A ROAD to Classification in High Dimensional Space

Jianqing Fan[1], Yang Feng[2] and Xin Tong[1]

[1] *Department of Operations Research & Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.*

[2] *Department of Statistics, Columbia University, New York, NY 10027, U.S.A.*

Summary. For high-dimensional classification, it is well known that naively performing the Fisher discriminant rule leads to poor results due to diverging spectra and noise accumulation. Therefore, researchers proposed independence rules to circumvent the diverging spectra, and sparse independence rules to mitigate the issue of noise accumulation. However, in biological applications, there are often a group of correlated genes responsible for clinical outcomes, and the use of the covariance information can significantly reduce misclassification rates. In theory the extent of such error rate reductions is unveiled by comparing the misclassification rates of the Fisher discriminant rule and the independence rule. To materialize the gain based on finite samples, a Regularized Optimal Affine Discriminant (ROAD) is proposed. ROAD selects an increasing number of features as the regularization relaxes. Further benefits can be achieved when a screening method is employed to narrow the feature pool before hitting the ROAD. An efficient Constrained Coordinate Descent algorithm (CCD) is also developed to solve the associated optimization problems. Sampling properties of oracle type are established. Simulation studies and real data analysis support our theoretical results and demonstrate the advantages of the new classification procedure under a variety of correlation structures. A delicate result on continuous piecewise linear solution path for the ROAD optimization problem at the population level justifies the linear interpolation of the CCD algorithm.

Keywords: High Dimensional Classification, LDA, Regularized Optimal Affine Discriminant, Fisher Discriminant, Independence Rule.

1. Introduction

Technological innovations have had deep impact on society and on various areas of scientific research. High-throughput data from microarray and proteomics technologies are frequently used in many contemporary statistical studies. In the case of microarray data, the dimensionality is frequently in thousands or beyond, while the sample size is typically in the order of tens. The large-p-small-n scenario poses challenges for the classification problems. We refer to Fan and Lv (2010) for an overview of statistical challenges associated with high dimensionality.

When the feature space dimension p is very high compared to the sample size n, the Fisher discriminant rule performs poorly due to diverging spectra as demonstrated by Bickel and Levina (2004). These authors showed that the independence rule in which the covariance structure is ignored performs better than the naive Fisher rule (NFR) in the high dimensional setting. Fan and Fan (2008) demonstrated further that even for the independence rules, a procedure using all the features can be as poor as random guessing due to noise accumulation in estimating population centroids in high-dimensional feature space. As a result, Fan and Fan (2008) proposed the Features Annealed Independence Rule (FAIR) that selects a subset of important features for classification. Dudoit *et al.* (2002) reported that for microarray data, ignoring correlations between genes leads to better classification results. Tibshirani *et al.* (2002) proposed

the Nearest Shrunken Centroid (NSC) which likewise employs the working independence structure. Similar problems are also studied in the machine learning community such as Domingos and Pazzani (1997) and Lewis (1998).

In microarray studies, correlation among different genes is an essential characteristic of the data and usually not negligible. Other examples include proteomics, and metabolomics data where correlation among biomarkers is commonplace. More details can be found in Ackermann and Strimmer (2009). Intuitively, the independence assumption among genes leads to loss of critical information and hence is suboptimal. We believe that in many cases, the crucial point is not whether to consider correlations, but how we can incorporate the covariance structure into the analysis with a bullet proof vest against diverging spectra and significant noise accumulation effect.

The setup of the objective classification problem is now introduced. We assume in the following that the variability of data under consideration can be described reasonably well by the means and variances. To be more precise, suppose that random variables representing two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ follow p-variate normal distributions: $\mathbf{X}|Y = 1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{X}|Y = 2 \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ respectively. Moreover, assume $\mathbb{P}(Y = 1) = 1/2$. This Gaussian discriminant analysis setup is known for its good performance despite its rigid model structure. For any linear discriminant rule

$$\delta_{\mathbf{w}}(\mathbf{X}) = \mathbb{I}\{\mathbf{w}^T(\mathbf{X} - \boldsymbol{\mu}_a) > 0\}, \tag{1}$$

where $\boldsymbol{\mu}_a = (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2$, and $\mathbb{I}$ denotes the indicator function with value 1 corresponds to assigning $\mathbf{X}$ to class $\mathcal{C}_2$ and 0 class $\mathcal{C}_1$, the misclassification rate of the (pseudo) classifier $\delta_{\mathbf{w}}$ is

$$W(\delta_{\mathbf{w}}) = \frac{1}{2}P_2(\delta_{\mathbf{w}}(\mathbf{X}) = 0) + \frac{1}{2}P_1(\delta_{\mathbf{w}}(\mathbf{X}) = 1) = 1 - \Phi(\mathbf{w}^T\boldsymbol{\mu}_d/(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{1/2}), \tag{2}$$

where $\boldsymbol{\mu}_d = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/2$, and P_i is the conditional distribution of $\mathbf{X}$ given its class label i. We will focus on such linear classifier $\delta_{\mathbf{w}}(\cdot)$, and the mission is to find a good data projection direction $\mathbf{w}$. Note that the Fisher discriminant

$$\delta_F(\mathbf{X}) = \mathbb{I}\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d)^T(\mathbf{X} - \boldsymbol{\mu}_a) > 0\} \tag{3}$$

is the *Bayes rule*. There are two fundamental difficulties in applying the Fisher discriminant whose missclassification rate is

$$1 - \Phi\left((\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d)^{1/2}\right). \tag{4}$$

The first difficulty arises from the noise accumulation effect in estimating the population centroids (Fan and Fan, 2008) when p is large. The second challenge is more severe: estimating the inverse of covariance matrix $\boldsymbol{\Sigma}$ when $p > n$ (Bickel and Levina, 2004). As a result, much previous researches focus on the independence rules, which act as if $\boldsymbol{\Sigma}$ is diagonal. However, correlation matters!

