u}_{1A} - \lambda(n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1}\tilde{u}_{1A}\\
&\quad + (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} d_A \rho_2 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} - E_{d1}\rho_2(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&\quad + \lambda(n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} d_A \rho_2^2 d_A^\top S_{1AA}^{-1}\tilde{u}_{1A}(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&\quad - \lambda E_{d1}\rho_2^2 d_A^\top S_{1AA}^{-1}\tilde{u}_{1A}(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&= -\lambda\Sigma_{1A^\complement A}\Sigma_{1AA}^{-1}\tilde{u}_{1A} + (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} d_A \rho_2 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&\quad - E_{d1}\rho_2(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} - \lambda E_{d1}\rho_2^2 d_A^\top S_{1AA}^{-1}\tilde{u}_{1A}(1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1}\\
&\quad - \lambda(n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1}(I + \rho_2^2 d_A d_A^\top S_{1AA}^{-1})^{-1}\tilde{u}_{1A}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
(S_{2A^\complement A}\rho_1^2 d_{A^\complement} d_A^\top)\tilde{v}_{2A} - \rho_1 d_{A^\complement} &= -\lambda\Sigma_{2A^\complement A}\Sigma_{2AA}^{-1}\tilde{u}_{2A} + (n_2 - 1)^{-1} E_{U2} U_{2A}^\top S_{2AA}^{-1} d_A \rho_1 (1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1}\\
&\quad - E_{d2}\rho_1(1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1} - \lambda E_{d2}\rho_1^2 d_A^\top S_{2AA}^{-1}\tilde{u}_{2A}(1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1}\\
&\quad - \lambda(n_2 - 1)^{-1} E_{U2} U_{2A}^\top S_{2AA}^{-1}(I + \rho_1^2 d_A d_A^\top S_{2AA}^{-1})^{-1}\tilde{u}_{2A}.
\end{aligned}$$

Therefore, using the triangle inequality,

$$\begin{aligned}&\left\|(S_{1A^{\complement}A}+\rho_2^2 d_{A^{\complement}}d_A^\top)\tilde{v}_{1A}-\rho_2 d_{A^{\complement}}\ ,\ (S_{2A^{\complement}A}+\rho_1^2 d_{A^{\complement}}d_A^\top)\tilde{v}_{2A}-\rho_1 d_{A^{\complement}}\right\|_{\infty,2}\\ &\quad\le \lambda\|\Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\tilde{u}_{1A},\Sigma_{2A^{\complement}A}\Sigma_{2AA}^{-1}\tilde{u}_{2A}\|_{\infty,2}+I_1+I_2+I_3+I_4,\end{aligned}$$

where

$$\begin{aligned}I_1 &= \|\rho_2(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}E_{d1},\rho_1(1+\rho_1^2 d_A^\top S_{2AA}^{-1}d_A)^{-1}E_{d2}\|_{\infty,2},\\ I_2 &= \left\|(n_1-1)^{-1}\frac{\rho_2 E_{U1}U_{1A}^\top S_{1AA}^{-1}d_A}{1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A},(n_2-1)^{-1}\frac{\rho_1 E_{U2}U_{2A}^\top S_{2AA}^{-1}d_A}{1+\rho_1^2 d_A^\top S_{2AA}^{-1}d_A}\right\|_{\infty,2},\\ I_3 &= \left\|\frac{E_{U1}U_{1A}^\top S_{1AA}^{-1}}{n_1-1}(I+\rho_2^2 d_A d_A^\top S_{1AA}^{-1})^{-1}\tilde{u}_{1A},\frac{E_{U2}U_{2A}^\top S_{2AA}^{-1}}{n_2-1}(I+\rho_1^2 d_A d_A^\top S_{2AA}^{-1})^{-1}\tilde{u}_{2A}\right\|_{\infty,2},\\ I_4 &= \left\|\frac{\rho_2^2}{1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A}E_{d1}d_A^\top S_{1AA}^{-1}\tilde{u}_{1A},\frac{\rho_1^2}{1+\rho_1^2 d_A^\top S_{2AA}^{-1}d_A}E_{d2}d_A^\top S_{2AA}^{-1}\tilde{u}_{2A}\right\|_{\infty,2}.\end{aligned}$$

By the irrepresentability condition (Assumption 3), there exists $\alpha\in(0,1]$ such that

$$\|\Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\tilde{u}_{1A},\Sigma_{2A^{\complement}A}\Sigma_{2AA}^{-1}\tilde{u}_{2A}\|_{\infty,2}\le 1-\alpha.$$

To conclude the proof, it is sufficient to show that with probability at least $1-\eta$, each $I_k\le\lambda\alpha/4$ for all $k\in\{1,\dots,4\}$. Next, we consider each of these four terms separately.

1. Show $I_1\le\lambda\alpha/4$ with probability at least $1-\eta/4$. By Lemma 2, $e_j^\top E_{dg}\sim\mathcal{N}(0,\sigma_{jdg}^2/n_g)$. Applying the standard normal concentration inequality, there exists a constant $C>0$ such that

$$\Pr\Big(\bigcap_{j\in A^{\complement}}\Big\{|e_j^\top E_{dg}|\ge C\max_{j\in A^{\complement}}\sigma_{jdg}\Big[n_g^{-1}\ln\{(p-s)\eta^{-1}\}\Big]^{1/2}\Big\}\Big)\le\eta/4.$$

Since

$$\begin{aligned}&\|\rho_2(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}E_{d1},\rho_1(1+\rho_1^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}E_{d2}\|_{\infty,2}\\ &\quad\le\ \sqrt{2}\max\Big\{\rho_2(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}\|E_{d1}\|_\infty,\rho_1(1+\rho_1^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}\|E_{d2}\|_\infty\Big\}\\ &\quad\le\ \sqrt{2}\max(\|E_{d1}\|_\infty,\|E_{d2}\|_\infty),\end{aligned}$$

it follows that there exists a constant $C>0$ such that

$$\Pr\Big(I_1\ge C\max_{g=1,2;\ j\in A^{\complement}}\sigma_{jdg}\Big[\ln\{(p-s)\eta^{-1}\}/\min(n_1,n_2)\Big]^{1/2}\Big)\le\eta/4.$$

Therefore, $I_1\le\lambda\alpha/4$ with probability at least $1-\eta/4$ under the conditions of the theorem.

