ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 4, *pp.* 827–851

Sparsifying the Fisher linear discriminant by rotation

Ning Hao and Bin Dong

University of Arizona, Tucson, USA

and Jianqing Fan

Princeton University, USA

[Received May 2013. Final revision August 2014]

Summary. Many high dimensional classification techniques have been proposed in the literature based on sparse linear discriminant analysis. To use them efficiently, sparsity of linear classifiers is a prerequisite. However, this might not be readily available in many applications, and rotations of data are required to create the sparsity needed. We propose a family of rotations to create the sparsity required. The basic idea is to use the principal components of the sample covariance matrix of the pooled samples and its variants to rotate the data first and then to apply an existing high dimensional classifier. This rotate-and-solve procedure can be combined with any existing classifiers and is robust against the level of sparsity of the true model. We show that these rotations do create the sparsity that is needed for high dimensional classifications and we provide theoretical understanding why such a rotation works empirically. The effectiveness of the method proposed is demonstrated by several simulated and real data examples, and the improvements of our method over some popular high dimensional classification rules are clearly shown.

Keywords: Classification; Equivariance; High dimensional data; Linear discriminant analysis; Principal components; Rotate-and-solve procedure

1. Introduction

Linear discriminant analysis (LDA) is a useful classical tool for classification. Consider two p-dimensional normal distributions with the same covariance matrix, $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for class 1 and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for class 2. Given a random vector $\mathbf{X}$ which is from one of these distributions with equal prior probabilities, a *linear discriminant rule*

$$\psi_{\boldsymbol{\omega},\boldsymbol{\nu}}(\mathbf{X}) = I\{(\mathbf{X}-\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{\omega} \geqslant 0\}, \qquad \boldsymbol{\omega}, \boldsymbol{\nu} \in \mathbb{R}^p, \tag{1.1}$$

assigns $\mathbf{X}$ to class 1 when $\psi_{\boldsymbol{\omega},\boldsymbol{\nu}}(\mathbf{X}) = 1$ and to class 2 otherwise. Geometrically, the equation $(\mathbf{x}-\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{\omega} = 0$ defines an affine space passing through a point $\boldsymbol{\nu}$ with a normal vector $\boldsymbol{\omega}$, which is the discriminant boundary of the classification rule.

When $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known, the optimal classifier, namely the Fisher linear discriminant rule, is

$$\psi_{\mathrm{F}}(\mathbf{X}) = I\{(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \geqslant 0\}, \tag{1.2}$$

where $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ and $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. In practice, these parameters are unknown and replaced by their estimates. Let $\{\mathbf{X}_i^{(1)} : 1 \leqslant i \leqslant n_1\}$ and $\{\mathbf{X}_i^{(2)} : 1 \leqslant i \leqslant n_2\}$ be independent and identically

Address for correspondence: Ning Hao, Department of Mathematics, University of Arizona, Tucson, AZ 85721-0089, USA.
E-mail: nhao@math.arizona.edu

 1369–7412/15/77827

distributed observations from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ respectively. In the classical setting with $n_1, n_2 \gg p$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}^{-1}$ are usually estimated by sample means $\hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{X}}^{(1)}$ and $\hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{X}}^{(2)}$ and the inverse pooled sample covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}$. Standard LDA uses an empirical version of rule (1.2):

$$\psi_{\hat{\mathrm{F}}}(\mathbf{X}) = I\{(\mathbf{X} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\delta}} \geqslant 0\}, \tag{1.3}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)$ and $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$.

Although standard LDA has been widely used in applications, it does not work well for high dimensional data when p is comparable with or larger than the sample size. The reason is that, with a limited number of observations, it is impossible to estimate too many parameters simultaneously and accurately. In particular, $\hat{\boldsymbol{\Sigma}}$ is singular and not invertible when $n_1 + n_2 < p - 1$. One may use the pseudoinverse $\hat{\boldsymbol{\Sigma}}^{-}$, but Bickel and Levina (2004) showed that LDA performs as poorly as random guessing when $p/(n_1 + n_2) \to \infty$. Since the work of Bickel and Levina (2004), a series of LDA-based methods have been proposed for the high dimensional classification problem. The main idea is to find methods which work well when the original classification problem is (nearly) sparse so that $\boldsymbol{\mu}$ or $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ in the optimal rule (1.2) can be well estimated. Ignoring the covariances between the features, Bickel and Levina (2004) proposed an independence rule (IR) which outperforms standard LDA in the high dimensional setting. Fan and Fan (2008) proposed the features annealed IR that selects a subset of features before applying the IR. In spite of the clear interpretations of the sparsity of the covariance matrix $\boldsymbol{\Sigma}$ and difference of centroids $\boldsymbol{\delta}$, in practice, it might be more efficient to find the sparse discriminant affine space directly (see Trendafilov and Jolliffe (2007), Wu *et al*. (2009), Cai and Liu (2011), Fan *et al*. (2012) and Mai *et al*. (2012) among others). Here, a sparse discriminant affine space is an affine space with a sparse normal vector. In particular, Fan *et al*. (2012) and Cai and Liu (2011) clearly illustrated the advantages of their direct approaches over the IR and features annealed IR, which oversimplify the problem in many scenarios.

For all the aforementioned LDA-based high dimensional classification rules, various explicit sparsity conditions on one or some of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are crucial to the classification accuracy. For example, the IR (Bickel and Levina, 2004) works well only when $\boldsymbol{\Sigma}$ is nearly diagonal; the features annealed IR (Fan and Fan, 2008) needs ideally diagonal $\boldsymbol{\Sigma}$ and sparse $\boldsymbol{\delta}$; the regularized optimal affine discriminant (ROAD) (Fan *et al*., 2012) and the linear programming discriminant (LPD) (Cai and Liu, 2011) need $\boldsymbol{\beta}$ to be sparse to achieve optimal classification. We shall refer to all these methods as sparse LDA methods. They are efficient when the corresponding sparsity conditions are granted. However, they may not work well when the sparsity conditions are violated. Although these sparse assumptions make sense in some applications, they can be too restrictive in many scenarios (see Hall *et al*. (2014) and reference therein). It is a natural and challenging question how and to what extent we can sparsify a possibly non-sparse problem.

To solve a non-sparse model, a natural idea is to rotate the data to a nearly sparse setting before applying sparse LDA methods. For example, the classification problem can be easily solved by the ROAD and LPD if the normal vector of the optimal discriminant affine space, $\boldsymbol{\beta}$, is sparse after a rotation. To do this, we need an oracle that can rotate the data to such a sparse setting. For the ideal case when $\boldsymbol{\beta}$ is known, there are infinitely many orthogonal matrices which can rotate $\boldsymbol{\beta}$ to a sparse vector $(\|\boldsymbol{\beta}\|_2, 0, \ldots, 0)^{\mathrm{T}}$. However, it is not realistic to approximate such rotations before estimating $\boldsymbol{\beta}$ itself. An alternative way might be to make $\boldsymbol{\Sigma}$ diagonal after a rotation, which is related to principal component analysis. However, such a rotation does not combine the information of the centroids and cannot accurately estimate directions with small variances, which may actually be crucial for classification.

In this paper, we propose a class of rotations which balance both mean and variance infor-

mation. Intuitively, both $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ should play essential roles in a rotation to make $\boldsymbol{\beta}$ sparse. In particular, if $\boldsymbol{\Sigma}$ is spiked (Johnstone, 2001), its principal components and $\boldsymbol{\delta}$ span a linear space, which contains key information on the rotation. Following this intuition, we define $\boldsymbol{\Sigma}_\rho^{\text{tot}} = \boldsymbol{\Sigma} + \rho\boldsymbol{\delta}\boldsymbol{\delta}^{\text{T}}$ for $\rho > 0$, whose principal components are determined by those of $\boldsymbol{\Sigma}$ as well as $\boldsymbol{\delta}$. Consider an orthogonal matrix $\mathbf{U}_\rho$, formed by the eigenvectors of $\boldsymbol{\Sigma}_\rho^{\text{tot}}$, which diagonalizes $\boldsymbol{\Sigma}_\rho^{\text{tot}}$. We shall show that $\mathbf{U}_\rho^{\text{T}}\boldsymbol{\beta}$ is sparse when the covariance matrix $\boldsymbol{\Sigma}$ is spiked. In other words, the eigenvectors of $\boldsymbol{\Sigma}_\rho^{\text{tot}}$ are good directions to rotate. Similarly, we can define the empirical version $\hat{\mathbf{U}}_\rho$ which diagonalizes $\hat{\boldsymbol{\Sigma}}_\rho^{\text{tot}} = \hat{\boldsymbol{\Sigma}} + \rho\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\text{T}}$. The rotation $\hat{\mathbf{U}}_\rho$ is a reasonably good approximation to $\mathbf{U}_\rho$ when $p \ll n$ (Johnstone and Lu, 2009) or $p > n$ with some additional conditions (Zou *et al.*, 2006; Fan *et al.*, 2013). In other words, under some conditions on $\boldsymbol{\Sigma}$, $\hat{\mathbf{U}}_\rho^{\text{T}}\boldsymbol{\beta}$ is nearly sparse, regardless of the level of sparsity of the original $\boldsymbol{\beta}$. Therefore, we propose to rotate the data by $\hat{\mathbf{U}}_\rho^{\text{T}}$ first before applying the ROAD or LPD, when the level of sparsity of $\boldsymbol{\beta}$ is unknown. Although our original motivation is to make $\boldsymbol{\beta}$ sparse by rotation, we find that our procedure is equivariant with respect to orthogonal transformation group $O(p)$ consisting of all rotations. This feature makes our method robust against the level of sparsity of $\boldsymbol{\beta}$. The advantage of our method is illustrated by numerous simulated and real data examples. An implementation of the proposed rotations by using MATLAB can be found on the Web site `http://math.arizona.edu/~dongbin/Data/HDF_Rotation.m`.

The rest of our paper is organized as follows. Section 2 introduces a family of ideal rotations and analyses their theoretical properties. In Section 3, we study a rotate-and-solve (RS) procedure for classification. Numerical studies on both simulated and real data are demonstrated in Section 4. All proofs are given in Appendix A. Various norms of vectors and matrices appear frequently in the paper. For a vector $\mathbf{a}$, $\|\mathbf{a}\|_p$ denote the standard l_p-norm. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ is the spectral norm.

2. A family of oracle rotations and their properties

As mentioned in Section 1, the performance of sparse LDA methods depends highly on the sparsity of $\boldsymbol{\beta}$, which is unknown and difficult to verify in practice. High dimensional classifiers will work more efficiently if an oracle rotates the data to a sparse setting before applying sparse LDA methods. If $\boldsymbol{\beta}$ is known, we can easily rotate $\boldsymbol{\beta}$ to a sparse vector $(\|\boldsymbol{\beta}\|_2, 0, \ldots, 0)^{\text{T}}$. Of course, it is meaningless to mimic such an oracle, which motivates us to find other ideal rotations that can be estimated more easily.

Recall that the distributions of two classes are $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for class 1 and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for class 2. Let $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and

$$\boldsymbol{\Sigma}_\rho^{\text{tot}} = \boldsymbol{\Sigma} + \rho\boldsymbol{\delta}\boldsymbol{\delta}^{\text{T}}, \qquad \text{for a given } \rho > 0.$$

Consider an orthogonal matrix $\mathbf{U}_\rho$, which is formed by the eigenvectors of $\boldsymbol{\Sigma}_\rho^{\text{tot}}$, which diagonalizes $\boldsymbol{\Sigma}_\rho^{\text{tot}}$. Then, without loss of generality by rearranging columns in $\mathbf{U}_\rho$, we assume that $\mathbf{U}_\rho^{\text{T}}\boldsymbol{\Sigma}_\rho^{\text{tot}}\mathbf{U}_\rho = \mathbf{D}_\rho$ where $\mathbf{D}_\rho = \text{diag}(\eta_1, \ldots, \eta_p)$ is the diagonal matrix, consisting of eigenvalues in descending order.

Let $\{\lambda_j\}_{j=1}^p$ be eigenvalues of $\boldsymbol{\Sigma}$, arranged from the largest to the smallest, and $\{\boldsymbol{\xi}_j\}_{j=1}^p$ be their corresponding eigenvectors. For repeated eigenvalues, say $\lambda_r = \lambda_{r+1} = \ldots = \lambda_s$, $\{\boldsymbol{\xi}_j\}_{j=r}^s$ can be chosen as any orthonormal basis of the corresponding eigenspace. Johnstone (2001) considered a spiked covariance model, where a few large eigenvalues clearly stand out from the rest.

Condition 1 (spiked covariance structure). Assume that $\lambda_1 \geqslant \ldots \geqslant \lambda_k > \lambda_{k+1} = \ldots = \lambda_p$ for some integer $k < p$.

Theorem 1. Under condition 1, we have $\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_0 \leqslant k+1$.

Theorem 1 shows the sparsity property of $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ when $\boldsymbol{\Sigma}$ is spiked and $k+1<p$. In particular, it implies that $\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_2 \leqslant \sqrt{(k+1)}$ by Cauchy–Schwarz inequality. The boundedness of the l_0- or l_1-norm is crucial for sparse LDA methods such as the ROAD and LPD to be efficient. For a vector that is randomly picked on the unit sphere in $\mathbb{R}^p$, the expectation of its l_1-norm is of order $\sqrt{p}$. Therefore, both l_0- and l_1-norms of $\boldsymbol{\beta}$ have been greatly reduced after rotation when $k \ll p$.

The condition of theorem 1 can still be relaxed somehow while keeping $\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_2$ bounded. This is shown in theorem 2 below.