To illustrate this point, consider a case when $p = 2$. These two features can be selected from the original thousands of features, and we can estimate the correlation between two variables with reasonable accuracy. Let

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $\rho \in [0, 1)$ and $\boldsymbol{\mu}_d = (\mu_1, \mu_2)^T$. Without loss of generality, assume $|\mu_1| \geq |\mu_2| > 0$. The misclassification rate of Fisher discriminant depends on

$$\Delta_p(\rho) = \boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d = \frac{1}{1 - \rho^2}(\mu_1^2 + \mu_2^2 - 2\rho\mu_1\mu_2). \tag{5}$$

Note that

$$\Delta_p'(\rho) > 0 \Leftrightarrow \mu_1\mu_2\rho^2 - (\mu_1^2 + \mu_2^2)\rho + \mu_1\mu_2 < 0.$$

Therefore, when $\mu_1\mu_2 < 0$, $\Delta_p'(\rho) > 0$ for all $\rho \in [0, 1)$. On the other hand, when $\mu_1\mu_2 > 0$, $\Delta_p(\rho)$ decreases on $\rho \in (0, \frac{\mu_2}{\mu_1})$, and increases on $(\frac{\mu_2}{\mu_1}, 1)$. Notice that when $\rho \to 1$, $\Delta_p \to \infty$ regardless of signs for $\mu_1\mu_2$, which in turn leads to vanishing classification error. On the other hand, if we use independence rule (also called naive Bayes rule), the optimal misclassification rate

$$1 - \Phi\left(\frac{\|\boldsymbol{\mu}_d\|_2^2}{(\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}\boldsymbol{\mu}_d)^{1/2}}\right) \tag{6}$$

depends on $\Gamma(\rho) = \|\boldsymbol{\mu}_d\|_2^4/\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}\boldsymbol{\mu}_d$, which is monotonically decreasing for $\rho \in [0, 1)$, with the limit $(\mu_1^2 + \mu_2^2)^2/(\mu_1 + \mu_2)^4$ that is smaller than unity when μ_1 and μ_2 have the same sign. Hence, the optimal classification error using the independence rule actually increases as correlation among features increases.

The above simple example shows that by incorporating correlation information, the gain in terms of classification error can be substantial. Elaboration on this point in more realistic scenarios is provided in Section 2. Now it seems wise to use at least a part of covariance structure to improve the performance of a classifier. So there is a need to estimate the covariance matrix $\boldsymbol{\Sigma}$. Without structural assumptions on $\boldsymbol{\Sigma}$, the pooled sample covariance $\hat{\boldsymbol{\Sigma}}$ is one natural estimate. But for $p > n$, it is not considered as a good estimate of $\boldsymbol{\Sigma}$ in general. We are lucky here because our mission is not constructing a good estimate of the covariance matrix, but finding a good direction $\mathbf{w}$ that leads to a good classifier. To mimic the optimal data projection direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$, we do not adopt a direct plug-in approach, simply because it is unlikely that a product is a good estimate when at least one of its components is not. Instead, we find the data projection direction $\mathbf{w}$ by directly minimizing the classification error subject to a capacity constraint on $\mathbf{w}$. From a broad spectrum of simulated and real data analysis, we are convinced that this approach leads to a robust and efficient sparse linear classifier.

Admittedly, our work is far from the first to use covariance for classification; support vector machines (Vapnik, 1995), for example, implicitly utilize covariance between covariates. Another notable work is "shrunken centroids regularized discriminant analysis" (SCRDA) (Guo *et al.*, 2005), which calls for a version of regularized sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\text{reg}}$, and soft-thresholds on $\hat{\boldsymbol{\Sigma}}_{\text{reg}}^{-1}\hat{\mathbf{x}}_i$. Shao *et al.* (2011) consider a sparse linear discriminant analysis, assuming the sparsity on both the covariance matrix and the mean difference vector so that they can be regularized. They show that such a regularized estimator is asymptotically optimal under some conditions. However, to the best of our knowledge, this work is the first to select features by directly optimizing the misclassification rates, to explicitly use un-regularized sample covariance information, and to establish the oracle inequality and risk approximation theory.

There is a huge literature on high dimensional classification. Examples include principal component analysis in Bair *et al.* (2006) and Zou *et al.* (2006), partial least squares in Nguyen and Rocke (2002), Huang (2003) and Boulesteix (2004), and sliced inverse regression in Li (1991) and Antoniadis *et al.* (2003).

The rest of our paper is organized as follows. Section 2 provides some insights on the performances of naive Bayes, Fisher discriminant and restricted Fisher discriminants. In Section 3, we propose the Regularized Optimal Affine Discriminant (ROAD) and variants of ROAD. An efficient algorithm Constrained Coordinate Descent (CCD) is constructed in Section 4. Main risk approximation results and continuous piecewise linear property of the solution path are established in Section 5. We conduct simulation and empirical studies in Section 6. A discussion is given in Section 7, and all proofs are relegated to the appendix.