2. Show $I_2\le\lambda\alpha/4$ with probability at least $1-\eta/4$. By Lemma 2, $E_{Ug}\sim\mathcal{N}(0,\Sigma_{gA^{\complement}A^{\complement}:A}\otimes I_{n_g-1})$ for $g\{1,2\}$, and is independent of U_{gA} and d. Hence,

$$\begin{aligned}&\rho_2(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}e_j^\top(n_1-1)^{-1}E_{U1}U_{1A}^\top S_{1AA}^{-1}d_A|U_{1A},d_A\\ &\qquad\sim\mathcal{N}\Big\{0,\sigma_{1jj:A}^2(n_1-1)^{-1}\rho_2^2 d_A^\top S_{1AA}^{-1}d_A(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-2}\Big\}.\end{aligned}$$

Define $L=(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-2}\rho_2^2 d_A^\top S_{1AA}^{-1}d_A$. Using standard normal concentration inequality, there exists a constant $C>0$ such that conditionally on L, the event

$$\bigcap_{j\in A^{\complement}}\Big\{\rho_2(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}|e_j^\top(n_1-1)^{-1}E_{U1}U_{1A}^\top S_{1AA}^{-1}d_A|\ge C\max_{j\in A^{\complement}}\sigma_{1jj:A}\Big[Ln_1^{-1}\ln\{(p-s)\eta^{-1}\}\Big]^{1/2}\Big\}$$

has probability at most $\eta/4$. Since $L=(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-2}\rho_2^2 d_A^\top S_{1AA}^{-1}d_A\le(1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A)^{-1}\le 1$, it follows that with probability at least $1-\eta/4$

$$\frac{\rho_2}{1+\rho_2^2 d_A^\top S_{1AA}^{-1}d_A}\left\|\frac{E_{U1}U_{1A}^\top S_{1AA}^{-1}d_A}{n_1-1}\right\|_\infty\le C\Big[\max_{j\in A^{\complement}}\sigma_{1jj:A}n_1^{-1}\ln\{(p-s)\eta^{-1}\}\Big]^{1/2}.$$

The case $g = 2$ is similar, leading to the desired bound under the conditions of the theorem.

3. Show $I_3 \le \alpha/4$ with probability at least $1 - \eta/4$. Similar to part 2,

$$\begin{aligned} &e_j^\top (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} (I + \rho_2^2 d_A d_A^\top S_{1AA}^{-1})^{-1} \tilde{u}_{1A} | U_{1A}, \tilde{u}_{1A}, d_A \\ &\quad \sim \mathcal{N}\left(0, (n_1 - 1)^{-1} \sigma^2_{1jj:A} \tilde{u}_{1A}^\top (S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1} S_{1AA} (S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1} \tilde{u}_{1A}\right). \end{aligned}$$

Define $L = \tilde{u}_{1A}^\top (S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1} S_{1AA} (S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1} \tilde{u}_{1A}$. As in part 2, there exists a constant $C > 0$ such that conditionally on L the event

$$\bigcap_{j \in A^\complement} \left\{ |e_j^\top (n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} (I + \rho_2^2 d_A d_A^\top S_{1AA}^{-1})^{-1} \tilde{u}_{1A}| \ge C \max_{j \in A^\complement} \sigma_{1jj:A} \left[L n_1^{-1} \ln\{(p - s)\eta^{-1}\} \right]^{1/2} \right\}$$

has probability at most $\eta/4$. Furthermore,

$$\begin{aligned} L &\le \|\tilde{u}_{1A}\|_2^2 \|(S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1} S_{1AA} (S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1}\|_2 \\ &\le s \|S_{1AA}^{-1/2} (I + \rho_2^2 S_{1AA}^{-1/2} d_A d_A^\top S_{1AA}^{-1/2})^{-2} S_{1AA}^{-1/2}\|_2^2 \\ &\le s \|S_{1AA}^{-1}\|_2, \end{aligned}$$

where in the last inequality we used $\|\tilde{u}_{1A}\|_2^2 + \|\tilde{u}_{2A}\|_2^2 \le s$ by definition of subgradient. By Lemma 3, there exists a constant $C > 0$ such that with probability at least $1 - \eta/4$,

$$\|S_{1AA}^{-1}\|_2 \le \|\Sigma_{1AA}^{-1}\|_2 \left[1 + C \left\{ n_1^{-1} \ln(\eta^{-1}) \right\}^{1/2} \right].$$

Combining the above displays leads to

$$\|(n_1 - 1)^{-1} E_{U1} U_{1A}^\top S_{1AA}^{-1} (I + \rho_2^2 d_A d_A^\top S_{1AA}^{-1})^{-1} \tilde{u}_{gA}\|_\infty \le C \max_{j \in A^\complement} \sigma_{1jj:A} \left[\|\Sigma_{1AA}^{-1}\|_2 n_1^{-1} s \ln\{(p - s)\eta^{-1}\} \right]^{1/2}$$

with probability at least $1 - \eta/4$. The proof for $g = 2$ is similar leading to the desired bound.

4. Show $I_4 \le \alpha/4$ with probability at least $1 - \eta/4$.

By Lemma 2, $e_j^\top E_{dg} \sim \mathcal{N}(0, n_g^{-1} \sigma_{jdg}^2)$, where σ_{jdg} is from Lemma 2. Then

$$\rho_2^2 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} e_j^\top E_{d1} d_A^\top S_{1AA}^{-1} \tilde{u}_{1A} | U_{1A}, \tilde{u}_{1A}, d_A \sim \mathcal{N}\left(0, \frac{\sigma_{jd1}^2 \rho_2^4}{n_1 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^2} \tilde{u}_{1A}^\top S_{1AA}^{-1} d_A d_A^\top S_{1AA}^{-1} \tilde{u}_{1A}\right).$$