Condition 2 (quasi-spiked covariance structure). Assume that $\lambda_k \geqslant \lambda_{k+1}+d$ and $\lambda_{k+1}-\lambda_p \leqslant \varepsilon$ for some integer $k<p$, where $d, \varepsilon > 0$.

Let $\mathbf{W}_1$ and $\mathbf{W}_2$ be two linear spaces spanned by $\{\boldsymbol{\xi}_j\}_{1\leqslant j\leqslant k}$ and $\{\boldsymbol{\xi}_j\}_{k+1\leqslant j\leqslant p}$ respectively. Then, we have $\mathbb{R}^p = \mathbf{W}_1 \oplus \mathbf{W}_2$ and the mean difference vector $\boldsymbol{\delta}$ can be decomposed as $\boldsymbol{\delta} = \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2$ with $\boldsymbol{\delta}_1 \in \mathbf{W}_1$ and $\boldsymbol{\delta}_2 \in \mathbf{W}_2$.

Theorem 2. If $\boldsymbol{\delta} \in \mathbf{W}_1$ and $\lambda_k > \lambda_{k+1}$, then $\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_0 \leqslant k$ and

$$\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_2 \leqslant \sqrt{k}.$$

If $\boldsymbol{\delta} \notin \mathbf{W}_1$ and condition 2 holds, then

$$\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_2 \leqslant \sqrt{(k+1)} + \sqrt{(p-k-1)}\frac{\lambda_p+\varepsilon}{\lambda_p}\left\{\frac{\varepsilon}{\lambda_p} + \sqrt{\left(\frac{\varepsilon}{\tilde{d}-2\varepsilon}\right)}\right\},$$

provided that $\varepsilon < \tilde{d}/2$, where $\tilde{d} = d\rho\|\boldsymbol{\delta}_2\|_2^2/(d+\rho\|\boldsymbol{\delta}\|_2^2)$.

Theorem 1 and the first part of theorem 2 demonstrate that the sparsity can be achieved after rotation even measured by the strong notion of l_0-norm. However, the weaker measure of sparsity by using the l_1-norm is needed to obtain more general results, as shown in the second part of theorem 2.

As a direct consequence of theorem 2, we have the following corollary.

Corollary 1. If $\varepsilon/\lambda_p = O\{\sqrt{(k/p)}\}$ and $\varepsilon\|\boldsymbol{\delta}\|_2^2/d\|\boldsymbol{\delta}_2\|_2^2 = O(k/p)$, then $\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_2 = O(\sqrt{k})$.

Note that the construction of $\mathbf{U}_\rho$ is independent of k, and conclusions of theorem 2 hold for any k satisfying the technical conditions. Define $d_k = \lambda_k - \lambda_{k+1}$, $\varepsilon_k = \lambda_{k+1} - \lambda_p$ and $\mathbf{W}_1^k = \mathrm{span}(\boldsymbol{\xi}_j)_{1\leqslant j\leqslant k}$ and $\mathbf{W}_2^k = \mathrm{span}(\boldsymbol{\xi}_j)_{k+1\leqslant j\leqslant p}$, $\boldsymbol{\delta} = \boldsymbol{\delta}_1^k + \boldsymbol{\delta}_2^k$ with $\boldsymbol{\delta}_m^k \in \mathbf{W}_m^k$, $m = 1, 2$. Let $\tilde{d}_k = d_k\rho\|\boldsymbol{\delta}_2^k\|_2^2/(d_k+\rho\|\boldsymbol{\delta}\|_2^2)$. Define

$$C_k = \sqrt{(k+1)} + \sqrt{(p-k-1)}\frac{\lambda_p+\varepsilon_k}{\lambda_p}\left\{\frac{\varepsilon_k}{\lambda_p} + \sqrt{\left(\frac{\varepsilon_k}{\tilde{d}_k-2\varepsilon_k}\right)}\right\}$$

if $\tilde{d}_k - 2\varepsilon_k > 0$, and $C_k = \infty$ otherwise. Theorem 2 implies the following corollary.

Corollary 2. If K is the least integer such that $\boldsymbol{\delta} \in \mathbf{W}_1^K$, then $\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}\|_2 \leqslant \min\{C, \sqrt{K}\}$, where $C = \min_{1\leqslant k<K}\{C_k\}$.

Theorems 1 and 2 show that the classification problem is reduced to a sparse problem after rotation by $\mathbf{U}_\rho^{\mathrm{T}}$ when the covariance structure is spiked. And the level of sparsity of $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ can be controlled by the spiked covariance structure (k and eigenvalue distribution in conditions 1 and 2).

Moreover, the procedure is invariant under orthonormal transformations. In other words, the normal vector of the optimal discriminant affine space after rotation, i.e. $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$, is invariant with respect to any rotation. Indeed, when the data are rotated by an arbitrary orthogonal matrix $\mathbf{V}$, then the new mean vectors and common covariance matrix are $\mathbf{V}\boldsymbol{\mu}_1$, $\mathbf{V}\boldsymbol{\mu}_2$ and $\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$. Since

$$\mathbf{D}_\rho = \mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\Sigma}_\rho^{\mathrm{tot}}\mathbf{U}_\rho = (\mathbf{V}\mathbf{U}_\rho)^{\mathrm{T}}\mathbf{V}\boldsymbol{\Sigma}_\rho^{\mathrm{tot}}\mathbf{V}^{\mathrm{T}}(\mathbf{V}\mathbf{U}_\rho),$$

the rotation matrix should be $(\mathbf{V}\mathbf{U}_\rho)^{\mathrm{T}}$, and the rotated normal vector $(\mathbf{V}\mathbf{U}_\rho)^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} = \mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$, which is independent of $\mathbf{V}$.

3. A rotate-and-solve procedure

In this section, we introduce a two-stage RS procedure for classification. The idea is to mimic the oracle rotations in the previous section and to rotate the data such that $\boldsymbol{\beta}$ is nearly sparse. Namely, we first use the orthogonal matrix $\hat{\mathbf{U}}_\rho$, consisting of the eigenvectors of the empirical total covariance $\hat{\boldsymbol{\Sigma}}_\rho^{\mathrm{tot}} = \hat{\boldsymbol{\Sigma}} + \rho\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\mathrm{T}}$ to rotate the data and then to apply sparse LDA methods such as the ROAD and LPD to the rotated data.

Let $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ be the sample mean vectors of classes 1 and 2 respectively. Set

$$\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2,$$
$$\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2.$$

Similarly, let $\hat{\boldsymbol{\Sigma}}^{(1)}$ and $\hat{\boldsymbol{\Sigma}}^{(2)}$ be their sample covariance matrices and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1+n_2}(n_1\hat{\boldsymbol{\Sigma}}^{(1)} + n_2\hat{\boldsymbol{\Sigma}}^{(2)})$$

be the pooled sample covariance matrix. The degree of freedom can be adjusted, but the version of the maximum likelihood estimate is used here to facilitate the expression in remark 1 below. We then estimate $\boldsymbol{\Sigma}_\rho^{\mathrm{tot}}$ by

$$\hat{\boldsymbol{\Sigma}}_\rho^{\mathrm{tot}} = \hat{\boldsymbol{\Sigma}} + \rho\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\mathrm{T}}.$$

Perform singular value decomposition

$$\hat{\mathbf{U}}_\rho^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_\rho^{\mathrm{tot}}\hat{\mathbf{U}}_\rho = \hat{\mathbf{D}}_\rho, \tag{3.1}$$

where $\hat{\mathbf{D}}_\rho = \mathrm{diag}(\hat{\eta}_1, \ldots, \hat{\eta}_p)$ is the diagonal matrix with sorted eigenvalues.

The two-stage RS procedure can be implemented as follows.

Stage 1: calculate $\hat{\mathbf{U}}_\rho$ and rotate the data to obtain $\{\hat{\mathbf{U}}_\rho^{\mathrm{T}}\mathbf{X}_i^{(m)}\}_{i=1}^{n_m}$ for $m=1$ and $m=2$.
Stage 2: apply the ROAD, LPD or other sparse LDA methods to the rotated data $\mathcal{X}\hat{\mathbf{U}}_\rho$ to obtain a prediction rule.

Remark 1. Define

$$\bar{\mathbf{X}} = \frac{1}{n_1+n_2}(n_1\bar{\mathbf{X}}^{(1)} + n_2\bar{\mathbf{X}}^{(2)}),$$
$$\hat{\boldsymbol{\Sigma}}_{\mathrm{sample}}^{\mathrm{tot}} = \frac{1}{n_1+n_2}\sum_{m=1}^{2}\sum_{i=1}^{n_m}(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}})(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}})^{\mathrm{T}}$$

which is the sample total covariance (ignoring the classes). It is straightforward to see that

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{sample}}^{\mathrm{tot}} = \hat{\boldsymbol{\Sigma}} + \frac{n_1 n_2}{(n_1+n_2)^2}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\mathrm{T}}.$$

When $p \ll n = n_1 + n_2$, $\hat{\mathbf{U}}_\rho$ and $\mathbf{U}_\rho$ are similar when the eigenvalues are separated from each other, and hence $\hat{\mathbf{U}}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ is similar to $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$. The property of $\hat{\mathbf{U}}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ is much more complicated when $p \sim n$ or $p \gg n$. In this case, it is difficult to guarantee that all estimated eigenvectors are close to the true eigenvectors. However, the eigenvectors that correspond to spiked eigenvalues can be consistently estimated. See for example Zou *et al.* (2006), Karoui (2008), Johnstone and Lu (2009), Agarwal *et al.* (2012), Fan *et al.* (2013) and Shen *et al.* (2013). As these eigenvectors point at most important directions, the consistent estimation of these directions ensures the correct rotations in these important directions. This explains our empirical results that the RS procedure performs very well compared with several state of the art methods, even when $p \gg n$.

To understand better the mathematics behind the excellent performance of the RS procedure, the classification error of the idealized Fisher classifier depends on $\gamma \equiv \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. Let $\mathbf{U}_{\rho 1}$ be a $(k+1) \times p$ matrix, consisting of the eigenvectors of $\boldsymbol{\Sigma}_\rho^{\mathrm{tot}}$ that correspond to the largest $k+1$ eigenvalues $\{\eta_j\}_{j=1}^{k+1}$. If we restrict the information to the first $k+1$ dimensions of the rotated data $\mathbf{U}_{\rho 1}^{\mathrm{T}}\mathbf{X}|_m \sim N(\mathbf{U}_{\rho 1}^{\mathrm{T}}\boldsymbol{\mu}_m, \mathbf{U}_{\rho 1}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}_{\rho 1})$, $m = 1, 2$, then the classification error depends on

$$\gamma_1 \equiv (\mathbf{U}_{\rho 1}^{\mathrm{T}}\boldsymbol{\delta})^{\mathrm{T}}(\mathbf{U}_{\rho 1}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}_{\rho 1})^{-1}(\mathbf{U}_{\rho 1}^{\mathrm{T}}\boldsymbol{\delta}).$$

Clearly, $\gamma_1 \leqslant \gamma$. How much is the information loss when $\{\eta_j\}_{j=1}^{k+1}$ are spiked? Under the conditions in theorem 1, there is no loss of information if the first $k+1$ most important features are used. Furthermore, Zou *et al.* (2006), Karoui (2008), Johnstone and Lu (2009), Agarwal *et al.* (2012), Fan *et al.* (2013) and Shen *et al.* (2013) give the conditions under which $\mathbf{U}_{\rho 1}$ can be consistently estimated.

The above argument is based on the fact that $\mathbf{U}_{\rho 1}^{\mathrm{T}}\boldsymbol{\delta}$ preserves the energy of $\boldsymbol{\delta}$. The result holds more generally for the covariance matrix $\boldsymbol{\Sigma}$ admitting spiked eigenvalues, including covariance matrices derived from approximate factor models (Fan *et al.*, 2013) or admitting low rank plus sparse matrix decomposition (Agarwal *et al.*, 2012). Recall that $\boldsymbol{\Sigma} = \Sigma_{i=1}^{p}\lambda_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}$ with $\boldsymbol{\xi}_i$ being the eigenvector of $\boldsymbol{\Sigma}$. Let $\lambda_i(\mathbf{B})$ be the ith largest eigenvalue of a symmetric matrix $\mathbf{B}$.

Theorem 3. If $\lambda_{k+1}(\Sigma_{i=1}^{k}\lambda_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}} + \rho\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}) > a\lambda_{k+1}$ for some $a > 2$, then

$$\|\mathbf{U}_{\rho 1}\boldsymbol{\delta}\|_2 \geqslant \frac{a-2}{a-1}\|\boldsymbol{\delta}\|_2$$

and

$$\gamma_1 \geqslant \frac{(a-2)^2}{(a-1)^2\lambda_1}\|\boldsymbol{\delta}\|_2^2.$$

The condition in theorem 3 holds relatively easily. We can take $k = 0$ when $\rho\|\boldsymbol{\delta}\|_2^2 \geqslant a\|\boldsymbol{\Sigma}\|^2$. This holds easily by taking a sufficiently large ρ.

Note that $\gamma \leqslant \lambda_p^{-1}\|\boldsymbol{\delta}\|_2^2$ and γ is usually significantly smaller than this upper bound. Therefore, when λ_1/λ_p is bounded, the loss of information by using rotated data is limited. Yet, we reduce significantly the accumulation of noise in classification (Fan and Fan, 2008). As noted above, the rotation $\mathbf{U}_{\rho 1}$ can be consistently estimated by regularization. These together provide theoretical endorsement of the advantages of using rotation.