2. Naive Bayes and Fisher Discriminant

To compare the naive Bayes and Fisher discriminant at the population level, we assume without loss of generality that variables have been marginally standardized so that $\mathbf{\Sigma}$ is a correlation matrix. Recall that the naive Bayes discriminant has error rate (6) and the Fisher discriminant has error rate (4). Let $\Gamma_p = \|\boldsymbol{\mu}_d\|_2^4/\boldsymbol{\mu}_d^T\mathbf{\Sigma}\boldsymbol{\mu}_d$ and $\Delta_p = \boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$. Denote by $\{\lambda_i\}_{i=1}^p$ the eigenvalues and $\{\boldsymbol{\xi}_i\}_{i=1}^p$ eigenvectors of the matrix $\mathbf{\Sigma}$. Decompose

$$\boldsymbol{\mu}_d = a_1\boldsymbol{\xi}_1 + \cdots + a_p\boldsymbol{\xi}_p, \tag{7}$$

where $\{a_i\}_{i=1}^p$ are the coefficients of $\boldsymbol{\mu}_d$ in this new orthonormal basis $\{\boldsymbol{\xi}_i\}_{i=1}^p$. Using the decomposition (7), we have

$$\Delta_p = \sum_{j=1}^p a_j^2/\lambda_j, \qquad \Gamma_p = \Big(\sum_{j=1}^p a_j^2\Big)^2 / \sum_{j=1}^p \lambda_j a_j^2. \tag{8}$$

The relative efficiency of Fisher discriminant over naive Bayes is characterized by Δ_p/Γ_p. By the Cauchy-Schwartz inequality,

$$\Delta_p/\Gamma_p \geq 1.$$

The naive Bayes method performs as well as the Fisher discriminant only when $\boldsymbol{\mu}_d$ is an eigenvector of $\mathbf{\Sigma}$.

In general, Δ_p/Γ_p can be much larger than unity. Since $\mathbf{\Sigma}$ is the correlation matrix, $\sum_{j=1}^p \lambda_j = \mathrm{tr}(\mathbf{\Sigma}) = p$. If $\boldsymbol{\mu}_d$ is equally loaded on $\boldsymbol{\xi}_j$, then the ratio

$$\Delta_p/\Gamma_p = p^{-2}\sum_{j=1}^p \lambda_j \sum_{j=1}^p \lambda_j^{-1} = p^{-1}\sum_{j=1}^p \lambda_j^{-1}. \tag{9}$$

More generally, if $\{a_j\}_{j=1}^p$ are realizations from a distribution with the second moment σ^2, then by the law of large numbers,

$$\sum_{j=1}^p a_j^2\lambda_j^{-1} \approx \sigma^2\sum_{j=1}^p 1/\lambda_j, \quad p^{-1}\sum_{j=1}^p a_j^2 \approx \sigma^2, \quad \sum_{j=1}^p \lambda_j a_j^2 \approx \sigma^2\sum_{j=1}^p \lambda_j.$$

Hence, (9) holds approximately in this case. In other words, the right hand side of (9) is approximately the relative efficiency of the Fisher discriminant over the naive Bayes. Now suppose further that half of the eigenvalues of $\mathbf{\Sigma}$ are c and the other half are $2-c$. Then, the right hand side of (9) is $(c^{-1}+(2-c)^{-1})/2$. For example when the condition number is 10, this ratio is about 3. A high ratio translates into a large difference in error rates: $1-\Phi(\Gamma_p^{1/2})$ for independence rule is much larger than $1-\Phi(3\Gamma_p^{1/2})$ for Fisher discriminant. For example, when $\Gamma_p^{1/2} = 0.5$, we have 30.9% and 6.7% error rates respectively for the naive Bayes and Fisher discriminant.

To put the above arguments under a visual inspection, consider a case in which $p = 1000$, $\boldsymbol{\mu}_d = (\boldsymbol{\mu}_s^T, \mathbf{0}^T)^T$ with $\boldsymbol{\mu}_s = (1,1,1,1,1,2,2,2,2,2)^T$ and $\mathbf{\Sigma}$ equals the equi-correlation matrix with pairwise correlation ρ. The vector $\boldsymbol{\mu}_d$ simulates the case in which 10 genes out of 1000 express mean differences. Figure 1 depicts the theoretical error rates of the Fisher discriminant and the naive Bayes rule as functions of ρ.

It is not surprising that the Fisher discriminant rule performs significantly better than the naive Bayes as ρ deviates away from zero. The error rate of the naive Bayes actually increases with ρ, whereas the error rate of the Fisher discriminant tends to zero as ρ approaches 1. This phenomenon is the same as what was shown analytically through the toy example in Section 1. To mimic Fisher discriminant by a plug-in estimator, we need to estimate $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$ with reasonable accuracy. This mission is difficult if not

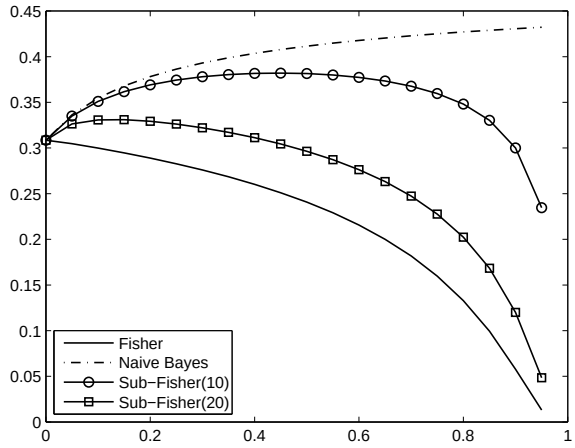


Fig. 1. Misclassification rates of Fisher discriminant, naive Bayes and restricted Fisher rules (10 and 20 features, respectively) against ρ.