Define $L = (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-2} \rho_2^4 \tilde{u}_{1A}^\top S_{1AA}^{-1} d_A d_A^\top S_{1AA}^{-1} \tilde{u}_{1A}$. Using the standard normal concentration inequality there exists a constant $C > 0$ such that conditionally on L, the event

$$\bigcap_{j \in A^\complement} \left\{ \frac{\rho_2^2}{1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A} e_j^\top E_{d1} d_A^\top S_{1AA}^{-1} \tilde{u}_{1A} \ge C \max_{j \in A^\complement} \sigma_{jd1} \left[L n_1^{-1} \ln\{(p - s)\eta^{-1}\} \right]^{1/2} \right\}$$

has probability at most $\eta/4$. Furthermore,

$$\begin{aligned} L = (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-2} \rho_2^4 (\tilde{u}_{1A}^\top S_{1AA}^{-1/2} S_{1AA}^{-1/2} d_A)^2 &\le \rho_2^2 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-2} \rho_2^2 d_A^\top S_{1AA}^{-1} d_A \tilde{u}_{1A}^\top S_{1AA}^{-1} \tilde{u}_{1A} \\ &\le \rho_2^2 \tilde{u}_{1A}^\top S_{1AA}^{-1} \tilde{u}_{1A} \\ &\le s \|S_{1AA}^{-1}\|_2, \end{aligned}$$

where in the last inequality we used $\|\tilde{u}_{1A}\|_2^2 + \|\tilde{u}_{2A}\|_2^2 \le s$ by definition of subgradient. Similar to part 3, this means that there exists a constant $C > 0$ such that

$$\left\| \frac{\rho_2^2}{1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A} E_{d1} d_A^\top S_{1AA}^{-1} \tilde{u}_{1A} \right\|_\infty \ge C \max_{j \in A^\complement} \sigma_{jd1} \left[\|\Sigma_{1AA}^{-1}\|_2 n_1^{-1} s \ln\{(p - s)\eta^{-1}\} \right]^{1/2}$$

with probability at most $\eta/4$. The proof for $g = 2$ is analogous, leading to the desired bound. □

Theorem 4 (Equivalent to Theorem 2)**.** *Assume the conditions of Theorem 3 hold. If in addition*

$$\psi_{\min} \gtrsim \lambda s^{1/2} \max_g \|\Sigma_{g,AA}^{-1}\|_2 (\max_g \delta_A^\top \Sigma_{gAA}^{-1} \delta_A \vee \gamma),$$

then $\Pr(\widehat{A} = A) \ge 1 - \eta$.

Proof of Theorem 4. Consider the oracle solution

$$\tilde{v}_{1A} = \rho_2 S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} - \lambda \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \tilde{u}_{1A},$$
$$\tilde{v}_{2A} = \rho_1 S_{2AA}^{-1} d_A (1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1} - \lambda \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \tilde{u}_{2A};$$

where $\tilde{u}_A$ is the subgradient. To show $\widehat{A} = A$, it is sufficient to show

$$\begin{aligned} &\min_{j \in A} \left\| \rho_2 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} e_j^\top S_{1AA}^{-1} d_A,\ \rho_1 (1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1} e_j^\top S_{2AA}^{-1} d_A \right\|_2 \\ &\quad \geq \lambda \max_{j \in A} \| e_j^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \tilde{u}_{1A},\ e_j^\top \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \tilde{u}_{2A} \|_2. \end{aligned} \tag{B.4}$$

Consider the right-hand side in (B.4)

$$\begin{aligned} &\max_{j \in A} \| e_j^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \tilde{u}_{1A},\ e_j^\top \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \tilde{u}_{2A} \|_2 \\ &= \max_{j \in A} \left[\left\{ e_j^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \tilde{u}_{1A} \right\}^2 + \left\{ e_j^\top \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \tilde{u}_{2A} \right\}^2 \right]^{1/2} \\ &\leq \max_{j \in A} \left\{ \| e_j^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \|_2^2 \| \tilde{u}_{1A} \|_2^2 + \| e_j^\top \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \|_2^2 \| \tilde{u}_{2A} \|_2^2 \right\}^{1/2} \\ &\leq \max_{j \in A} \left\{ \| e_j^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \|_2 \vee \| e_j^\top \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \|_2 \right\} \left(\| \tilde{u}_{1A} \|_2^2 + \| \tilde{u}_{2A} \|_2^2 \right)^{1/2} \\ &\leq \left\{ \| (S_{1AA} + \rho_2^2 d_A d_A^\top)^{-1} \|_2 \vee \| (S_{2AA} + \rho_1^2 d_A d_A^\top)^{-1} \|_2 \right\} s^{1/2}. \end{aligned}$$

Furthermore,

$$\| \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \|_2 = \| S_{1AA}^{-1/2} \left(I + \rho_2^2 S_{1AA}^{-1/2} d_A d_A^\top S_{1AA}^{-1/2} \right)^{-1} S_{1AA}^{-1/2} \|_2 \leq \| S_{1AA}^{-1} \|_2,$$

and similarly $\|(S_{2AA} + \rho_1^2 d_A d_A^\top)^{-1}\|_2 \leq \|S_{2AA}^{-1}\|_2$. Using Lemma 3, with probability at least $1 - \eta$

$$\max_{j \in A} \| e_j^\top \left(S_{1AA} + \rho_2^2 d_A d_A^\top \right)^{-1} \tilde{u}_{1A},\ e_j^\top \left(S_{2AA} + \rho_1^2 d_A d_A^\top \right)^{-1} \tilde{u}_{2A} \|_2 \leq \max_g \| \Sigma_{gAA}^{-1} \|_2 s^{1/2} \left[1 + C \{ s \ln(\eta^{-1}) / \min(n_1, n_2) \}^{1/2} \right].$$

Consider the left-hand side in (B.4). Applying Lemma 1 and Corollary 1, there exist constants C_1, C_2 such that with probability at least $1 - \eta$

$$\begin{aligned} &\min_{j \in A} \left\| \rho_2 (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} e_j^\top \Sigma_{1AA}^{-1} \delta_A,\ \rho_1 (1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1} e_j^\top \Sigma_{2AA}^{-1} \delta_A \right\|_2 \\ &\geq \left[1 + C_1 \max_g \delta_A^\top \Sigma_{gAA}^{-1} \delta_A + C_2 (\max_g \delta_A^\top \Sigma_{gAA}^{-1} \delta_A \vee \gamma) \left\{ s \ln(\eta^{-1}) / \min(n_1, n_2) \right\}^{1/2} \right]^{-1} \min_{j \in A} \left\| \pi_2 e_j^\top S_{1AA}^{-1} d_A,\ \pi_1 e_j^\top S_{2AA}^{-1} d_A \right\|_2. \end{aligned}$$