Remark 2 (dimensionality reduction). When $p > n$, $\hat{\mathbf{U}}_\rho$ is not unique since $\hat{\boldsymbol{\Sigma}}_\rho^{\mathrm{tot}}$ is singular. (The null space of $\hat{\boldsymbol{\Sigma}}_\rho^{\mathrm{tot}}$ is large and we can choose an arbitrary basis of the null space as the columns of $\hat{\mathbf{U}}_\rho$.) Since the last $p-n$ columns in $\hat{\mathbf{U}}_\rho$ are arbitrary and cannot be controlled, we define $\tilde{\mathbf{U}}_\rho$ as the first n columns (or even fewer) of $\hat{\mathbf{U}}_\rho$ and conduct classification on the rotated data $\{\tilde{\mathbf{U}}_\rho^{\mathrm{T}}\mathbf{X}_i^{(m)}\}_{i=1}^{n_m}$ for $m = 1$ and $m = 2$. From the theoretical analysis in the previous section, we see that, under ideal conditions, $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ is sparse with non-vanishing part concentrated

on the first $k+1$ components. This implies that only the first $k+1$ columns of the rotated data are useful to estimate $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$, which motivates us to use $\tilde{\mathbf{U}}_\rho$ instead of $\hat{\mathbf{U}}_\rho$ as a practical approach with reduced dimensionality. Theorem 3 further shows that the loss of classification power due to this dimensionality reduction is limited. Let $\tilde{\psi}$ be a classification rule constructed by some (fixed) sparse LDA method based on $\tilde{\mathcal{X}} = \mathcal{X}\tilde{\mathbf{U}}_\rho$. It is straightforward to see that $\tilde{\psi}$ is equivariant.

Remark 3 (computation of transform). When $p > n$, the computation of $\tilde{\mathbf{U}}_\rho$ can be performed as follows. First of all, $\hat{\boldsymbol{\Sigma}}_\rho^{\mathrm{tot}}$ can be written as $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ for a given $(n+1) \times p$ matrix $\mathbf{Y}$ (suitable scaling of centred observations and sample mean). Note that $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ and $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ have the same non-vanishing eigenvalues. Let $\tilde{\mathbf{U}}_\rho = \mathbf{Y}^{\mathrm{T}}\hat{\mathbf{V}}$, where $\hat{\mathbf{V}}$ is the orthogonal matrix consisting of eigenvectors of non-vanishing eigenvalues of the $(n+1) \times (n+1)$ matrix $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$. Then, the columns of $\tilde{\mathbf{U}}_\rho$ contain the eigenvectors of non-vanishing eigenvalues of $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ and are orthogonal. In other words, $\tilde{\mathbf{U}}_\rho$ can be used to transform the data. The reduction of computation cost is significant when $p \gg n$, since the singular value decomposition of $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ is much faster.

Remark 4 (sensitivity of ρ). Our empirical studies show that the RS procedure is not sensitive to ρ in a broad range. For a large range of choices of ρ, the classification errors are significantly improved over the existing LDA algorithms, as will be shown by our numerical experiments in the next section. Ideally, ρ can be estimated by using data-adaptive methods such as cross-validation. However, cross-validation on ρ may be computationally intractable for high dimensional data where p is huge. As noted from remark 3, we may use $\tilde{\mathbf{U}}_\rho$ to rotate the data, which reduces the dimension from p to n. Thus cross-validation on ρ is more tractable by using the modified RS procedure, and the classification quality can be noticeably improved as will be shown by our numerical experiments.

4. Numerical studies

In this section, we compare the RS procedure with several popular LDA-based methods including standard LDA (1.3) (using the Moore–Penrose pseudoinverse when $\hat{\boldsymbol{\Sigma}}$ is singular), IR, nearest shrunken centroids (NSCs) (Tibshirani *et al.*, 2002), the ROAD and LPD, via simulation and real data examples. For the RS procedure, two variants RS-ROAD and RS-LPD are included. For simulated examples using the toy models, we also consider the oracle RS methods (O-RS-ROAD and O-RS-LPD) where the oracle rotations that were shown in Section 2 are used to rotate the data. Moreover, the oracle Fisher rule (1.2) is used as a benchmark method. In all RS-related methods, the parameter ρ is fixed to $\frac{1}{2}$ unless explicitly defined. The same numbers of observations are generated for both classes for all simulated data in Section 4.1, i.e. $n_1 = n_2$. All simulation settings were repeated 100 times unless noted otherwise.

4.1. Simulated data

4.1.1. Toy models

We begin with several toy models with relatively small n and p to illustrate the performance of the RS procedures *versus* the aforementioned LDA methods. We consider the following three toy models:

(a) *toy model 1*, $\boldsymbol{\Sigma} = \mathbf{I}_p$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = a_1\mathbf{1}_p$;
(b) *toy model 2*, $\boldsymbol{\Sigma} = (\sigma_{i,j})$ with $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.5$ for $i \neq j$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = (a_2\mathbf{1}_l^{\mathrm{T}}, \mathbf{0}_{p-l}^{\mathrm{T}})^{\mathrm{T}}$, where $l = 5$;
(c) *toy model 3*, the setting is the same as for model 2 except that $l = p/2$ and $\boldsymbol{\mu}_2 = (a_3\mathbf{1}_l^{\mathrm{T}}, \mathbf{0}_{p-l}^{\mathrm{T}})^{\mathrm{T}}$.

The values of a_1, a_2 and a_3 in each of the toy models are chosen such that the expected clas-

sification errors of the oracle Fisher rule (1.2) are 1%, 5% and 10%. For each model, we take $p = 50$ and $n_1 = 20$ or $n = 30$. The same numbers of observations were collected independently as the testing set.

We apply the IR, standard LDA, NSCs, the ROAD, LPD, RS-ROAD, RS-LPD, O-RS-ROAD and O-RS-LPD to 100 replicates of every simulation scenario. Simulation results are presented in Fig. 1 (for $n_1 = 20$) and Fig. 2 (for $n_1 = 30$). The oracle rule always performs best and gives a benchmark for the other methods. The O-RS methods perform very well and are comparable with the oracle rule. For toy model 1, the features are independent, so the IR performs best besides the oracle rule. But RS methods are comparable with the IR. For model 2, the true $\boldsymbol{\beta}$ is nearly sparse. Therefore, the ROAD and LPD perform well but RS methods still improve on their performance. For toy model 3, neither the covariance matrix nor true $\boldsymbol{\beta}$ is sparse. RS methods work significantly better than their competitors. We observe that the RS methods are uniformly good in all the three models.

To see why RS methods outperform their direct sparse competitors, we plot the percentages of the sum of squares of the first several largest components of the true $\boldsymbol{\beta}$ before and after rotation. For a rotation $R = \mathbf{U}_\rho$ or $R = \hat{\mathbf{U}}_\rho$, define $\boldsymbol{\beta}^R = R^{\mathrm{T}}\boldsymbol{\beta}$. Denote by $|\beta|_{(1)}, \ldots, |\beta|_{(p)}$ and $|\beta^R|_{(1)}, \ldots, |\beta^R|_{(p)}$ the reversed order statistics (from largest to smallest) of $\{|\beta_j|\}_{j=1}^p$ and $\{|\beta_j^R|\}_{j=1}^p$ respectively. For each setting, we plot $\Sigma_{i=1}^k |\beta|_{(i)}^2/\|\boldsymbol{\beta}\|_2^2$, $\Sigma_{i=1}^k |\beta^{\mathbf{U}_\rho}|_{(i)}^2/\|\boldsymbol{\beta}^{\mathbf{U}_\rho}\|_2^2$ and $(1/100)\Sigma_{j=1}^{100}\Sigma_{i=1}^k |\beta^{\hat{\mathbf{U}}_{\rho j}}|_{(i)}^2/\|\boldsymbol{\beta}^{\hat{\mathbf{U}}_{\rho j}}\|_2^2$ for $k = 1, \ldots, p$, where $\hat{\mathbf{U}}_{\rho j}$ is the rotation matrix for the jth replicate and $\mathbf{U}_\rho$ is the oracle rotation matrix. In Fig. 3, we see that, after rotation, $\boldsymbol{\beta}$ is more concentrated in its largest components. $\mathbf{U}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ is extremely sparse and $\hat{\mathbf{U}}_\rho^{\mathrm{T}}\boldsymbol{\beta}$ is sparser than the original $\boldsymbol{\beta}$. Obviously, the ROAD or LPD is more efficient after the rotation.

4.1.2. More simulations

In our next numerical simulations, we consider the following three covariance structures:

(a) *model 1*, $\boldsymbol{\Sigma} = (\sigma_{i,j})$ with $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.5$ for $i \neq j$;
(b) *model 2*, $\boldsymbol{\Sigma} = (\sigma_{i,j})$ with $\sigma_{i,j} = 0.7^{|i-j|}$;
(c) *model 3*, $\boldsymbol{\Sigma} = \mathbf{I} + \mathbf{A}\mathbf{A}^{\mathrm{T}}$ where $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ is a $p \times 5$ matrix with entries generated independently from $\mathcal{N}(0, 1)$.

Without loss of generality, we set $\mu_1 = \mathbf{0}$ and $\mu_2 = (a\mathbf{1}_{p/2}^{\mathrm{T}}, \mathbf{0}_{p/2}^{\mathrm{T}})^{\mathrm{T}}$, where a is chosen specifically for each model such that the expected classification error of the oracle rule is 2%. Similarly to before, for each simulation, we generate $2n_1$ independent observations for each class, where n_1 observations are used as training data and the other n_1 observations are used for testing. Results for models 1, 2 and 3 are presented in Figs 4, 5 and 6 respectively, with various sample sizes and dimensionality. For models 1 and 3 where the covariance structure is spiked, the improvement of the RS methods over the ROAD or LPD is remarkable. For model 2 where the covariance structure is far from being spiked, the RS methods still generally outperform their counterparts.

To show that the improvement by applying RS is relatively general, we consider the following two scenarios with randomly generated covariance matrices:

(a) *random model 1*, $\boldsymbol{\Sigma} = (M/\|M\|)^{\mathrm{T}}(M/\|M\|) + \mathrm{diag}(v)$ with each entry of $p \times p$ matrix M being generated independently from $N(0, 1)$ and v from $\mathcal{U}(0, 1)$, where $|M|$ is the operator norm of M;
(b) *random model 2*, $\boldsymbol{\Sigma} = 4(M/\|M\|)^{\mathrm{T}}(M/\|M\|)$ with each entry of M being generated independently from $N(0, 1)$.

We fix $n_1 = 30$ and $p = 300$ and consider different levels of sparsity of $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_0/p = 5\%$, $10\%, \ldots, 95\%, 100\%$. We randomly generate $\boldsymbol{\beta}$ with a given level of sparsity, whose non-zero

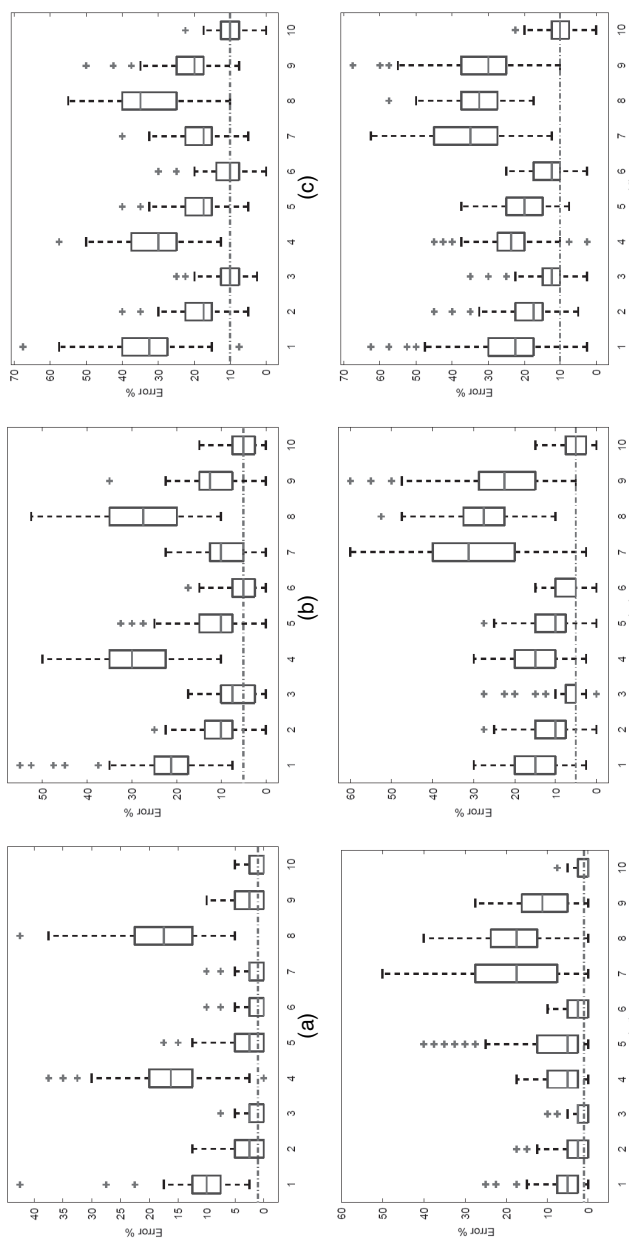


Fig. 1. Simulation results for the three toy models (a)–(c) 1, (d)–(f) 2 and (g)–(i) 3 with $n_1 = 20$ and $p = 50$ (1, ROAD; 2, RS-ROAD; 3, O-RS-ROAD; 4, LPD; 5, RS-LPD; 6, O-RS-LPD; 7, IR; 8, standard LDA; 9, NSCs; 10, oracle): (a), (d), (g) oracle error 1%; (b), (e), (h) oracle error 5%; (c), (f), (i) oracle error 10%