impossible. On the other hand, imitating a weaker oracle is more manageable. For example, when the samples are of reasonable size, we can select the 10 variables with differences in means by applying a two-sample t-test. Restricting to the best linear classifiers based on these $s = 10$ variables, we have the optimal error rate

$$1 - \Phi((\boldsymbol{\mu}_s^T \boldsymbol{\Sigma}_s^{-1} \boldsymbol{\mu}_s)^{1/2}),$$

where the classification rule is $\delta_{\mathbf{w}^R}$ and $\mathbf{w}^R = ((\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\mu}_s)^T, \mathbf{0}^T)^T$. The performance of this oracle classifier is depicted by the sub-Fisher (10 features) in Figure 1. It performs much better than the naive Bayes method. One can also employ the naive Bayes rule to the restricted feature space, but this method has exactly the same performance as the naive Bayes method in the whole space. Thus, the restricted Fisher discriminant outperforms both the naive Bayes method with restricted features and the naive Bayes method using all features.

Mimicking the performance of the restricted Fisher discriminant is feasible. Instead of estimating a 1000×1000 covariance matrix, we only need to gauge a 10×10 submatrix. However, this restricted Fisher rule is not powerful enough, as shown in Figure 1. We can improve its performance by including 10 most correlated variables to each of those selected features to further account for the correlation effect, giving rise to a 20-dimensional feature space. Since the variables are equally correlated in this example, we are free to choose any 10 variables among the other 990. The performance of such an enlarged restricted Fisher discriminant is represented by sub-Fisher (20 features) in Figure 1. It performs closely to the Fisher discriminant which uses the whole feature space, and it is feasible to implement with finite samples.

3. Regularized Optimal Affine Discriminant

The misclassification rate of Fisher discriminant is $1 - \Phi(\Delta_p^{1/2})$, where $\Delta_p = \boldsymbol{\mu}_d^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$. However, for high dimensional data, it is impossible to achieve such a performance empirically. Among other reasons, the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ is ill-conditioned or not invertible. One solution is to focus only on the $s(<< p)$ most important features for classification. Ideally, the best s features should be the ones with the largest Δ_s among all $\binom{p}{s}$ possibilities, where Δ_s is the counterpart of Δ_p when only s variables

are considered. Naive search for the best subset of size s is NP-hard. Thus, we develop a regularized method to circumvent these two problems.

3.1. ROAD

Recall that by (2), minimizing the classification error $W(\delta_{\mathbf{w}})$ is the same as maximizing $\mathbf{w}^T\boldsymbol{\mu}_d/(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{1/2}$, which is equivalent to minimizing $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ subject to $\mathbf{w}^T\boldsymbol{\mu}_d = 1$. We would like to add a penalty function for capacity control. There are many ways to do regularization; for the literature on penalized methods, refer to LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Elastic net (Zou and Hastie, 2005), MCP (Zhang, 2010) and related methods (Zou, 2006; Zou and Li, 2008). As our primary interest is classification error (the risk of the procedure), an L_1 constraint $\|\mathbf{w}\|_1 \le c$ is added for regularization, so the problem can be recast as

$$\mathbf{w}_c = \underset{\|\mathbf{w}\|_1 \le c, \mathbf{w}^T\boldsymbol{\mu}_d = 1}{\operatorname{argmin}} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}. \tag{10}$$

We name the classifier $\delta_{\mathbf{w}_c}(\cdot)$ the Regularized Optimal Affine Discriminant(ROAD). The existence of a feasible solution in (10) dictates

$$c \ge 1/\max_{1\le i\le p} |\mu_{d,i}|. \tag{11}$$

When c is small, we obtain a sparse solution and achieve feature selection using covariance information. When $c \ge \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d\|_1/\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$, the L_1 constraint is no longer binding and $\delta_{\mathbf{w}_c}$ reduces to the Fisher discriminant, which can be denoted by $\delta_{\mathbf{w}_\infty}$ $(= \delta_F)$. Therefore we have provided a family of linear discriminants, indexed by c, using from only one feature to all features. In some applications such as portfolio selection, the choice of c reflects the investor's tolerance upper bound on gross exposure. In other applications, when the user does not have a such a preference, the choice of c can be data-driven. To accommodate both application scenarios, we propose a coordinate descent algorithm (Section 4) to implement our ROAD proposal.

3.2. Variants of ROAD

At the sample level, NSC (Tibshirani *et al.*, 2002) and FAIR (Fan and Fan, 2008) both use shrunken versions of standardized mean difference to find the s features. In the same spirit, we consider the following Diagonal Regularized Optimal Affine Discriminant(D-ROAD) $\delta_{\mathbf{w}_c^I}$, where

$$\mathbf{w}_c^I = \underset{\|\mathbf{w}\|_1 \le c, \mathbf{w}^T\boldsymbol{\mu}_d = 1}{\operatorname{argmin}} \mathbf{w}^T\operatorname{diag}(\boldsymbol{\Sigma})\mathbf{w}. \tag{12}$$

The D-ROAD will be compared with NSC (Tibshirani *et al.*, 2002) and FAIR (Fan and Fan, 2008) in the simulation studies, and all these independence based rules will be compared with ROAD and its two variants defined below.