Furthermore,

$$\begin{aligned} &\min_{j \in A} \left\| \pi_2 e_j^\top S_{1AA}^{-1} d_A,\ \pi_1 e_j^\top S_{2AA}^{-1} d_A \right\|_2 \\ &\quad = \min_{j \in A} \left\{ \pi_2^2 (e_j^\top S_{1AA}^{-1} d_A)^2 + \pi_1^2 (e_j^\top S_{2AA}^{-1} d_A)^2 \right\}^{1/2} \\ &\quad = \min_{j \in A} \left[\pi_2^2 \{ e_j^\top (S_{1AA}^{-1} d_A - \Sigma_{1AA}^{-1} \delta_A + \Sigma_{1AA}^{-1} \delta_A) \}^2 + \pi_1^2 \{ e_j^\top (S_{2AA}^{-1} d_A - \Sigma_{2AA}^{-1} \delta_A + \Sigma_{2AA}^{-1} \delta_A) \}^2 \right]^{1/2} \\ &\quad \geq \min_{j \in A} \left\| \pi_2 e_j^\top \Sigma_{1AA}^{-1} \delta_A,\ \pi_1 e_j^\top \Sigma_{2AA}^{-1} \delta_A \right\|_2 - \max_g \left(\| S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1} \delta_A \|_\infty \right) \\ &\quad = \psi_{\min} - \max_g \left(\| S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1} \delta_A \|_\infty \right), \end{aligned}$$

where in the last inequality we used $\pi_1^2 + \pi_2^2 \leq 1$. Using Lemma 8, with probability at least $1 - \eta$

$$\max_g \left(\| S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1} \delta_A \|_\infty \right) \leq C \Big[\max_{j \in A, g} \left\{ (\Sigma_{gAA}^{-1})_{jj} (\delta_A^\top \Sigma_{gAA}^{-1} \delta_A \vee \gamma) \right\} s \ln(\eta^{-1}) / \min(n_1, n_2) \Big]^{1/2}.$$

Therefore, to have $A \subseteq \widehat{A}$, it is sufficient to have

$$\psi_{\min} > C\Big[\max_{j\in A,g}\Big\{(\Sigma_{gAA}^{-1})_{jj}(\delta_A^\top \Sigma_{gAA}^{-1}\delta_A \vee \gamma)\Big\} s\ln(\eta^{-1})/\min(n_1,n_2)\Big]^{1/2}$$
$$+\left[1 + C_1 \max_g \delta_A^\top \Sigma_{gAA}^{-1}\delta_A + C_2(\max_g \delta_A^\top \Sigma_{gAA}^{-1}\delta_A \vee \gamma)\Big\{s\ln(\eta^{-1})/\min(n_1,n_2)\Big\}\right]$$
$$\times \lambda \max_g \|\Sigma_{gAA}^{-1}\|_2 s^{1/2}\left[1 + C\Big\{s\ln(\eta^{-1})/\min(n_1,n_2)\Big\}^{1/2}\right].$$

Using the conditions on λ, and the fact that $\gamma \geq 1$, it follows that the second term above is the dominant term, and therefore it is sufficient to have, for some constant $C > 0$,

$$\psi_{\min} > C\lambda s^{1/2} \max_g \|\Sigma_{gAA}^{-1}\|_2 (\max_g \delta_A^\top \Sigma_{gAA}^{-1}\delta_A \vee \gamma).$$

□

Appendix E. Supporting theorems and lemmas

Theorem 5 (Oracle solution)**.** *Consider an oracle estimator* $[\tilde{v}_{1A}\ \tilde{v}_{2A}]$ *from* (B.3)*. Let* $\rho_1 = n_1/n$, $\rho_2 = n_2/n$. *Then*

$$\tilde{v}_{1A} = \rho_2 S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} - \lambda\left(S_{1AA} + \rho_2^2 d_A d_A^\top\right)^{-1} \tilde{u}_{1A},$$
$$\tilde{v}_{2A} = \rho_1 S_{2AA}^{-1} d_A (1 + \rho_1^2 d_A^\top S_{2AA}^{-1} d_A)^{-1} - \lambda\left(S_{2AA} + \rho_1^2 d_A d_A^\top\right)^{-1} \tilde{u}_{2A};$$

where $\tilde{u}_A$ *is the subgradient of* $\sum_{j=1}^s (v_{1Aj}^2 + v_{2Aj}^2)^{1/2}$.

Proof. We present the proof only for $\tilde{v}_{1A}$, the proof for $\tilde{v}_{2A}$ is analogous. From Section 2.3,

$$[\tilde{v}_{1A}\ \tilde{v}_{2A}] = \operatorname*{argmin}_{v_{1A}, v_{2A}\in\mathbb{R}^s} \Big\{\widehat{L}_{\psi_1}(v_{1A}) + \widehat{L}_{\psi_2}(v_{2A}) + \lambda \sum_{j=1}^s (v_{1Aj}^2 + v_{2Aj}^2)^{1/2}\Big\},$$
$$\widehat{L}_{\psi_1}(v_{1A}) + \widehat{L}_{\psi_2}(v_{2A}) = v_{1A}^\top S_{1AA} v_{1A}/2 + \left(n_2/n d_A^\top v_{1A} - 1\right)^2/2 + v_{2A}^\top S_{2AA} v_{2A}/2 + \left(n_2/n d_A^\top v_{2A} - 1\right)^2/2.$$

Using the optimality conditions, the oracle solution must satisfy

$$\tilde{v}_{1A} = \left(S_{1AA} + \rho_2^2 d_A d_A^\top\right)^{-1}(\rho_2 d_A - \lambda \tilde{u}_{1A}),$$

where $\tilde{u}_A$ is the subgradient of $\sum_{j=1}^s (v_{1Aj}^2 + v_{2Aj}^2)^{1/2}$. By Sherman–Morrison identity,

$$(S_{1AA} - \rho_2^2 d_A d_A^\top)^{-1} = S_{1AA}^{-1} - (1 + \rho_2^2 d_A^\top S_{1AA}^{-1} d_A)^{-1} \rho_2^2 S_{1AA}^{-1} d_A d_A^\top S_{1AA}^{-1}.$$

The statement follows by combining the above two displays. □

Lemma 1. *There exists a constant* $C > 0$ *such that with probability at least* $1 - \eta$

$$\left|n_g/n - \pi_g\right| \leq C\Big\{\ln(\eta^{-1})/n\Big\}^{1/2} \quad (g = 1, 2), \quad |n_1/n_2 - \pi_1/\pi_2| \leq C\Big\{\ln(\eta^{-1})/n\Big\}^{1/2}.$$