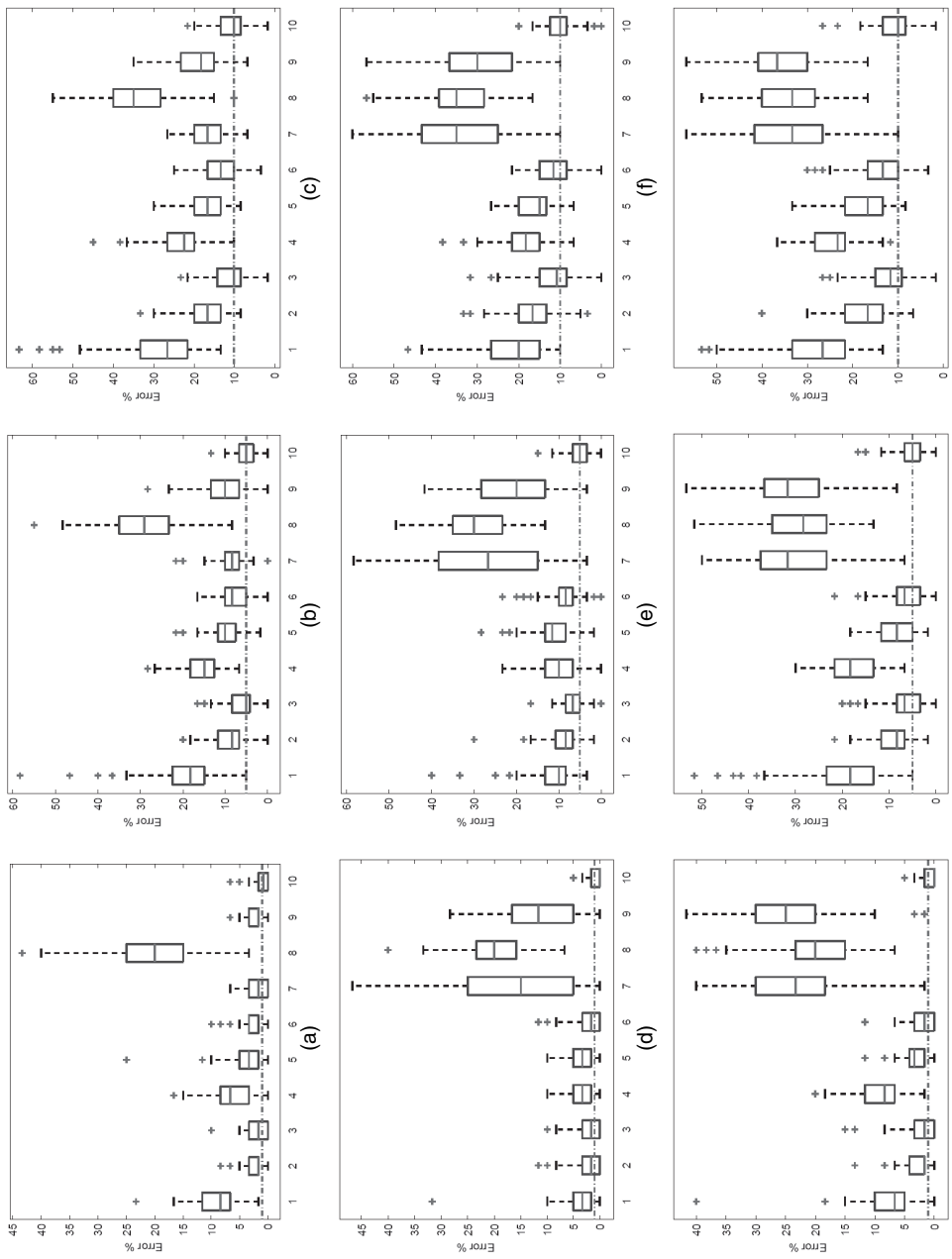


Fig. 2. Simulation results for the three toy models (a)–(c) 1, (d)–(f) 2 and (g)–(i) 3 with $n_1 = 30$ and $p = 50$ (1, ROAD; 2, RS-ROAD; 3, O-RS-ROAD; 4, LPD; 5, RS-LPD; 6, O-RS-LPD; 7, IR; 8, standard LDA; 9, NSC; 10, oracle): (a), (d), (g) oracle error 1%; (b), (e), (h) oracle error 5%; (c), (f), (i) oracle error 10%

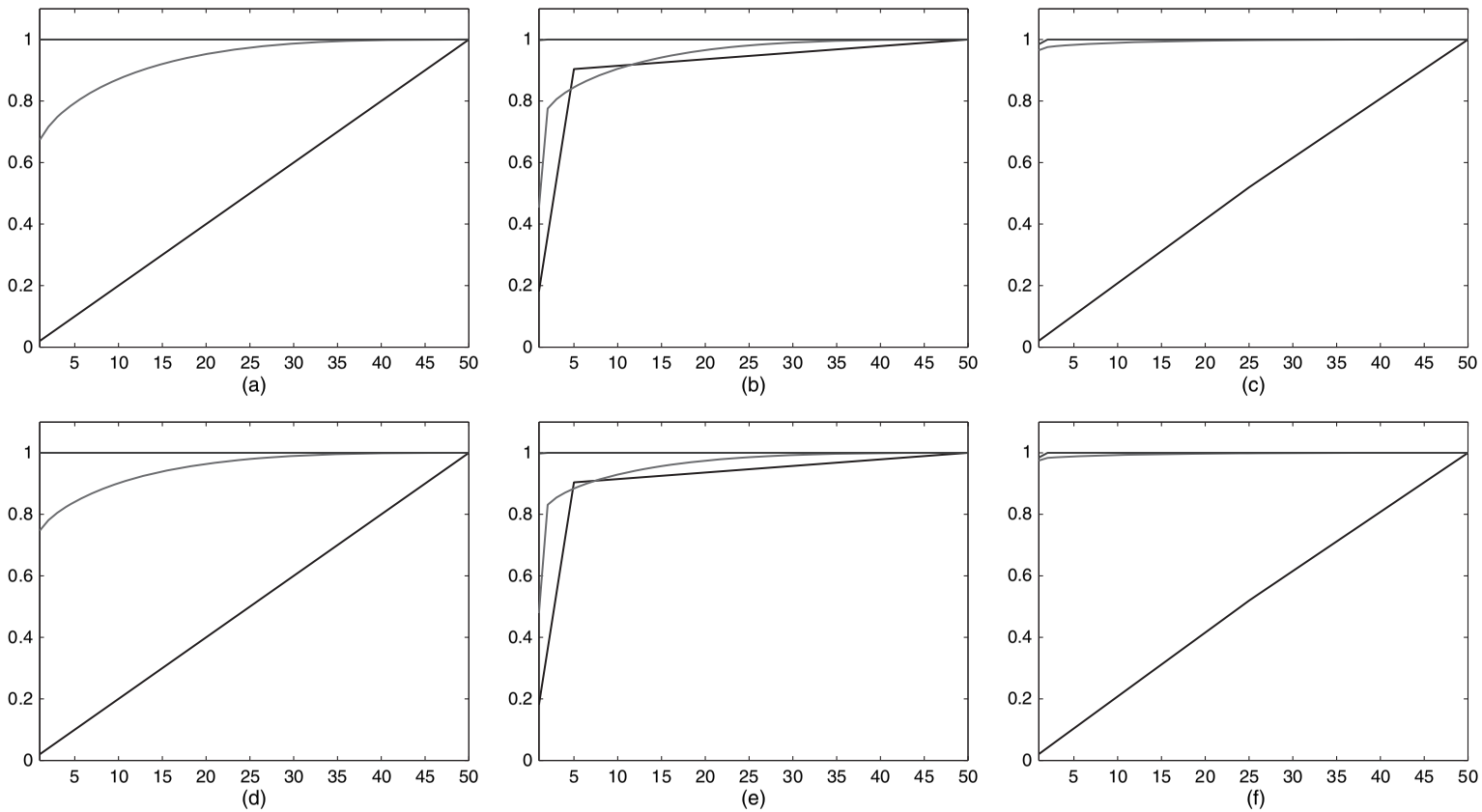


Fig. 3. Sparsity levels of β before (———) and after rotation using $\mathbf{U}_\rho$ (———) and $\hat{\mathbf{U}}_\rho$ (———) for (a)–(c) $n_1 = 20$ and $p = 50$ and (d)–(f) $n_1 = 30$ and $p = 50$: (a), (d) toy model 1; (b), (e) toy model 2; (c), (f) toy model 3

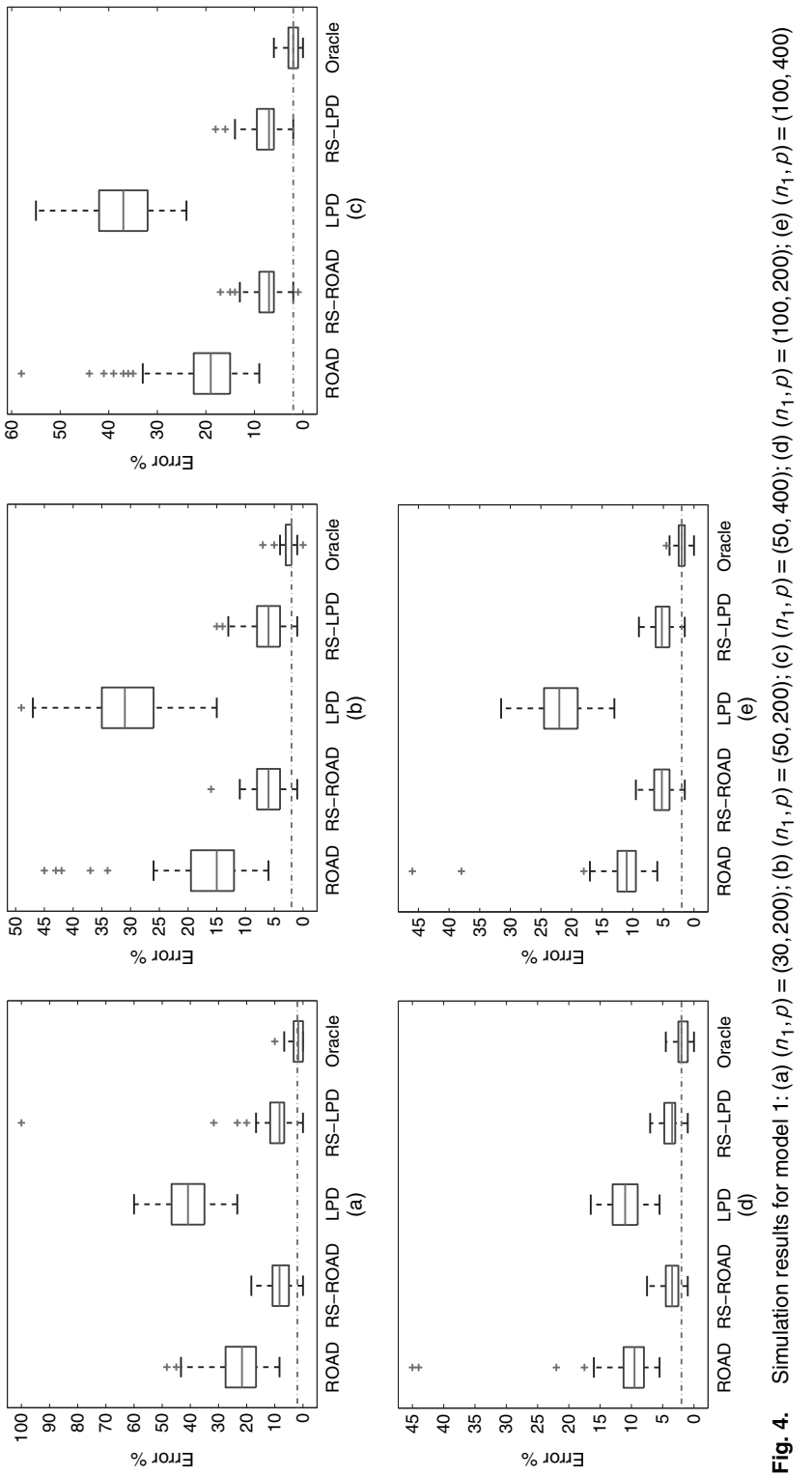


Fig. 4. Simulation results for model 1: (a) $(n_1, p) = (30, 200)$; (b) $(n_1, p) = (50, 200)$; (c) $(n_1, p) = (50, 400)$; (d) $(n_1, p) = (100, 200)$; (e) $(n_1, p) = (100, 400)$