A screening-based variant (to be proposed) of ROAD aims at mimicking the performance of sub-Fisher (10 features) in Figure 1. A fast way to select features is the independence screening, which uses the marginal information such as the two-sample t-test. We can also enlarge the selected feature subspace by incorporating the features which are most correlated to what have been chosen. This additional variant of ROAD tracks the performance of sub-Fisher (20 features) in Figure 1. We will refer to the two variants of ROAD as S-ROAD1 and S-ROAD2. More description of these procedures, along with their theoretical properties and numerical investigations, will be detailed in Sections 5 and 6.

A hint of the rationale behind including correlated features that do not show a difference in means between the two classes, is revealed through the two-feature example in the introduction. Suppose $\mu_2 = 0$.

Then, by (5), the power of the discriminant using two features is $1-\Phi(\Delta_2^{1/2})$ where $\Delta_2=\mu_1^2/(1-\rho^2)$, whereas with the first feature alone the misclassification rate is $1-\Phi(\Delta_1^{1/2})$ where $\Delta_1=\mu_1^2$. Therefore when the correlation $|\rho|$ is large, using two correlated features is far more powerful than employing only one feature, even though the second feature has no marginal discrimination power. More intuition is granted by this observation: at the population level, the best s features are not necessarily those with largest standardized mean differences. In other words, with the two class Gaussian model in mind, when $\mathbf{\Sigma}$ is the correlation matrix, the most powerful s features for classification are not necessarily the coordinates of $\boldsymbol{\mu}_d$ with largest absolute values. This is illustrated by the next stylized example.

Let $\mathbf{X}|Y=0\sim\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{\Sigma})$ and $\mathbf{X}|Y=1\sim\mathcal{N}(\boldsymbol{\mu}_2,\mathbf{\Sigma})$, where $\boldsymbol{\mu}_1=(0,0,0)^T$, $\boldsymbol{\mu}_2=(4,0.5,1)^T$, and

$$\mathbf{\Sigma}=\begin{pmatrix}1 & -0.25 & 0\\ -0.25 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}.$$

Suppose the objective is to choose 2 out of 3 variables for classification. If we rank features by marginal information, for example by the absolute value of standardized mean differences, then we would choose the 1st and 3rd features. On the other hand, denote $\boldsymbol{\mu}_{d,ij}$ the mean difference vector for features i and j, $\mathbf{\Sigma}_{ij}$ the covariance matrix of features i and j, then the classification power using features i and j depends on $\Gamma_{ij}=\boldsymbol{\mu}_{d,ij}^T\mathbf{\Sigma}_{ij}^{-1}\boldsymbol{\mu}_{d,ij}$. Simple calculation leads to

$$\Gamma_{12}=18.4>17=\Gamma_{13}\,.$$

Hence the most powerful two features for classification are not the 1st and 3rd.

4. Constrained Coordinate Descent

With a Lagrangian argument, we reformulate problem (10) as

$$\bar{\mathbf{w}}_\lambda=\underset{\mathbf{w}^T\boldsymbol{\mu}_d=1}{\operatorname{argmin}}\,\frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}+\lambda\|\mathbf{w}\|_1. \tag{13}$$

In this section, we propose a Constrained Coordinate Descent (CCD) algorithm that is tailored for solving our minimization problem with linear constraints. Optimization (13) is a constrained quadratic programming problem and can be solved by existing softwares such as MOSEK. Although these softwares are well regarded in practice, they are slow for our application. The structure of (13) could be exploited in order to obtain a more efficient algorithm. In line with the LARS algorithm, we will exploit the fact that the solution path has a piecewise-linear property.

In the compressed sensing literature, it is common to replace an affine constraint by a quadratic penalty. We borrow this idea and consider the following approximation to (13):

$$\tilde{\mathbf{w}}_{\lambda,\gamma}=\operatorname{argmin}\frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}+\lambda\|\mathbf{w}\|_1+\frac{1}{2}\gamma(\mathbf{w}^T\boldsymbol{\mu}_d-1)^2\,. \tag{14}$$

In practice, we replace $\mathbf{\Sigma}$ by the pooled sample covariance $\hat{\mathbf{\Sigma}}$ and $\boldsymbol{\mu}$ by the sample mean difference vector $\hat{\boldsymbol{\mu}}_d$. By Theorem 6.7 in Ruszczynski (2006), we have

$$\tilde{\mathbf{w}}_{\lambda,\gamma}\to\bar{\mathbf{w}}_\lambda \text{ when } \gamma\to\infty.$$

Note that we do not have to enforce the affine constraint strictly, because it only serves to normalize our problem. In the optimization problem (14), when $\lambda=0$, the solution $\tilde{\mathbf{w}}_{0,\gamma}$ is always in the direction of

$\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$, the Fisher discriminant, regardless of the value of γ. In addition, this observation is confirmed in the data analysis (Section 6.2) by the insensitivity of choice for γ. Therefore we hold γ as a constant in practice.

We solve (14) by coordinate descent. Non-gradient algorithms seem to be less popular for convex optimization. For instance, the popular textbook *Convex Optimization* by Boyd and Vandenberghe (2004) does not even have a section on these methods. Coordinate descent method is an algorithm, in which the p search directions are just unit vectors $e_1, \cdots, e_p$, where e_i denotes the ith element in the standard basis of $\mathbb{R}^p$. These unit vectors are used as search directions in each search cycle until some convergence criterion is met.

What makes the coordinate descent algorithm particularly attractive for (14) is that there is an explicit formula for each coordinate update. For a given γ, fix τ and K, then do the optimization on a grid (of log-scale) of λ values: $\tau\lambda_{\max} = \lambda_K < \lambda_{K-1} < \cdots < \lambda_1 = \lambda_{\max}$. The $\lambda_{\max}$ is the minimum λ value such that no variables enter the model; this is analogous to the minimum requirement on c in (11). In our implementation, we take $\tau = 0.001$ and $K = 100$. The problem is solved backwards from $\lambda_{\max}$. When $\lambda = \lambda_{i+1}$, we use the solution from $\lambda = \lambda_i$ as the initial value. This kind of "warm start" is very effective in improving computational efficiency.