Proof. Given that $n_g \sim \mathrm{Bin}(n, \pi_g)$, by Hoeffding's inequality $\Pr(|\pi_g - n_g/n| \geq \varepsilon) \leq 2\exp(-2n\varepsilon^2)$. Let $\eta = 2\exp(-2n\varepsilon^2)$, then $2n\varepsilon^2 = \ln(2\eta^{-1})$, $\varepsilon = C\{\ln(\eta^{-1})/n\}^{1/2}$ and $n_g/n = \pi_g + O_p\{n^{-1/2}\}$. Let $f(x) = x/(1-x)$, which is non-decreasing for $x \in (0, 1)$. Since $n_1/n_2 = f(n_1/n)$, the second inequality in the lemma follows from the first. □

Lemma 2. *Let* $E_{Ug} = U_{gA^\complement} - \Sigma_{gA^\complement A}\Sigma_{gAA}^{-1} U_{gA}$, $E_{dg} = d_{A^\complement} - \Sigma_{gA^\complement A}\Sigma_{gAA}^{-1} d_A$, $g = 1, 2$. *Then* E_{Ug} *is independent from* U_{gA}, $E_{Ug} \sim \mathcal{N}(0, \Sigma_{gA^\complement A^\complement :A} \otimes I_{n_g - 1})$, $\quad e_j^\top E_{dg} \sim \mathcal{N}\left(0, n_g^{-1}\sigma_{jdg}^2\right)$; *where* $\sigma_{jdg}^2 = e_j^\top \Sigma_{d_g} e_j$, *and* $\Sigma_{gA^\complement A^\complement :A}$, Σ_{d_g} *are defined in* (B.2).

Proof. Since E_{dg}, E_{Ug} are formed by applying linear transformation to normal d, U_1, U_2, it follows that E_{dg}, E_{Ug} are also normally distributed. It remains to verify the form of the means and covariance matrices. We consider $g = 1$, the proof for $g = 2$ is similar.

Consider E_{U1}. By definition, the columns of U_1 satisfy $u_{1i} \sim \mathcal{N}(0, \Sigma_1)$. Since

$$E_{U1} = (-\Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\ I_{p-s})\begin{pmatrix} U_{1A} \\ U_{1A^{\complement}} \end{pmatrix},$$

it follows that $\mathrm{E}(E_{U1}) = 0$, and

$$\begin{aligned}\mathrm{var}(E_{U1}) &= (-\Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\ I_{p-s})\begin{pmatrix} \Sigma_{1AA} & \Sigma_{1AA^{\complement}} \\ \Sigma_{1A^{\complement}A} & \Sigma_{1A^{\complement}A^{\complement}} \end{pmatrix}(-\Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\ I_{p-s})^{\top} \otimes I_{n_1-1} \\ &= (\Sigma_{1A^{\complement}A^{\complement}} - \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}}) \otimes I_{n_1-1}.\end{aligned}$$

Consider E_{d1}. Since $\Sigma_1^{-1}\delta = \psi_1 = (\psi_{1A}^{\top}, 0)^{\top}$, by rewriting $\Sigma_1\Sigma_1^{-1}\delta = \delta$, and using block matrices of Σ_1 and Σ_1^{-1}, it follows that $\Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\delta_A = \delta_{A^{\complement}}$. Then $\mathrm{E(E_{d1})} = \delta_{\mathrm{A}^{\complement}} - \Sigma_{1\mathrm{A}^{\complement}\mathrm{A}}\Sigma_{1\mathrm{AA}}^{-1}\delta_{\mathrm{A}} = 0$. Furthermore,

$$\begin{aligned}\mathrm{var}(E_{d1}) &= \mathrm{var}(d_{A^{\complement}} - \Sigma_{1A^{\complement}A}\Sigma_{1,AA}^{-1}d_A) \\ &= \mathrm{var}(d_{A^{\complement}}) + \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\,\mathrm{var}(d_A)\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}} - \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\,\mathrm{cov}(d_A, d_{A^{\complement}}) - \mathrm{cov}(d_{A^{\complement}}, d_A)\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}} \\ &= n_1^{-1}\Sigma_{1A^{\complement}A^{\complement}} + n_2^{-1}\Sigma_{2A^{\complement}A^{\complement}} + \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\left(n_1^{-1}\Sigma_{1AA} + n_2^{-1}\Sigma_{2AA}\right)\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}} \\ &\quad - \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\left(n_1^{-1}\Sigma_{1AA^{\complement}} + n_2^{-1}\Sigma_{2AA^{\complement}}\right) - \left(n_1^{-1}\Sigma_{1A^{\complement}A} + n_2^{-1}\Sigma_{2A^{\complement}A}\right)\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}} \\ &= n_1^{-1}\Sigma_{1A^{\complement}A^{\complement}:A} + n_2^{-1}\Big(\Sigma_{2A^{\complement}A^{\complement}} + \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{2AA}\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}} - \Sigma_{1A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{2AA^{\complement}} - \Sigma_{2A^{\complement}A}\Sigma_{1AA}^{-1}\Sigma_{1AA^{\complement}}\Big).\end{aligned}$$

□

Lemma 3. *Let S_{gAA} be a submatrix of the sample covariance matrix for group $g \in \{1, 2\}$ corresponding to variables in A, with $s = \mathrm{card}(A)$. Let Σ_{gAA} be the corresponding submatrix of population covariance matrix. Under Assumption 1, there exist constants $C_1, C_2 > 0$ such that with probability at least $1 - \eta$*

$$\|\Sigma_{gAA}^{1/2}S_{gAA}^{-1}\Sigma_{gAA}^{1/2} - I\|_2 \leq C_1\Big\{s\ln(\eta^{-1})/n_g\Big\}^{1/2},\ \|S_{gAA}^{-1}\|_2 \leq \|\Sigma_{gAA}^{-1}\|_2\Big[1 + C_2\Big\{s\ln(\eta^{-1})/n_g\Big\}^{1/2}\Big].$$

Proof. Using normality, the sample covariance matrices satisfy $S_{gAA} = (n_g - 1)^{-1}W_gW_g^{\top}$ with $W_g \in \mathbb{R}^{s\times(n_g-1)}$ having independent columns $w_{gi} \sim \mathcal{N}(0, \Sigma_{gAA})$. Then the desired bounds follow from Lemma 9 in Wainwright [50]. □