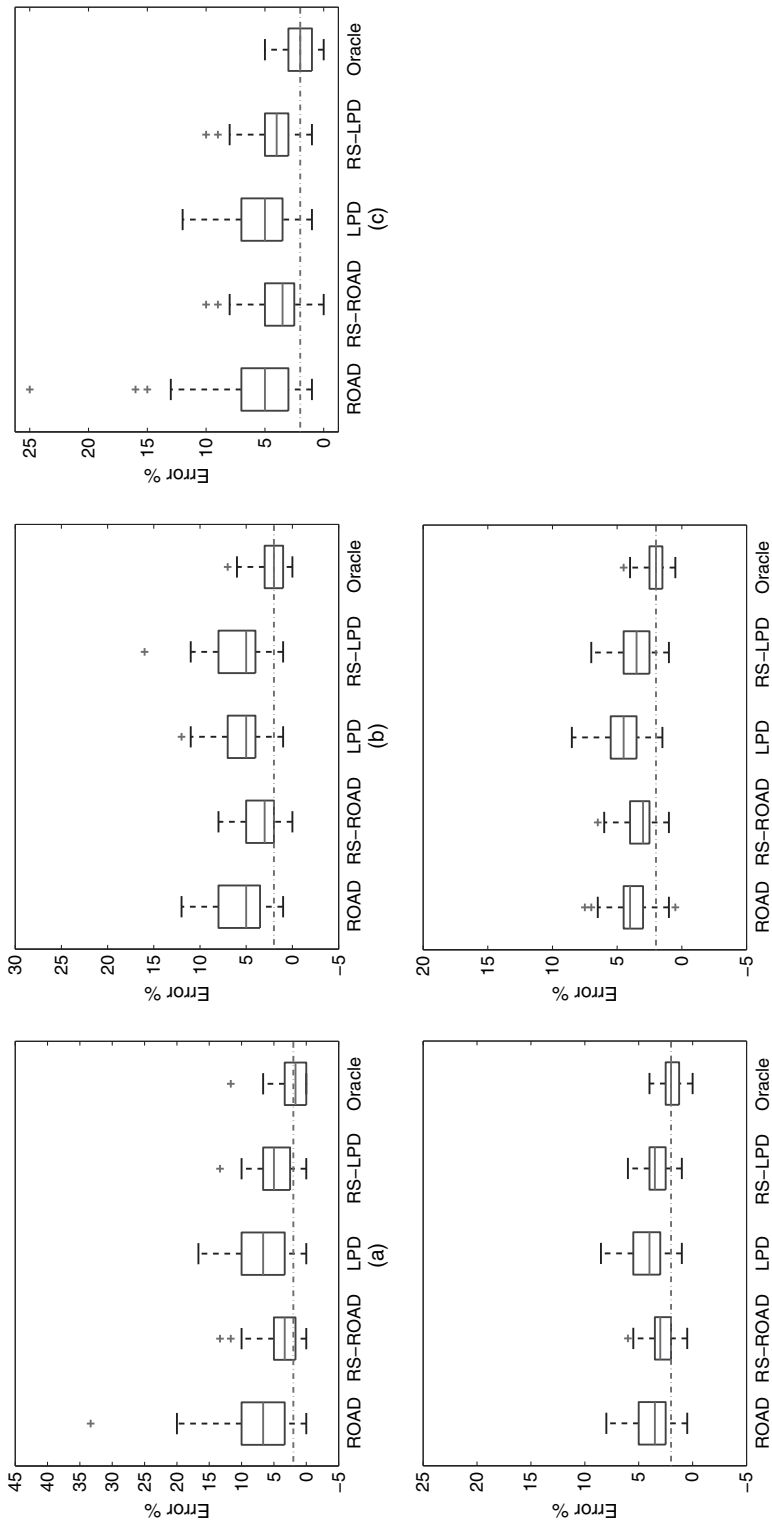


Fig. 5. Simulation results for model 2: (a) $(n_1, p) = (30, 200)$; (b) $(n_1, p) = (50, 200)$; (c) $(n_1, p) = (50, 400)$; (d) $(n_1, p) = (100, 200)$; (e) $(n_1, p) = (100, 400)$

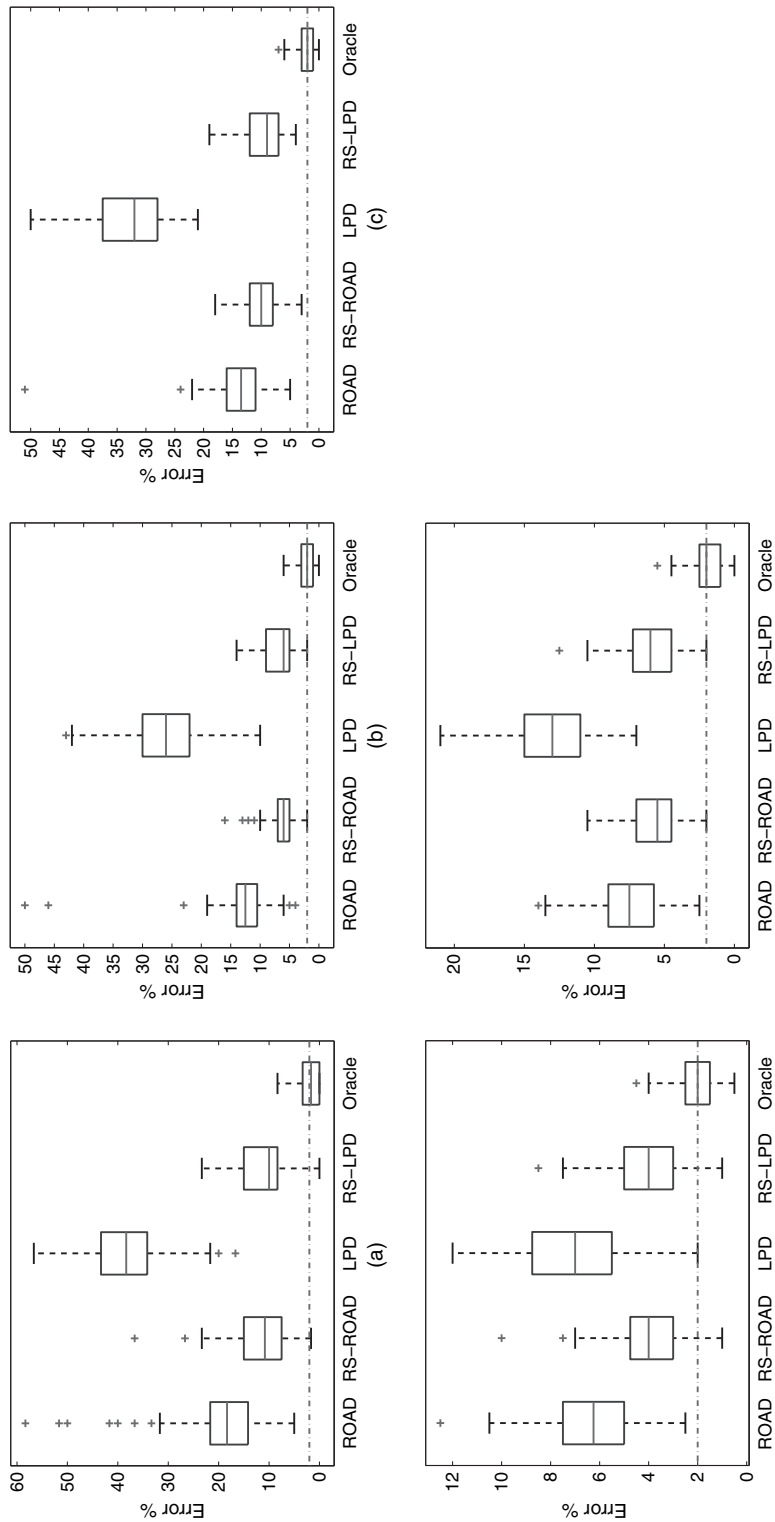


Fig. 6. Simulation results for model 3: (a) $(n_1, p) = (30, 200)$; (b) $(n_1, p) = (50, 200)$; (c) $(n_1, p) = (50, 400)$; (d) $(n_1, p) = (100, 200)$; (e) $(n_1, p) = (100, 400)$

entries are independent and identically distributed from $N(0, 1)$. We then normalize β such that $\beta^T \Sigma \beta = 12$. We fix $\mu_1 = \mathbf{0}$ and let $\mu_2 = -\Sigma\beta$. We repeat our data generation and classification 100 times for each scenario and record the average classification errors and their standard deviations.

We compare the results of the ROAD and RS-ROAD in Fig. 7. As we can see when β is very sparse, the ROAD outperforms RS-ROAD as expected. However, the performance of the ROAD depends greatly on the level of sparsity. In contrast, RS-ROAD has significantly smaller overall error rates, and has the same qualitative behaviour as the oracle. In particular, RS-ROAD is robust against the level of sparsity of the true data-generating procedure.

4.2. Real data: leukaemia and lung cancer

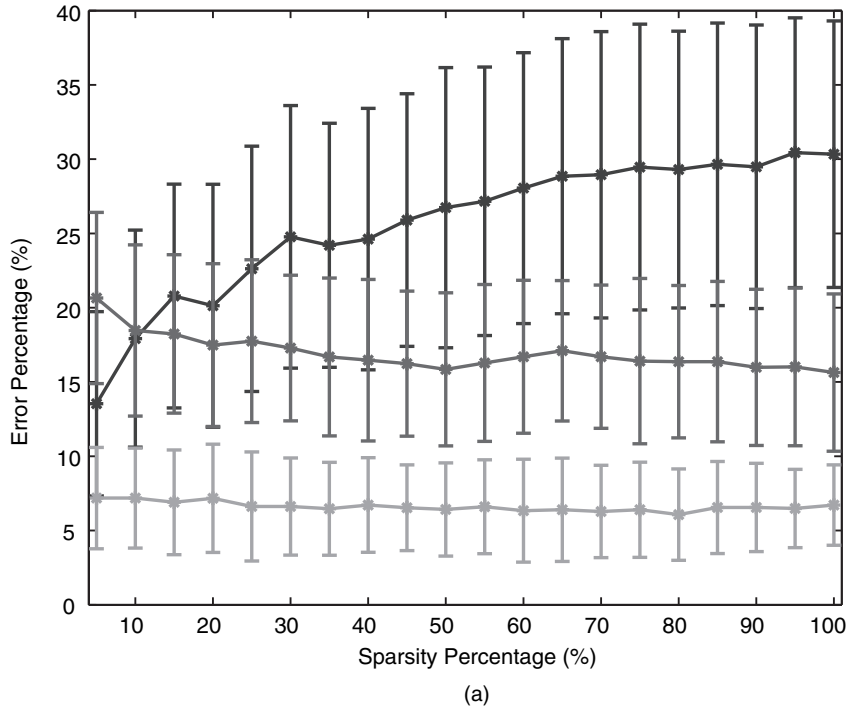
We now evaluate the performance of our proposed RS procedure on two popular gene expression data sets: the leukaemia (Golub *et al.*, 1999) and lung cancer (Gordon *et al.*, 2002) data. The two data sets come with separate training and testing sets of data vectors. The leukaemia data set contains $p = 7129$ genes with $n_1 = 27$ acute lymphoblastic leukaemia and $n_2 = 11$ acute myeloid leukaemia vectors in the training set. The testing set includes 20 acute lymphoblastic and 14 acute myeloid leukaemia vectors. The lung cancer data set contains $p = 12\,533$ genes with $n_1 = 16$ adenocarcinoma and $n_2 = 16$ mesothelioma training vectors. The testing set has 134 adenocarcinoma and 15 mesothelioma vectors. These two data sets can be downloaded from the Web sites `http://www.bioconductor.org/packages/release/data/experiment/html/golubEsets.html` and `http://www.chestsurg.org/publications/2002-microarray.aspx` respectively.

In our experiments, we put all the 47 (27 training plus 20 testing data) acute lymphoblastic leukaemia vectors and 25 (11 training plus 14 testing data) acute myeloid leukaemia vectors together and randomly select 23 acute lymphoblastic leukaemia and 12 acute myeloid leukaemia as training and the rest as testing data. We repeat the experiments 20 times. We conduct a similar experiment on the lung cancer data by randomly selecting 75 adenocarcinoma and 15 mesothelioma data vectors as training and the rest as testing, and repeat 20 times. The classification results of the aforementioned experiments using the IR, NSC, the ROAD and RS-ROAD are presented in Table 1, where RS-ROAD has the best overall performance.

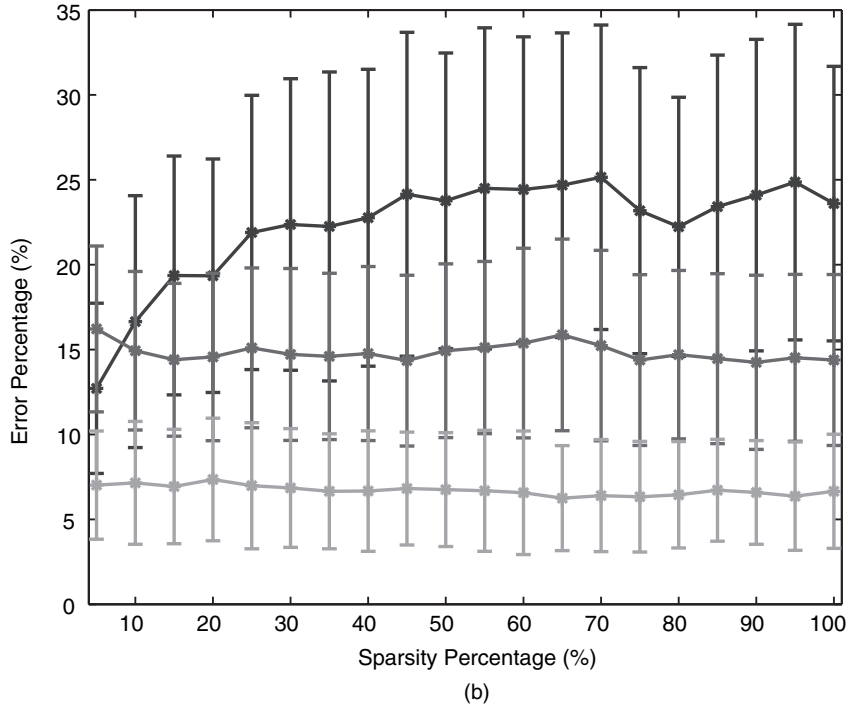
4.3. Real data: shape classifications

We also evaluate the performance of RS on shape classification, which is one of the most fundamental and important problems in computer vision and machine learning. All the shapes are represented by two-dimensional binary images. We downloaded the MPEG-7 CE shape-1 part-B data set (Thakoor *et al.*, 2007) from the Web site `http://visionlab.uta.edu/shape_data.htm` and selected a subset of it for our tests. Since the images in the data set generally have different sizes, we resized them to the same size 50×50 (i.e. $p = 2500$) by using the MATLAB command `imresize` with bicubic interpolation. All the selected and resized shape images are shown in Fig. 8.