Consider a coordinate descent step to solve (14). Without loss of generality, suppose that $\tilde{w}_j$ for all $j \geq 2$ are given, and we need to optimize with respect to w_1. The objective function now becomes

$$g(w_1) = \frac{1}{2}\begin{pmatrix} w_1^T & \tilde{\mathbf{w}}_2^T \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} w_1 \\ \tilde{\mathbf{w}}_2 \end{pmatrix} + \lambda|w_1| + \lambda|\tilde{\mathbf{w}}_2|_1 + \frac{1}{2}\gamma(\mathbf{w}^T\boldsymbol{\mu}_d - 1)^2.$$

When $w_1 \neq 0$, we have

$$\begin{aligned} g'(w_1) &= \Sigma_{11}w_1 + \mathbf{\Sigma}_{12}\tilde{\mathbf{w}}_2 + \lambda\ \mathrm{sign}(w_1) + \gamma(\mathbf{w}^T\boldsymbol{\mu}_d - 1)\mu_{d1} \\ &= (\Sigma_{11} + \gamma\mu_{d1}^2)w_1 + (\mathbf{\Sigma}_{12} + \gamma\mu_{d1}\boldsymbol{\mu}_{d2}^T)\tilde{\mathbf{w}}_2 + \lambda\ \mathrm{sign}(w_1) - \gamma\mu_{d1}. \end{aligned}$$

By simple calculation (Donoho and Johnstone, 1994), the coordinate-wise update has the form

$$\tilde{w}_1 = \frac{S\left(\gamma\mu_{d1} - (\mathbf{\Sigma}_{12} + \gamma\mu_{d1}\boldsymbol{\mu}_{d2}^T)\tilde{\mathbf{w}}_2, \lambda\right)}{\Sigma_{11} + \gamma\mu_{d1}^2},$$

where $S(z, \lambda) = \mathrm{sign}(z)(|z| - \lambda)^+$ is the soft-thresholding operator.

Now, we consider the convergence property of the coordinate descent algorithm. Here, although the objective function is not strictly convex, it is strictly convex in each of the coordinates.

To show $g(w_1)$ is strictly convex in w_1, we decompose it as follows:

$$g(w_1) = g_1(w_1) + g_2(w_1),$$

where $g_2(w_1) = \lambda|w_1|$ and

$$g_1(w_1) = \frac{1}{2}\begin{pmatrix} w_1^T & \tilde{\mathbf{w}}_2^T \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} w_1 \\ \tilde{\mathbf{w}}_2 \end{pmatrix} + \lambda|\tilde{\mathbf{w}}_2|_1 + \frac{1}{2}\gamma(\mathbf{w}^T\boldsymbol{\mu}_d - 1)^2\,.$$

Note that $g_1(w_1)$ is a quadratic function of w_1 and $g_1''(w_1) = \Sigma_{11} + \gamma\mu_{d1}^2 > 0$ for all $w_1 \in \mathbb{R}$. Therefore, the function $g_1(\cdot)$ is strictly convex on $\mathbb{R}$. Also, it is clear that g_2 is convex on $\mathbb{R}$. Therefore $g = g_1 + g_2$ is a strictly convex function on $\mathbb{R}$.

Combining the coordinate-wise strict convexity with the fact that the non-differentiable part of the objective function is separable, Theorem 5.1 of Tseng (2001) guarantees that coordinate descent algorithms

converge to coordinate-wise minima. Moreover, since all directional derivatives exist, every coordinate-wise minimum is also a local minimum. A similar study on the convergence of the coordinate descent algorithm can be found in Breheny and Huang (2011).

In each coordinate update, the computational complexity is $\mathcal{O}(p)$. A complete cycle through all p variables costs $\mathcal{O}(p^2)$ operations. From our experience, CCD converges quickly after a few cycles if "warm start" is used for the initial solution. Let C denote the average number of cycles until convergence for each λ. Then our algorithm CCD enjoys computational complexity $\mathcal{O}(CKp^2)$. The D-ROAD can be similarly implemented by replacing the covariance matrix with its diagonal.

5. Asymptotic Property

5.1. Risk Approximation

Let $\hat{\mathbf{w}}_c$ be a sample version of $\mathbf{w}_c$ in (10),

$$\hat{\mathbf{w}}_c \in \underset{\|\mathbf{w}\|_1 \le c, \mathbf{w}^T \hat{\boldsymbol{\mu}}_d = 1}{\operatorname{argmin}} \mathbf{w}^T \hat{\boldsymbol{\Sigma}} \mathbf{w}. \tag{15}$$