Lemma 4. *Let a random vector $X \in \mathbb{R}^s$ be such that $X \sim \mathcal{N}(0, n^{-1}A)$. Then there exists a constant $C > 0$ such that with probability at least $1 - \eta$*

$$\|X\|_2 \leq C\Big\{\|A\|_2 n^{-1}s\ln(\eta^{-1})\Big\}^{1/2}.$$

Proof. Since $A^{-1/2}X \sim \mathcal{N}(0, n^{-1}I_s)$, by Hsu et al. [22, Proposition 1.1], with probability at least $1 - \eta$

$$\|A^{-1/2}X\|_2^2 \leq s/n + 2\Big\{s\ln(\eta^{-1})\Big\}^{1/2}/n + 2\ln(\eta^{-1})/n.$$

For small η it follows that there exist $C > 0$ such that $\|A^{-1/2}X\|_2^2 \leq Cn^{-1}s\ln(\eta^{-1})$ with probability at least $1 - \eta$. The statement of the lemma follows since

$$\|X\|_2^2 = X^{\top}X = X^{\top}A^{-1/2}AA^{-1/2}X \leq \|A\|_2\|A^{-1/2}X\|_2^2.$$

□

Lemma 5. *There exist constant $C > 0$ such that with probability at least $1 - \eta$, and γ in (B.1)*

$$\max_g \|\Sigma_{gAA}^{-1/2}(d_A - \delta_A)\|_2 \leq C\Big\{\gamma s\ln(\eta^{-1})/\min(n_1, n_2)\Big\}^{1/2}.$$

Proof. Since $d_A - \delta_A \sim \mathcal{N}(0, n_1^{-1}\Sigma_{1AA} + n_2^{-1}\Sigma_{2AA})$, it follows that

$$\Sigma_{1AA}^{-1/2}(d_A - \delta_A) \sim \mathcal{N}\left[0, n_1^{-1}\left(I + n_2^{-1}n_1\Sigma_{1AA}^{-1/2}\Sigma_{2AA}\Sigma_{1AA}^{-1/2}\right)\right].$$

Applying Lemmas 1 and 4 concludes the proof. The case $g = 2$ is analogous. □

Lemma 6. *There exist constants C_1, C_2 such that with probability at least $1-\eta$ for $g = 1, 2$*

$$d_A^\top S_{gAA}^{-1} d_A \le C_1 d_A^\top \Sigma_{gAA}^{-1} d_A \left[1 + C_2\Big\{\ln(\eta^{-1})/(n_g - s)\Big\}^{1/2}\right].$$

Proof. We prove for $g = 1$, case $g = 2$ is analogous. Since $(n_1 - 1)S_{1AA} \sim W_s(n_1 - 1, \Sigma_{1AA})$, and d_A is independent of S_{1AA}, by Theorem 3.2.12 in [38]

$$(n_1 - 1)\frac{d_A^\top \Sigma_{1AA}^{-1} d_A}{d_A^\top S_{1AA}^{-1} d_A} \sim \chi^2_{n_1 - s}.$$

Using Lemma 1 in [29],

$$\Pr\left[(n_1 - 1)\frac{d_A^\top \Sigma_{1AA}^{-1} d_A}{d_A^\top S_{1AA}^{-1} d_A} \ge (n_1 - s) - 2\Big\{(n_1 - s)\ln(\eta^{-1})\Big\}^{1/2}\right] \ge 1 - \eta.$$

Therefore, with probability at least $1 - \eta$

$$d_A^\top S_{1AA}^{-1} d_A \le (n_1 - 1)(n_1 - s)^{-1} d_A^\top \Sigma_{1AA}^{-1} d_A \left[1 - 2\Big\{\ln(\eta^{-1})/(n_1 - s)\Big\}^{1/2}\right]^{-1}.$$

Hence, there exist constants $C_1, C_2 > 0$ such that with probability at least $1 - \eta$

$$d_A^\top S_{1AA}^{-1} d_A \le C_1 d_A^\top \Sigma_{1AA}^{-1} d_A \left[1 + C_2\Big\{\ln(\eta^{-1})/(n_1 - s)\Big\}^{1/2}\right].$$

□

Lemma 7. *There exists a constant $C > 0$ such that with probability at least $1 - \eta$, and γ in (B.1)*

$$d_A^\top \Sigma_{gAA}^{-1} d_A \le C\left\{\delta_A^\top \Sigma_{gAA}^{-1}\delta_A + \gamma n_g^{-1} s \ln(\eta^{-1})\right\} \quad (g = 1, 2).$$

Proof. We prove the result for $g = 1$, the case $g = 2$ is similar. Consider

$$\begin{aligned} d_A^\top \Sigma_{1AA}^{-1} d_A &= \delta_A^\top \Sigma_{1AA}^{-1}\delta_A + 2(d_A - \delta_A)^\top \Sigma_{1AA}^{-1}\delta_A + (d_A - \delta_A)^\top \Sigma_{1AA}^{-1}(d_A - \delta_A) \\ &\le 2\delta_A^\top \Sigma_{1AA}^{-1}\delta_A + 2(d_A - \delta_A)^\top \Sigma_{1AA}^{-1}(d_A - \delta_A). \end{aligned}$$

By Lemma 5, there exists a constant $C \ge 0$ such that with probability at least $1 - \eta$,

$$(d_A - \delta_A)^\top \Sigma_{1AA}^{-1}(d_A - \delta_A) \le C\gamma n_1^{-1} s \ln(\eta^{-1}).$$

The result follows by combining the above displays. □

Corollary 1. *There exist constants $C_1, C_2, C_3 > 0$ such that with probability at least $1 - \eta$ for $g = 1, 2$ and γ in (B.1)*

$$d_A^\top S_{gAA}^{-1} d_A \le C_1 \delta_A^\top \Sigma_{gAA}^{-1}\delta_A \left[1 + C_2\left\{\ln(\eta^{-1})/(n_g - s)\right\}^{1/2}\right] + C_3\gamma n_g^{-1} s \ln(\eta^{-1}).$$