There are 20 images for each shape class. After being loaded, each image is a matrix, with elements taking integer values in $[0, 255]$. To test the robustness of the classifiers, we also added Gaussian noise $N(0, 50^2)$ to all the images selected. For every pair of shapes, we randomly select 10 from each class as testing data and the rest as training data (i.e. $n_1 = n_2 = 10$). We repeat this 50 times for each of the shape pairs. The average classification errors by the IR, NSCs, the ROAD and RS-ROAD are summarized in Table 2. We observe that RS-ROAD has the best overall performance, and it consistently improves on ROAD in all scenarios.



(a)



(b)

Fig. 7. Average classification errors of (a) random model 1 and (b) random model 2 for $\|\beta\|_0/p = 5\%, 10\%, \ldots, 95\%, 100\%$, for $n = 30$ and $p = 200$ ($\perp$, standard deviations of classification errors across 100 simulations): ———, ROAD; ———, RS-ROAD; ———, oracle

Table 1. Classification errors for the cancer data

Data set	*Errors (%) (and standard deviations (%)) for the following methods:*			
	IR	*NSCs*	*ROAD*	*RS-ROAD*
Leukaemia	4.2708 (2.9998)	8.5135 (8.4232)	6.3514 (5.9650)	4.4595 (3.0721)
Lung cancer	3.4669 (1.4381)	10.4396 (7.2675)	1.3736 (1.0621)	0.9341 (0.8931)

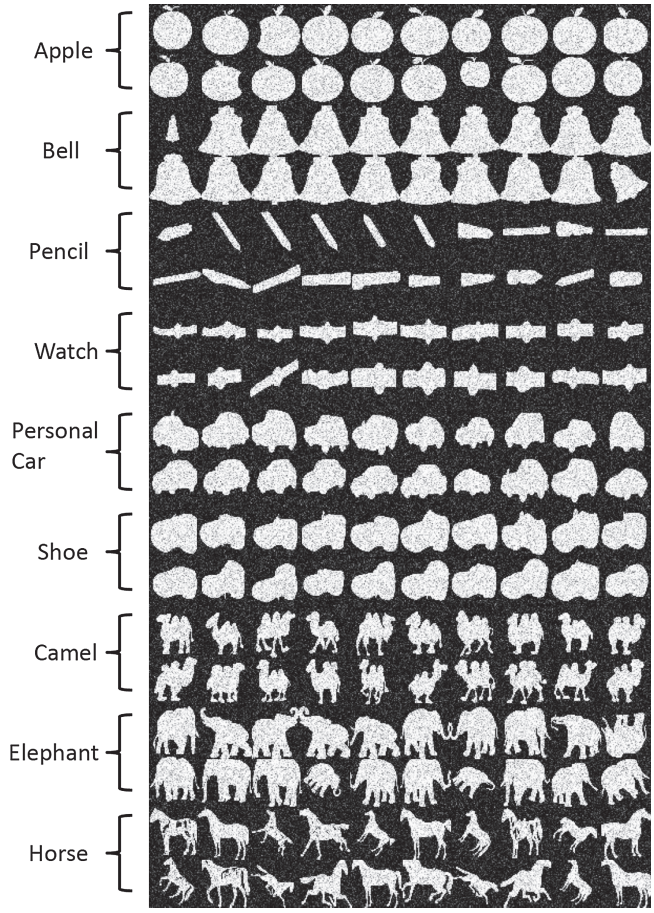


Fig. 8. Selected shape images, resized to 50 × 50 with additive Gaussian noise

4.4. Choice of ρ

Here we shall mainly discuss two issues related to the choice of ρ in $\mathbf{\Sigma}_{\rho}^{\mathrm{tot}}$:

(a) the sensitivity of the classification results to the choices of ρ;
(b) data-adaptive selection of ρ by cross-validation.

4.4.1. Sensitivity to ρ

In the following simulations, we take the toy models 1–3 with a_is chosen such that the oracle

Table 2. Classification errors for shapes

Shape pair	*Errors (%) for the following methods:*							
	IR		*NSCs*		*ROAD*		*RS-ROAD*	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
Apple and bell	7.9	3.0	7.7	3.1	8.3	4.5	7.8	3.4
Pencil and watch	19.2	6.1	20.4	7.1	18.2	6.9	16.0	7.1
Personal car and shoe	7.7	5.1	11.3	6.6	13.2	6.9	6.1	4.2
Camel and elephant	8.8	6.6	12.1	9.1	20.3	11.0	6.9	4.0
Camel and horse	9.6	7.1	11.8	9.5	22.0	10.6	7.7	5.6
Elephant and horse	8.8	6.4	11.9	8.4	15.1	10.1	6.9	5.9

error rate is 10%, and we use the method RS-ROAD as an example. Let $\hat{\mathbf{U}}_\rho$ be the eigenvectors of $\hat{\boldsymbol{\Sigma}} + \rho\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\mathrm{T}}$ with various values of ρ. The average classification errors (among 100 replicates) of method RS-ROAD with various ρ are shown in Fig. 9, where the dark curves show the errors that are associated with ρ and the light horizontal lines indicate the errors of ROAD. As we can see, the best choice of ρ depends on the scenario. Although it seems that choosing ρ optimally is a complicated issue, the plots in Fig. 9 indicate that, for a large range of ρ, the classification results are significantly improved over a non-rotated classifier such as the ROAD. This also indicates the robustness of the RS procedure to the choices of the parameter ρ. In general, any reasonable positive value of ρ should work well in most applications (Fig. 9 shows the workable range of $\log(\rho) \in [-1, 10]$), if one does not have the resources or time to perform cross-validation.

4.4.2. Cross-validation choice of ρ

Cross-validation on ρ is computationally expensive when p is large. See remark 4 for reduction of computation. When $\boldsymbol{\Sigma}$ has a (quasi-)spiked covariance structure, i.e. there are k eigenvalues that are significantly larger than the other $p-k$ eigenvalues, and if k is much less than the number of observations n, then we may use $\tilde{\mathbf{U}}_\rho$ to rotate the data instead of using $\hat{\mathbf{U}}_\rho$. Recall that $\tilde{\mathbf{U}}_\rho$ is the collections of the n eigenvectors of $\hat{\boldsymbol{\Sigma}}^{\mathrm{tot}}$ corresponding to the n largest eigenvalues. Then, after rotating the data by using $\tilde{\mathbf{U}}_\rho$, we reduce the dimension of the problem from p to n, which will be a significant reduction when $n \ll p$ (e.g. the real data considered in the previous two sections). We can also take $\tilde{\mathbf{U}}_\rho$ to be principal components, with dimensionality much less than n.

Our first simulations show that using $\tilde{\mathbf{U}}_\rho$ instead of $\hat{\mathbf{U}}_\rho$ does not hurt the classification error. We take the toy models 1–3 with a_is chosen such that the oracle error rate is 10%, and we use the method RS-ROAD as an example. We set $n_1 = n_2 = 10$ (i.e. $n = 20$) and $p = 50$. The results are summarized in Table 3.

The previous simulation shows that we can reduce the size of the problem from p to n without sacrificing much of the quality of classification. Since the computational cost can be greatly reduced in this way, cross-validation on ρ is now a computationally viable approach. In our next experiments, we take the leukaemia and lung cancer data in Section 4.2 and conduct an experiment that is similar to the experiment that we did before, except that we use $\tilde{\mathbf{U}}_\rho$ and choose

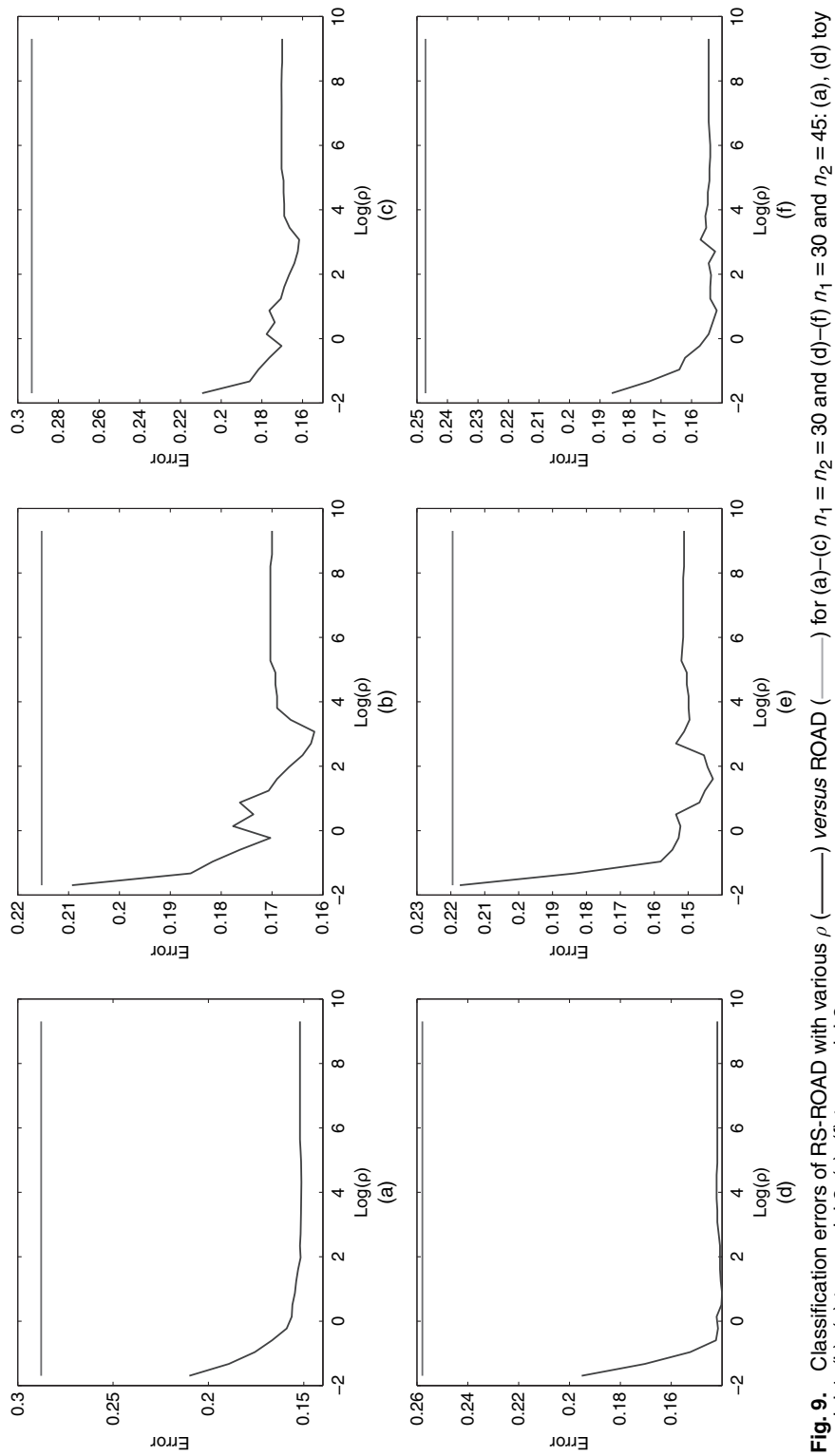


Fig. 9. Classification errors of RS-ROAD with various ρ (———) *versus* ROAD (———) for (a)–(c) $n_1 = n_2 = 30$ and (d)–(f) $n_1 = 30$ and $n_2 = 45$: (a), (d) toy model 1; (b), (e) toy model 2; (c), (f) toy model 3

Table 3. Classification errors and their standard deviations

Method	*Errors (%) (and standard deviations (%)) for the following models:*		
	Toy model 1	*Toy model 2*	*Toy model 3*
Using $\hat{\mathbf{U}}_\rho$	24.9500 (11.3817)	26.8500 (12.6861)	26.7000 (12.5171)
Using $\tilde{\mathbf{U}}_\rho$	25.0000 (11.5470)	26.5000 (12.5831)	26.9500 (12.5508)

Table 4. Classification errors and their standard deviations for the leukaemia and lung cancer data

	Errors (%) (and standard deviations (%)) for the following data:	
	Leukaemia cancer	*Lung cancer*
Without cross-validation	4.4595 (3.0721)	0.9341 (0.8931)
Cross-validation	4.0541 (3.9702)	0.6593 (0.7479)
Estimated ρ	0.2241 (0.2630)	0.0848 (0.0890)

Table 5. Classification errors for shapes and the average values of ρ

Shape pair	*Without cross-validation results*	*Cross-validation results*	*Estimated ρ*
	Error mean (standard deviation)	*Error mean (standard deviation)*	*Mean (standard deviation)*
Apple and bell	7.8 (3.4)	7.4 (3.4)	0.0806 (0.0810)
Pencil and watch	16.0 (7.1)	15.0 (6.5)	0.2561 (0.2757)
Personal car and shoe	6.1 (4.2)	5.5 (3.4)	0.0639 (0.0808)
Camel and elephant	6.9 (4.0)	7.1 (4.1)	0.0723 (0.0763)
Camel and horse	7.7 (5.6)	6.0 (6.1)	0.1196 (0.1829)
Elephant and horse	6.9 (5.9)	6.5 (4.9)	0.0667 (0.0725)

ρ by using fivefold cross-validation. The classification results are summarized in Table 4, where we also reproduce the results in Table 1 for comparison. We also presented therein the average values of ρ chosen by cross-validation along with their standard deviations. We repeat the same simulation with the shape data that we presented in Section 4.3 and present comparisons and the estimated values of ρ in Table 5. As we can see that the choice of ρ is generally different for different types of data, and using cross-validation to select ρ, we can further reduce the classification errors.