The fact that $\hat{\boldsymbol{\Sigma}}$ is only positive semi-definite leads to potential non-uniqueness of $\hat{\mathbf{w}}_c$. Now, we have three different classifiers: $\delta_{\mathbf{w}_\infty} = \mathbb{I}\{\mathbf{w}_\infty^T(\mathbf{X} - \boldsymbol{\mu}_a) > 0\}$, $\delta_{\mathbf{w}_c} = \mathbb{I}\{\mathbf{w}_c^T(\mathbf{X} - \boldsymbol{\mu}_a) > 0\}$ and $\hat{\delta}_{\mathbf{w}_c} = \mathbb{I}\{\hat{\mathbf{w}}_c^T(\mathbf{X} - \hat{\boldsymbol{\mu}}_a) > 0\}$. The first two are oracle classifiers, requiring the knowledge of unknown parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$, while the third one is the feasible classifier, ROAD, based on the sample. Their classification errors are given by (2). Explicitly, the error rates are respectively $W(\delta_{\mathbf{w}_\infty})$ [see (4)], $W(\delta_{\mathbf{w}_c})$, and $W(\hat{\delta}_{\mathbf{w}_c})$. By (2), an obvious estimator of the misclassification rate of $\hat{\delta}_{\mathbf{w}_c}$ is

$$W_n(\hat{\delta}_{\mathbf{w}_c}) = 1 - \Phi\left(\frac{\hat{\mathbf{w}}_c^T \hat{\boldsymbol{\mu}}_d}{(\hat{\mathbf{w}}_c^T \hat{\boldsymbol{\Sigma}} \hat{\mathbf{w}}_c)^{1/2}} \right). \tag{16}$$

Two questions arise naturally:

(i) how close is $W(\hat{\delta}_{\mathbf{w}_c})$, the misclassification error of $\hat{\delta}_{\mathbf{w}_c}$, to that of its oracle $W(\delta_{\mathbf{w}_c})$?
(ii) does $W_n(\hat{\delta}_{\mathbf{w}_c})$ estimate $W(\hat{\delta}_{\mathbf{w}_c})$ well?

Theorem 1 addresses these two questions. We introduce an intermediate optimization problem for convenience:

$$\mathbf{w}_c^{(1)} = \underset{\|\mathbf{w}\|_1 \le c, \mathbf{w}^T \hat{\boldsymbol{\mu}}_d = 1}{\operatorname{argmin}} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}.$$

THEOREM 1. *Let $s_c = \|\mathbf{w}_c\|_0$, $s_c^{(1)} = \|\mathbf{w}_c^{(1)}\|_0$, and $\hat{s}_c = \|\hat{\mathbf{w}}_c\|_0$. Assume that $\lambda_{\min}(\boldsymbol{\Sigma}) \ge \sigma_0^2 > 0$, $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty = O_p(a_n)$ and $\|\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|_\infty = O_p(a_n)$ for a given sequence $a_n \to 0$. Then, we have*

$$W(\hat{\delta}_{\mathbf{w}_c}) - W(\delta_{\mathbf{w}_c}) = O_p(d_n)\,,$$

and

$$W_n(\hat{\delta}_{\mathbf{w}_c}) - W(\hat{\delta}_{\mathbf{w}_c}) = O_p(b_n)\,,$$

where $b_n = \left(c^2 \vee s_c \vee s_c^{(1)}\right) a_n$ and $d_n = b_n \vee (\hat{s}_c a_n)$.

REMARK 1. *In Theorem 1, $\|\cdot\|_\infty$ is the element wise super-norm. When $\hat{\boldsymbol{\Sigma}}$ is the sample covariance, by Bickel and Levina (2004), $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty = O_p(\sqrt{(\log p)/n})$; hence we can take $a_n = \sqrt{(\log p)/n}$. The first result in Theorem 1 shows the difference between the misclassification rate of $\hat{\delta}_{\mathbf{w}_c}$ and its oracle version $\delta_{\mathbf{w}_c}$; the second result says about the error in estimating the true misclassification rate of ROAD.*

REMARK 2. *In view of (2), one intends to choose a* $\mathbf{w}$ *that makes* $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ *small and* $\mathbf{w}^T\boldsymbol{\mu}_d$ *large. A compromise of these dual objectives leads to a utility function*

$$U(\mathbf{w}) = -\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} + \xi\boldsymbol{\mu}_d^T\mathbf{w},$$

as a proxy of the objective function (2) for a fixed ξ*. For any* $\xi > 0$*, the optimal choice* $\mathbf{w}^* \in \operatorname{argmin} U(\mathbf{w})$ *leads to the Fisher discriminant rule. Consider also the regularized versions*

$$\mathbf{w}_c^* = \operatorname{argmin}_{\|\mathbf{w}\|_1 \le c} U(\mathbf{w}), \quad \textit{and} \quad \hat{\mathbf{w}}_c^* = \operatorname{argmin}_{\|\mathbf{w}\|_1 \le c} \hat{U}(\mathbf{w}),$$

where $\hat{U}(\mathbf{w})$ *is the utility function with* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\mu}_d$ *estimated by* $\hat{\boldsymbol{\Sigma}}$ *and* $\hat{\boldsymbol{\mu}}_d$*. Then, it is easy to see the following utility approximation: for any* $\|\mathbf{w}\|_1 \le c$

$$|U(\mathbf{w}) - \hat{U}(\mathbf{w})| \le \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty c^2 + \xi c\|\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|_\infty$$

and

$$|U(\hat{\mathbf{w}}_c^*) - U(\mathbf{w}_c^*)| \le 2\Big(\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty c^2 + \xi c\|\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|_\infty\Big).$$

REMARK 3. *The most prominent technical challenge of our original problem (10) is due to different dualities of penalization problems. For the population version (10), it can be reduced, by the Lagrange multiplier method, to the utility* $U(\mathbf{w})$ *optimization problem in Remark 2 with a given* $\xi > 0$*, while for the sample version (15), it can be reduced to the utility* $\hat{U}(\mathbf{w})$ *optimization problem with a different* $\hat{\xi}$*. Therefore, the problem is not the same as the utility optimization problem in Remark 2:* $\hat{\xi}$ *is hard to bound. In fact, it is much harder and yields more complicated results.*

We now show how different the data projection direction in the regularized oracle can be from that in the Fisher discriminant. To gain better insight, we reformulate the L_1 constraint problem as the following penalized version:

$$\mathbf{w}^\lambda = \operatorname*{argmin}_{\mathbf{w}:\boldsymbol{\mu}_d^T\mathbf{w}=1} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} + \lambda\|\mathbf{w}\|_1. \tag{17}$$

The following characterizes its convergence to the Fisher discriminant weight $\mathbf{w}_\infty$ as $\lambda \to 0$.