Proof. The result follows by combining results of Lemmas 6–7. □

Lemma 8. *There exists a constant $C > 0$ such that with probability at least $1 - \eta$ for $g \in \{1, 2\}$, and γ in (B.1)*

$$\|S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1}\delta_A\|_\infty \le C\Big\{\max_{j\in A}(\Sigma_{gAA}^{-1})_{jj}(\delta_A^\top \Sigma_{gAA}^{-1}\delta_A \vee \gamma) n_g^{-1} s \ln(\eta^{-1})\Big\}^{1/2}.$$

Proof. We prove the result for $g = 1$, the case $g = 2$ is similar. Consider

$$\begin{aligned} &|e_j^\top S_{1AA}^{-1} d_A - e_j^\top \Sigma_{1AA}^{-1}\delta_A| \\ &\quad= |e_j^\top (S_{1AA}^{-1} - \Sigma_{1AA}^{-1})(d_A - \delta_A) + e_j^\top (S_{1AA}^{-1} - \Sigma_{1AA}^{-1})\delta_A + e_j^\top \Sigma_{1AA}^{-1}(d_A - \delta_A)| \\ &\quad\le (e_j^\top \Sigma_{1AA}^{-1} e_j)^{1/2}\|(\Sigma_{1AA}^{1/2} S_{1AA}^{-1}\Sigma_{1AA}^{1/2} - I)\Sigma_{1AA}^{-1/2}(d_A - \delta_A)\|_2 + (e_j^\top \Sigma_{1AA}^{-1} e_j)^{1/2}\|(\Sigma_{1AA}^{1/2} S_{1AA}^{-1}\Sigma_{1AA}^{1/2} - I)\Sigma_{1AA}^{-1/2}\delta_A\|_2 \\ &\qquad+ (e_j^\top \Sigma_{1AA}^{-1} e_j)^{1/2}\|\Sigma_{1AA}^{-1/2}(d_A - \delta_A)\|_2. \end{aligned}$$

Let $m_1 = \|\Sigma_{1AA}^{1/2} S_{1AA}^{-1} \Sigma_{1AA}^{1/2} - I\|_2$ and $m_2 = \|\Sigma_{1AA}^{-1/2}(d_A - \delta_A)\|_2$. Using the above display

$$\|S_{1AA}^{-1} d_A - \Sigma_{1AA}^{-1}\delta_A\|_\infty \leq \max_{j\in A}(\Sigma_{1AA}^{-1})_{jj}^{1/2}\left\{m_1 m_2 + m_1(\delta_A^\top \Sigma_{1AA}^{-1}\delta_A)^{1/2} + m_2\right\}. \quad \text{(C.1)}$$

Using Lemma 3, there exists a constant $C_1 > 0$ such that $m_1 \leq C_1\{s\ln(\eta^{-1})/n_1\}^{1/2}$ with probability at least $1-\eta$. Using Lemma 5, there exists a constant $C_2 > 0$ such that $m_2 \leq C_2\{\gamma s\ln(\eta^{-1})/n_1\}^{1/2}$ with probability at least $1-\eta$. Combining these bounds with (C.1), there exist constant $C > 0$ such that with probability at least $1-\eta$

$$\|S_{1AA}^{-1} d_A - \Sigma_{1AA}^{-1}\delta_A\|_\infty \leq C\Big\{\max_{j\in A}(\Sigma_{1AA}^{-1})_{jj}(\delta_A^\top \Sigma_{1AA}^{-1}\delta_A \vee \gamma)n_1^{-1} s\ln(\eta^{-1})\Big\}^{1/2}.$$

□

[1] F. R. Bach, Consistency of the Group Lasso and Multiple Kernel Learning, Journal of Machine Learning Research 9 (2008) 1179–1225.

[2] R. F. Barber, M. Drton, Exact block-wise optimization in group lasso and sparse group lasso for linear regression, arXiv.org (2010).

[3] S. P. Boyd, L. Vandenberghe, Convex Optimization, Cambridge Univ Press, Cambridge, 2004.

[4] P. Breheny, J. Huang, Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, Statistics and computing 25 (2015) 173–187.

[5] T. Cai, W. Liu, A direct estimation approach to sparse linear discriminant analysis, Journal of the American Statistical Association 106 (2011) 1566–1577.

[6] H.-W. Chen, H.-C. Huang, Y.-S. Lin, K.-J. Chang, W.-H. Kuo, H.-L. Hwa, F.-J. Hsieh, H.-F. Juan, Comparison and identification of estrogen-receptor related gene expression profiles in breast cancer of different ethnic origins., Breast cancer : basic and clinical research 1 (2008) 35–49.

[7] K. Chin, S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. M. Ljung, L. Esserman, D. G. Albertson, F. M. Waldman, J. W. Gray, Genomic and transcriptional aberrations linked to breast cancer pathophysiologies., Cancer Cell 10 (2006) 529–541.

[8] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, A. Mazumder, Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative., The Journal of molecular diagnostics : JMD 8 (2006) 31–39.

[9] L. Clemmensen, D. M. Witten, T. J. Hastie, B. Ersbøll, Sparse Discriminant Analysis, Technometrics 53 (2011) 406–413.

[10] P. Danaher, JGL: Performs the Joint Graphical Lasso for sparse inverse covariance estimation on multiple classes, 2013. R package version 2.3.

[11] P. Danaher, P. Wang, D. M. Witten, The joint graphical lasso for inverse covariance estimation across multiple classes, Journal of the Royal Statistical Society, Ser. B 76 (2014) 373–397.

[12] S. Dudoit, J. Fridlyand, T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97 (2002) 77–87.

[13] J. H. Friedman, Regularized discriminant analysis, Journal of the American Statistical Association 84 (1989) 165–175.

[14] I. Gaynanova, MGSDA: Multi-Group Sparse Discriminant Analysis, 2016. R package version 1.4.

[15] I. Gaynanova, J. G. Booth, M. T. Wells, Simultaneous sparse estimation of canonical vectors in the $p >> N$ setting, Journal of the American Statistical Association 111 (2016) 696–706.

[16] I. Gaynanova, J. G. Booth, M. T. Wells, Penalized Versus Constrained Generalized Eigenvalue Problems, Journal of Computational and Graphical Statistics 26 (2017) 379–387.

[17] I. Gaynanova, M. Kolar, Optimal variable selection in multi-group sparse discriminant analysis, Electronic Journal of Statistics 9 (2015) 2007–2034.