Acknowledgements

The authors are grateful to the Joint Editor, the Associate Editor and two referees for helpful

comments, which have led to many improvements. Hao's research is supported by National Science Foundation grant DMS-1309507 and an AMS-Simons travel grant. Fan's research is supported by the National Institute of General Medical Sciences of the National Institutes of Health through grants R01-GM072611-9 and R01-GMR01GM100474 and National Science Foundation grants DMS-1206464 and DMS-1406266.

Appendix A

The parameter ρ is a fixed positive constant throughout the appendix. So for easy presentation, we drop it from the notations $\mathbf{\Sigma}_\rho^{\text{tot}}$, $\mathbf{U}_\rho$, etc.

A.1. Proof of theorem 1

Let $a=\lambda_p>0$ and $a_i=\lambda_i-\lambda_p>0$. It then follows directly from condition 1 and the singular value decomposition that

$$\mathbf{\Sigma}=a\mathbf{I}+\sum_{i=1}^{k} a_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}} \tag{A.1}$$

and

$$\mathbf{\Sigma}^{\text{tot}}=a\mathbf{I}+\rho\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}+\sum_{i=1}^{k} a_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}. \tag{A.2}$$

It can be shown that

$$\left(a\mathbf{I}+\sum_{i=1}^{k} a_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\right)^{-1}=a^{-1}\mathbf{I}-\sum_{i=1}^{k}\frac{a_i}{a(a+a_i)}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}. \tag{A.3}$$

This can be directly verified by

$$\begin{aligned}\left(a\mathbf{I}+\sum_{i=1}^{k} a_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\right)\left\{a^{-1}\mathbf{I}-\sum_{i=1}^{k}\frac{a_i}{a(a+a_i)}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\right\}&=\mathbf{I}+\sum_{i=1}^{k} a^{-1}a_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}-\sum_{i=1}^{k}\frac{a_i}{a+a_i}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}-\sum_{i=1}^{k}\frac{a_i^2}{a(a+a_i)}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\\&=\mathbf{I},\end{aligned}$$

using the orthogonality

$$\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{\xi}_j=\begin{cases}0, & i\neq j,\\ 1, & i=j.\end{cases}$$

By equation (A.3),

$$\boldsymbol{\beta}=\mathbf{\Sigma}^{-1}\boldsymbol{\delta}=a^{-1}\boldsymbol{\delta}-\sum_{i=1}^{k}\frac{a_i\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{\delta}}{a(a+a_i)}\boldsymbol{\xi}_i. \tag{A.4}$$

In other words, $\boldsymbol{\beta}$ is in the space spanned by $\{\boldsymbol{\delta}, \boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k\}$. However, by equation (A.2), it is easy to see that the space that is spanned by eigenvectors of $\mathbf{\Sigma}^{\text{tot}}$ corresponding to eigenvalues greater than a is exactly the space spanned by $\{\boldsymbol{\delta}, \boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k\}$. Therefore, $\boldsymbol{\beta}$ is perpendicular to the $(p-k-1)$-dimensional eigenspace corresponding to eigenvalue a, i.e. $\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}\|_0\leqslant k+1$. □

Before proving theorem 2, we need a couple of results on the eigenvalues and eigenspaces of Hermitian or symmetric matrices.

Lemma 1 (Weyl, 1912). If A and B are symmetric $p\times p$ matrices that differ by a matrix of rank at most r, then their eigenvalues (in descending order) $\{\alpha_j\}_{1\leqslant j\leqslant p}$ and $\{\gamma_j\}_{1\leqslant j\leqslant p}$ satisfy

$$\alpha_{j+r}\leqslant\gamma_j \quad\text{and}\quad \gamma_{j+r}\leqslant\alpha_j \qquad\text{for}\quad 1\leqslant j, j+r\leqslant p.$$

In particular, if $r=1$ and $A\geqslant B$, it implies an interlacing property

$$\alpha_1\geqslant\gamma_1\geqslant\alpha_2\geqslant\ldots\geqslant\alpha_p\geqslant\gamma_p.$$

Lemma 2 (Davis and Kahan, 1970). Let A and B be symmetric matrices with $A-B=H$ and eigenvalues $\{\alpha_j\}_{1\leqslant j\leqslant p}$ and $\{\gamma_j\}_{1\leqslant j\leqslant p}$ respectively. If there is a subset $\mathcal{S}\subset\{1,\ldots,p\}$, an interval $[s,t]$ and a positive constant z, such that $\alpha_j,\gamma_j\in[s,t]$ when $j\in\mathcal{S}$ and $\alpha_j,\gamma_j\in(-\infty,s-z]\cup[t+z,\infty)$ when $j\notin\mathcal{S}$, then $\|P-Q\|\leqslant\|H\|/z$, where P and Q are projection matrices to the subspaces spanned by eigenvectors corresponding to $\{\alpha_j\}_{j\in\mathcal{S}}$ and $\{\gamma_j\}_{j\in\mathcal{S}}$ respectively.

The following lemmas are crucial in the proof of theorem 2.

Lemma 3. Under condition 2, if $\boldsymbol{\delta}\in\mathbf{W}_1$, then the eigenvalues of $\boldsymbol{\Sigma}^{\text{tot}}$ satisfy

$$\eta_1\geqslant\eta_2\geqslant\ldots\geqslant\eta_k\geqslant\eta_{k+1}+d>\eta_{k+1}\geqslant\ldots\geqslant\eta_p; \tag{A.5}$$

otherwise,

$$\eta_1\geqslant\eta_2\geqslant\ldots\geqslant\eta_{k+1}\geqslant\eta_{k+2}+d\frac{\rho\|\boldsymbol{\delta}_2\|_2^2}{d+\rho\|\boldsymbol{\delta}\|_2^2}-\varepsilon\geqslant\eta_{k+2}\geqslant\ldots\geqslant\eta_p. \tag{A.6}$$

Proof. Recall that $\{\lambda_j\}_{1\leqslant j\leqslant p}$ are eigenvalues of $\boldsymbol{\Sigma}$ in descending order and $\boldsymbol{\xi}_j$ is the eigenvector corresponding to λ_j. $\mathbf{W}_1$ and $\mathbf{W}_2$ are linear spaces spanned by $\{\boldsymbol{\xi}_j\}_{1\leqslant j\leqslant k}$ and $\{\boldsymbol{\xi}_j\}_{k+1\leqslant j\leqslant p}$ respectively. $\boldsymbol{\delta}=\boldsymbol{\delta}_1+\boldsymbol{\delta}_2$ with $\boldsymbol{\delta}_m\in\mathbf{W}_m$, $m=1,2$.

If $\boldsymbol{\delta}\in\mathbf{W}_1$, then $\boldsymbol{\delta}\perp\boldsymbol{\xi}_j$ for $k+1\leqslant j\leqslant p$. Therefore, $\{\boldsymbol{\xi}_j\}_{k+1\leqslant j\leqslant p}$ are eigenvectors of $\boldsymbol{\Sigma}^{\text{tot}}=\boldsymbol{\Sigma}+\rho\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}$ as well, and the corresponding eigenvalues satisfy $\eta_j=\lambda_j$ for $k+1\leqslant j\leqslant p$. Moreover, by lemma 1, $\eta_1\geqslant\lambda_1\geqslant\eta_2\geqslant\ldots\geqslant\eta_k\geqslant\lambda_k$. Thus, condition 2 implies that

$$\eta_1\geqslant\eta_2\geqslant\ldots\geqslant\eta_k\geqslant\eta_{k+1}+d>\eta_{k+1}\geqslant\ldots\geqslant\eta_p.$$

If $\boldsymbol{\delta}\notin\mathbf{W}_1$, i.e. $\boldsymbol{\delta}_2\neq0$, define $\mathbf{W}=\mathbf{W}_1\oplus\boldsymbol{\delta}_2$. For all $\mathbf{w}\in\mathbf{W}$, with $\|\mathbf{w}\|_2=1$, we may write $\mathbf{w}=\mathbf{w}_1+\mathbf{w}_2$ where $\mathbf{w}_1\in\mathbf{W}_1$ and $\mathbf{w}_2=c\boldsymbol{\delta}_2\in\mathbf{W}_2$. It follows that

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{\text{tot}}\mathbf{w}&=\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}+\rho\mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}\mathbf{w}\\&=\mathbf{w}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}_1+\mathbf{w}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}_2+\rho\{(\mathbf{w}_1^{\mathrm{T}}+\mathbf{w}_2^{\mathrm{T}})(\boldsymbol{\delta}_1+\boldsymbol{\delta}_2)\}^2\\&=\mathbf{w}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}_1+\mathbf{w}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}_2+\rho(\mathbf{w}_1^{\mathrm{T}}\boldsymbol{\delta}_1+\mathbf{w}_2^{\mathrm{T}}\boldsymbol{\delta}_2)^2\\&\geqslant\lambda_k\|\mathbf{w}_1\|_2^2+\lambda_p\|\mathbf{w}_2\|_2^2+\rho(\mathbf{w}_1^{\mathrm{T}}\boldsymbol{\delta}_1+\mathbf{w}_2^{\mathrm{T}}\boldsymbol{\delta}_2)^2\\&\geqslant\lambda_p+d\|\mathbf{w}_1\|_2^2+\rho(|\mathbf{w}_1^{\mathrm{T}}\boldsymbol{\delta}_1|-|\mathbf{w}_2^{\mathrm{T}}\boldsymbol{\delta}_2|)^2.\end{aligned}$$

It is easy to see that

$$\inf\{\rho(|\mathbf{w}_1^{\mathrm{T}}\boldsymbol{\delta}_1|-|\mathbf{w}_2^{\mathrm{T}}\boldsymbol{\delta}_2|)^2\}=\begin{cases}0, & \text{if } \|\mathbf{w}_1\|_2\geqslant\|\boldsymbol{\delta}_2\|_2/\|\boldsymbol{\delta}\|_2,\\ \rho(\|\mathbf{w}_2\|_2\|\boldsymbol{\delta}_2\|_2-\|\mathbf{w}_1\|_2\|\boldsymbol{\delta}_1\|_2)^2, & \text{if } \|\mathbf{w}_1\|_2<\|\boldsymbol{\delta}_2\|_2/\|\boldsymbol{\delta}\|_2.\end{cases}$$

Therefore, if $\|\mathbf{w}_1\|_2\geqslant\|\boldsymbol{\delta}_2\|_2/\|\boldsymbol{\delta}\|_2$,

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{\text{tot}}\mathbf{w}\geqslant\lambda_p+d\|\mathbf{w}_1\|_2^2\geqslant\lambda_p+d\frac{\|\boldsymbol{\delta}_2\|_2^2}{\|\boldsymbol{\delta}\|_2^2};$$

if $\|\mathbf{w}_1\|_2<\|\boldsymbol{\delta}_2\|_2/\|\boldsymbol{\delta}\|_2$,

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{\text{tot}}\mathbf{w}&\geqslant\lambda_p+d\|\mathbf{w}_1\|_2^2+\rho(\|\mathbf{w}_2\|_2\|\boldsymbol{\delta}_2\|_2-\|\mathbf{w}_1\|_2\|\boldsymbol{\delta}_1\|_2)^2\\&\geqslant\lambda_p+d\|\mathbf{w}_1\|_2^2+\rho(\|\boldsymbol{\delta}_2\|_2-\|\mathbf{w}_1\|_2\|\boldsymbol{\delta}\|_2)^2\\&\geqslant\lambda_p+d\frac{\rho\|\boldsymbol{\delta}_2\|_2^2}{d+\rho\|\boldsymbol{\delta}\|_2^2}.\end{aligned}$$

Overall, we have

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{\text{tot}}\mathbf{w}\geqslant\lambda_p+\tilde{d}\qquad\text{for all } \mathbf{w}\in\mathbf{W}, \tag{A.7}$$

where $\tilde{d}=d\rho\|\boldsymbol{\delta}_2\|_2^2/(d+\rho\|\boldsymbol{\delta}\|_2^2)$. Since $\dim(\mathbf{W})=k+1$, result (A.7) implies that there are $k+1$ eigenvalues that are greater than $\lambda_p+\tilde{d}$ for $\boldsymbol{\Sigma}^{\text{tot}}$. Together with lemma 1, we conclude that

$$\eta_1\geqslant\eta_2\geqslant\ldots\geqslant\eta_k\geqslant\eta_{k+1}\geqslant\lambda_p+\tilde{d}>\lambda_{k+1}\geqslant\eta_{k+2}\geqslant\ldots\geqslant\eta_p.$$

which leads to result (A.6). □

Similarly, we have the following lemma.

Lemma 4. Under condition 1, if $\boldsymbol{\delta} \in \mathbf{W}_1$, then the eigenvalues of $\boldsymbol{\Sigma}^{\text{tot}}$ satisfy

$$\eta_1 \geqslant \eta_2 \geqslant \ldots \geqslant \eta_k \geqslant \eta_{k+1} + d > \eta_{k+1} = \ldots = \eta_p; \tag{A.8}$$

otherwise,

$$\eta_1 \geqslant \eta_2 \geqslant \ldots \geqslant \eta_k \geqslant \eta_{k+1} \geqslant \eta_{k+2} + d\frac{\rho\|\boldsymbol{\delta}_2\|_2^2}{d+\rho\|\boldsymbol{\delta}\|_2^2} \geqslant \eta_{k+2} = \ldots = \eta_p. \tag{A.9}$$

Proof. Note that condition 1 is a special case of condition 2 with $\varepsilon = 0$. Therefore expressions (A.8) and (A.9) are implied by expressions (A.5) and (A.6) respectively, except the equalities in the second halves of the sequences, which are implied by lemma 1 and the fact that $\lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p$.

A.2. Proof of theorem 2

Again, let $\boldsymbol{\xi}_j$ be the eigenvector of $\boldsymbol{\Sigma}$ corresponding to λ_j for $1 \leqslant j \leqslant p$. $a = \lambda_p$ and $a_j = \lambda_j - \lambda_p$.

A.2.1. Part I

$\boldsymbol{\delta} \in \mathbf{W}_1$ implies that $\boldsymbol{\delta} \perp \boldsymbol{\xi}_j$ for $k < j \leqslant p$, so the eigenvectors $\{\boldsymbol{\xi}_j\}_{k<j\leqslant p}$ are also eigenvectors of $\boldsymbol{\Sigma}^{\text{tot}}$. Write $\mathbf{U} = (\mathbf{U}_1\ \mathbf{U}_2)$ where $\mathbf{U}_2$ is a submatrix of $\mathbf{U}$, consisting of right $p-k$ columns. Then $\mathbf{U}_2 = (\boldsymbol{\xi}_{k+1}, \ldots, \boldsymbol{\xi}_p)$. Therefore,

$$\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \mathbf{D}_2^{-1}\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0},$$

where $\mathbf{D}_2 = \operatorname{diag}(\lambda_{k+1}, \ldots, \lambda_p)$.

A.2.2. Part II

Under condition 2, we can write $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Sigma}_0 = a\mathbf{I} + \Sigma_{j=1}^{k} a_j\boldsymbol{\xi}_j\boldsymbol{\xi}_j^{\mathrm{T}}$ and $\boldsymbol{\Delta} = \Sigma_{j=k+1}^{p} a_j\boldsymbol{\xi}_j\boldsymbol{\xi}_j^{\mathrm{T}}$. Thus, $\boldsymbol{\Sigma}_0$ satisfies condition 1, and $\boldsymbol{\Delta}$ is a semipositive matrix with maximal eigenvalue less than ε. Define

$$\boldsymbol{\Sigma}_0^{\text{tot}} = \boldsymbol{\Sigma}_0 + \rho\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}},$$
$$\boldsymbol{\Sigma}^{\text{tot}} = \boldsymbol{\Sigma} + \rho\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}.$$

And let $\{\eta_{0j}\}_{1\leqslant j\leqslant p}$ and $\{\eta_j\}_{1\leqslant j\leqslant p}$ be their eigenvalues, in descending order respectively. Moreover, let $\mathbf{V}$ and $\mathbf{U}$ be orthogonal matrices such that

$$\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{\text{tot}}\mathbf{V} = \mathbf{D}_0,$$
$$\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}^{\text{tot}}\mathbf{U} = \mathbf{D},$$

where $\mathbf{D}_0 = \operatorname{diag}(\eta_{01}, \ldots, \eta_{0p})$ and $\mathbf{D} = \operatorname{diag}(\eta_1, \ldots, \eta_p)$.

Here is the strategy of the proof. By theorem 1, $\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}$ is sparse so its l_1-norm can be well controlled. Because of the results on the separated eigenvalues (lemmas 3 and 4), we can show that $\mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}$ is similar to $\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}$ by using lemma 2. Therefore, the l_1-norm can be controlled as well.

Write $\mathbf{U} = (\mathbf{U}_1\ \mathbf{U}_2)$ and $\mathbf{V} = (\mathbf{V}_1\ \mathbf{V}_2)$ where $\mathbf{U}_2$ and $\mathbf{V}_2$ are submatrices of $\mathbf{U}$ and $\mathbf{V}$ respectively, consisting of right $p-k-1$ columns. Note that

$$\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}\|_1 = \|\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\beta}\|_1 + \|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\beta}\|_1,$$

where $\|\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\beta}\|_1 \leqslant \sqrt{(\|\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\beta}\|_0\|\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\beta}\|_2^2)} \leqslant \sqrt{(k+1)}\|\boldsymbol{\beta}\|_2$. So it is crucial to control $\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\beta}\|_1$. From the proof of theorem 1, we see that $\mathbf{V}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta} = \mathbf{0}$. Hence

$$\begin{aligned}
\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\beta}\|_2 &= \|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} - \mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta} + \mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2 \\
&\leqslant \|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} - \mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2 + \|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2 \\
&\leqslant \|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} - \mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2 + \sqrt{(\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2^2 - \|\mathbf{V}_2^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2^2)} \\
&\leqslant \|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}\|\|\boldsymbol{\delta}_2\|_2 + \sqrt{\{\boldsymbol{\delta}^{\mathrm{T}}(\boldsymbol{\Sigma}_0^{-1})^{\mathrm{T}}(\mathbf{U}_2\mathbf{U}_2^{\mathrm{T}} - \mathbf{V}_2\mathbf{V}_2^{\mathrm{T}})\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\}} \\
&= S_1 + S_2
\end{aligned}$$

and

$$\|\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_0^{-1}\|=\lambda_p^{-1}-\lambda_{k+1}^{-1}=a^{-1}-(a+a_{k+1})^{-1}=\frac{a_{k+1}}{a(a+a_{k+1})}\leqslant\frac{\varepsilon}{a^2}.$$

Thus,

$$S_1\leqslant\frac{\varepsilon}{a^2}\|\boldsymbol{\delta}_2\|_2\leqslant\frac{\varepsilon(a+\varepsilon)}{a^2}\|\boldsymbol{\beta}\|_2.$$

To control S_2, we must show that the spaces that are spanned by column vectors of $\mathbf{V}_2$ and $\mathbf{U}_2$ are close to each other. By lemmas 3 and 4, we have

$$\eta_1\geqslant\eta_2\geqslant\ldots\geqslant\eta_k\geqslant\eta_{k+1}\geqslant\eta_{k+2}+\tilde{d}-\varepsilon\geqslant\eta_{k+2}\geqslant\ldots\geqslant\eta_p,$$

$$\eta_{01}\geqslant\eta_{02}\geqslant\ldots\geqslant\eta_{0k}\geqslant\eta_{0,k+1}\geqslant\eta_{0,k+2}+\tilde{d}\geqslant\eta_{0,k+2}=\ldots=\eta_{0p},$$

where $\tilde{d}=d\rho\|\boldsymbol{\delta}_2\|_2^2/(d+\rho\|\boldsymbol{\delta}\|_2^2)$. Moreover, by lemma 1, $\eta_{k+2}\leqslant\lambda_{k+1}\leqslant\lambda_p+\varepsilon=a+\varepsilon$, $\eta_{0,k+2}=\lambda_p=a$. However, $\eta_{k+1}\geqslant\eta_{k+2}+\tilde{d}-\varepsilon\geqslant a+\tilde{d}-\varepsilon$ and $\eta_{0,k+1}\geqslant\eta_{0,k+2}+\tilde{d}=a+\tilde{d}$.

By lemma 2, $\|\mathbf{U}_2\mathbf{U}_2^{\mathrm{T}}-\mathbf{V}_2\mathbf{V}_2^{\mathrm{T}}\|\leqslant\|\boldsymbol{\Delta}\|/(\tilde{d}-2\varepsilon)\leqslant\varepsilon/(\tilde{d}-2\varepsilon)$.

$$\|\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}\|_2=\|\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{\beta}\|_2\leqslant\|\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\|\,\|\boldsymbol{\beta}\|_2\leqslant\frac{a+\varepsilon}{a}\|\boldsymbol{\beta}\|_2.$$

Thus,

$$S_2\leqslant\sqrt{\left(\frac{\varepsilon}{\tilde{d}-2\varepsilon}\right)}\frac{a+\varepsilon}{a}\|\boldsymbol{\beta}\|_2.$$

Therefore,

$$\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\beta}\|_2\leqslant\frac{a+\varepsilon}{a}\left\{\frac{\varepsilon}{a}+\sqrt{\left(\frac{\varepsilon}{\tilde{d}-2\varepsilon}\right)}\right\}\|\boldsymbol{\beta}\|_2.$$

$$\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\beta}\|_1\leqslant\sqrt{(p-k-1)}\frac{a+\varepsilon}{a}\left\{\frac{\varepsilon}{a}+\sqrt{\left(\frac{\varepsilon}{\tilde{d}-2\varepsilon}\right)}\right\}\|\boldsymbol{\beta}\|_2.$$

Finally,

$$\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2\leqslant\sqrt{(k+1)}+\sqrt{(p-k-1)}\frac{a+\varepsilon}{a}\left\{\frac{\varepsilon}{a}+\sqrt{\left(\frac{\varepsilon}{\tilde{d}-2\varepsilon}\right)}\right\}.$$

A.3. Proof of theorem 3

Let $\mathbf{V}_1$ be a matrix whose columns vectors are the eigenvectors corresponding to the non-vanishing eigenvalues of the matrix $\mathbf{A}=\Sigma_{i=1}^k\lambda_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}+\rho\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}$. Recall that $\lambda_i(\mathbf{B})$ is the ith largest eigenvalue of a symmetric matrix $\mathbf{B}$. Then, by lemma 2,

$$|\mathbf{U}_1-\mathbf{V}_1|\leqslant\frac{\|\boldsymbol{\Sigma}^{\mathrm{tot}}-\mathbf{A}\|}{\lambda_{k+1}(\mathbf{A})-\lambda_{k+2}(\boldsymbol{\Sigma}^{\mathrm{tot}})}=\frac{\lambda_{k+1}}{\lambda_{k+1}(\mathbf{A})-\lambda_{k+2}(\boldsymbol{\Sigma}^{\mathrm{tot}})}.$$

By lemma 1, $\lambda_{k+2}(\boldsymbol{\Sigma}^{\mathrm{tot}})\leqslant\lambda_{k+1}$. Hence,

$$\|\mathbf{U}_1-\mathbf{V}_1\|\leqslant\frac{\lambda_{k+1}}{\lambda_{k+1}(\mathbf{A})-\lambda_{k+1}}\leqslant\frac{1}{a-1},$$

Let $\mathbf{V}_2$ be the eigenvectors that are orthogonal to $\mathbf{V}_1$. Then, $\mathbf{V}_2^{\mathrm{T}}\boldsymbol{\delta}=0$, since the columns of $\mathbf{V}_1$ are linear combinations of $\boldsymbol{\delta}$ and $\{\boldsymbol{\xi}_i\}_{i=1}^k$. Consequently, $\|\mathbf{V}_1^{\mathrm{T}}\boldsymbol{\delta}\|_2=\|\boldsymbol{\delta}\|_2$ and

$$\|\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\delta}\|_2=\|\mathbf{V}_1^{\mathrm{T}}\boldsymbol{\delta}+(\mathbf{U}_1-\mathbf{V}_1)^{\mathrm{T}}\boldsymbol{\delta}\|_2\geqslant\|\boldsymbol{\delta}\|_2-\|\mathbf{U}_1-\mathbf{V}_1\|\|\boldsymbol{\delta}\|_2=\frac{a-2}{a-1}\|\boldsymbol{\delta}\|_2.$$

The second conclusion follows directly from the fact that $\|\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}_1\|\leqslant\|\boldsymbol{\Sigma}\|=\lambda_1$.

References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012) Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *Ann. Statist.*, **40**, 1171–1197.

Bickel, P. and Levina, E. (2004) Some theory for Fisher's linear discriminant function, naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.

Cai, T. and Liu, W. (2011) A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Ass.*, **106**, 1566–1577.

Davis, C. and Kahan, W. (1970) The rotation of eigenvectors by a perturbation: iii. *SIAM J. Numer. Anal.*, **7**, 1–46.

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

Fan, J., Feng, Y. and Tong, X. (2012) A road to classification in high dimensional space. *J. R. Statist. Soc.* B, **74**, 745–771.

Fan, J., Liao, Y. and Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Statist. Soc.* B, **75**, 603–680.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D. and Bueno, R. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, **62**, 49–63.

Hall, P., Jin, J. and Miller, H. (2014) Feature selection when there are many influential features. *Bernoulli*, **20**, 1647–1671.

Johnstone, I. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**, 295–327.

Johnstone, I. and Lu, A. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Ass.*, **104**, 682–693.

Karoui, N. (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, **36**, 2717–2756.

Mai, Q., Zou, H. and Yuan, M. (2012) A direct approach to sparse discriminant analysis in ultra-high dimensionals. *Biometrika*, **99**, 29–42.

Shen, D., Shen, H., Zhu, H. and Marron, J. (2013) Surprising asymptotic conical structure in critical sample eigen-directions. *Preprint arXiv:1303.6171.*

Thakoor, N., Gao, J. and Jung, S. (2007) Hidden markov model-based weighted likelihood discriminant for 2-d shape classification. *IEEE Trans. Image Process.*, **16**, 2707–2719.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA*, **99**, 6567–6572.

Trendafilov, N. T. and Jolliffe, I. T. (2007) Dalass: variable selection in discriminant analysis via the lasso. *Computnl Statist. Data Anal.*, **51**, 3718–3736.

Weyl, H. (1912) Das asymtotische Verteilungsgesetz der eigenwerte lineare partieller Differentialgleichungen. *Math. Ann.*, **71**, 441–479.

Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C. and Lin, X. (2009) Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, **25**, 1145–1151.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Computnl Graph. Statist.*, **15**, 265–286.