[18] E. Gravier, G. Pierron, A. Vincent-Salomon, N. Gruel, V. Raynal, A. Savignoni, Y. De Rycke, J.-Y. Pierga, C. Lucchesi, F. Reyal, A. Fourquet, S. Roman-Roman, F. Radvanyi, X. Sastre-Garau, B. Asselain, O. Delattre, A prognostic DNA signature for T1T2 node-negative breast cancer patients., Genes, chromosomes & cancer 49 (2010) 1125–1134.

[19] J. Guo, E. Levina, G. Michailidis, J. Zhu, Joint estimation of multiple graphical models, Biometrika 98 (2011) 1–15.

[20] F. Holst, Estrogen receptor alpha gene amplification in breast cancer: 25 years of debate., World journal of clinical oncology 7 (2016) 160–173.

[21] F. Holst, P. R. Stahl, C. Ruiz, O. Hellwinkel, Z. Jehan, M. Wendland, A. Lebeau, L. Terracciano, K. Al-Kuraya, F. Jänicke, G. Sauter, R. Simon, Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer., Nature genetics 39 (2007) 655–660.

[22] D. Hsu, S. M. Kakade, T. Zhang, A tail inequality for quadratic forms of subgaussian random vectors, Electronic Communications in Probability 17 (2012) no. 52–6.

[23] J. Huang, P. Breheny, S. Ma, A Selective Review of Group Selection in High-Dimensional Models, Statistical Science 27 (2012) 481–499.

[24] T. Iwamoto, D. Booser, V. Valero, J. L. Murray, K. Koenig, F. J. Esteva, N. T. Ueno, J. Zhang, W. Shi, Y. Qi, J. Matsuoka, E. J. Yang, G. N. Hortobagyi, C. Hatzis, W. F. Symmans, L. Pusztai, Estrogen receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry., JCO 30 (2012) 729–734.

[25] T. Kadota, L. Shepp, On the best finite set of linear observables for discriminating two Gaussian signals, IEEE Transactions on Information Theory 13 (1967) 278–284.

[26] M. Kolar, H. Liu, Optimal feature selection in high-dimensional discriminant analysis, IEEE Transactions on Information Theory 61 (2015) 1063–1083.

[27] S. Kullback, An application of information theory to multivariate analysis, Annals of Mathematical Statistics 23 (1952) 88–102.

[28] A.-V. Laenkholm, A. Knoop, B. Ejlertsen, T. Rudbeck, M.-B. Jensen, S. Müller, A. E. Lykkesfeldt, B. B. Rasmussen, K. V. Nielsen, ESR1 gene status correlates with estrogen receptor protein levels measured by ligand binding assay and immunohistochemistry., Molecular oncology 6 (2012) 428–436.

[29] B. Laurent, P. Massart, Adaptive Estimation of a Quadratic Functional by Model Selection, Annals of Statistics 28 (2000) 1302–1338.

[30] Y. Le, T. J. Hastie, Sparse Quadratic Discriminant Analysis and Community Bayes, arXiv.org (2014).

[31] Q. Li, J. Shao, Sparse quadratic discriminant analysis for high dimensional data, Statistica Sinica 25 (2015) 457–473.

[32] Y. Li, A. Ngom, Nonnegative least-squares methods for the classification of high-dimensional biological data, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 10 (2013) 447–456.
[33] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information theory 37 (1991) 145–151.
[34] Q. Mai, H. Zou, A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods, Technometrics 55 (2013) 243–246.
[35] Q. Mai, H. Zou, M. Yuan, A direct approach to sparse discriminant analysis in ultra-high dimensions, Biometrika 99 (2012) 29–42.
[36] K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, New York, 1979.
[37] O. Mersmann, microbenchmark: Accurate Timing Functions, 2015. R package version 2.1.
[38] R. J. Muirhead, Aspects of multivariate statistical theory, John Wiley and Sons, Inc., New York, 1982.
[39] Y. S. Niu, N. Hao, B. Dong, A new reduced-rank linear discriminant analysis method and its applications, Statistica Sinica 28 (2018) 189–202.
[40] G. Obozinski, M. J. Wainwright, M. I. Jordan, Support union recovery in high-dimensional multivariate regression, Annals of Statistics 39 (2011) 1–47.
[41] B. S. Price, RidgeFusion: R Package for Ridge Fusion in Statistical Learning, 2014. R package version 1.0-3.
[42] B. S. Price, C. J. Geyer, A. J. Rothman, Ridge Fusion in Statistical Learning, Journal of Computational and Graphical Statistics 24 (2014) 439–454.
[43] J. A. Ramey, datamicroarray: Collection of Data Sets for Classification, 2016. Https://github.com/ramhiser/datamicroarray, http://ramhiser.com.
[44] J. A. Ramey, C. K. Stein, P. D. Young, D. M. Young, High-Dimensional Regularized Discriminant Analysis, arXiv.org (2016).
[45] A. L. Rukhin, Generalized Bayes estimators of a normal discriminant function, Journal of Multivariate Analysis 41 (1992) 154–162.
[46] N. Simon, R. J. Tibshirani, Discriminant Analysis with Adaptively Pooled Covariance, arXiv.org (2011).
[47] N. Simon, R. J. Tibshirani, Standardization and the group Lasso penalty, Statistica Sinica 22 (2012) 983–1001.
[48] R. J. Tibshirani, T. J. Hastie, B. Narasimhan, G. Chu, Class prediction by nearest shrunken centroids, with applications to DNA microarrays, Statistical Science 18 (2003) 104–117.
[49] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, Journal of optimization theory and applications 109 (2001) 475–494.
[50] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), IEEE Transactions on Information Theory 55 (2009) 2183–2202.
[51] T. Wang, I. Gaynanova, DAP: Discriminant Analysis via Projections, 2018. R package version 1.0.
[52] H. Wickham, ggplot2: elegant graphics for data analysis, New York: Springer, 2016.
[53] D. M. Witten, R. J. Tibshirani, Penalized classification using Fisher's linear discriminant., Journal of the Royal Statistical Society, Ser. B 73 (2011) 753–772.
[54] M. C. Wu, L. Zhang, Z. Wang, D. C. Christiani, X. Lin, Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection., Bioinformatics 25 (2009) 1145–1151.
[55] Y. Wu, Y. Qin, M. Zhu, Quadratic Discriminant Analysis for High-Dimensional Data, Statistica Sinica to appear (2018).
[56] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, Ser. B 68 (2006) 49–67.