THEOREM 2. *Let* s *be the size of the set* $\{k : (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d)_k \ne 0\}$*. Then, we have*

$$\|\mathbf{w}^\lambda - \mathbf{w}_\infty\|_2 \le \frac{\lambda\sqrt{s}}{\lambda_{\min}(\boldsymbol{\Sigma})},$$

where $\mathbf{w}_\infty = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d/(\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d)$ *is the normalized Fisher discriminant, optimizing (17) with* $\lambda = 0$*.*

5.2. Screening-based ROAD (S-ROAD)

Following the idea of Sure Independence Screening in Fan and Lv (2008), we pre-screen all the features before hitting the ROAD. The advantage of this two-step procedure is that we have a control on the total number of features used in the final classification rule. A popular method for independent feature selection is the two-sample t-test (Tibshirani *et al.*, 2002; Fan and Fan, 2008), which is a specific case of marginal screening in Fan and Lv (2008). The sure screening property of such a method was demonstrated in Fan and Fan (2008), which selects consistently the features with different means in the same settings as ours.

Once the features are selected, we can hit the ROAD, producing the vanilla Screening-based Regularized Optimal Affine Discriminant (S-ROAD1):

(1) Employ a screening method to get k features.
(2) Apply ROAD to the k selected features.

In the first step, we use the t-statistics as the screening criteria and determine a data-driven threshold. This idea is motivated by a FDR criterion for choosing marginal screening threshold in Zhao and Li (2010). A random permutation π of $\{1, \cdots, n\}$ is used to decouple $\mathbf{X}_i$ and Y_i so that the resulting data $(\mathbf{X}_{\pi(i)}, Y_i)$ follow a null model, by which we mean that features have no prediction power for the class label. More specifically, the screening step is carried out as follows:

(i) Calculate the t-statistic t_j for each feature j, where $j = 1, \cdots, p$.
(ii) For the permuted data pairs $(\mathbf{X}_{\pi(i)}, Y_i)$, recalculate the t-statistic t_j^*, for $j = 1, \cdots, p$. (Intuitively, if j is the index of an important feature, $|t_j|$ should be larger than most of $|t_j^*|$, because the random permutation is meant to eliminate the prediction power of features.)
(iii) For $q \in [0, 1]$, let $\omega_{(q)}$ be the q^{th} quantile of $\{|t_j^*|, j = 1, 2, \cdots, p\}$. Then, the selected set $\mathcal{A}$ is defined as

$$\mathcal{A} = \{j || t_j | \geq \omega_{(q)}\}.$$

The choice of threshold is made to retain the features whose t-statistics are significant in the two sample t-test. Alternatively, if the user knows his k, (due to budget constraints, etc.), then he can just rank $|t_j|$'s and choose the threshold accordingly.

The S-ROAD1 tracks the performance of oracle procedures like sub-Fisher (10 features) in Figure 1. The feature space gotten by step (1) can be expanded by including those features which are most correlated with what have already been selected. This additional variant, S-ROAD2, aims at achieving the performance of sub-Fisher (20 features) type of procedure in Figure 1.

To elaborate on the theoretical properties of S-ROAD1, assume with no loss of generality that the first k variables are selected in the screening step. Denote by $\mathbf{\Sigma}_k$ the upper left $k \times k$ block of $\mathbf{\Sigma}$ and $\boldsymbol{\mu}_k$ the first k coordinates of $\boldsymbol{\mu}_d$. Let

$$\boldsymbol{\beta}_c = \underset{\|\boldsymbol{\beta}\|_1 \leq c, \boldsymbol{\beta}^T \boldsymbol{\mu}_k = 1}{\operatorname{argmin}} \boldsymbol{\beta}^T \mathbf{\Sigma}_k \boldsymbol{\beta}.$$

The quantities $\hat{\boldsymbol{\beta}}_c$ and $\boldsymbol{\beta}_c^{(1)}$ are defined similarly to $\hat{\mathbf{w}}_c$ and $\mathbf{w}_c^{(1)}$ (defined right before Theorem 1). Then denote by $\boldsymbol{y}_c = (\boldsymbol{\beta}_c^T, \mathbf{0}^T)^T$, $\hat{\boldsymbol{y}}_c = (\hat{\boldsymbol{\beta}}_c^T, \mathbf{0}^T)^T$ and $\boldsymbol{y}_c^{(1)} = (\mathbf{w}_c^{(1)}, \mathbf{0}^T)^T$. The next two theorems can be verified along lines similar to Theorems 1 and 2. Hence, the proofs are omitted.

THEOREM 3. *If* $\|\hat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k\|_\infty = O_p(\sqrt{\log k/n})$, $\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty = O_p(\sqrt{\log k/n})$, *and* $\lambda_{\min}(\mathbf{\Sigma}_k) \geq \delta_0 > 0$, *then we have*

$$W(\hat{\delta}_{\boldsymbol{y}_c}) - W(\delta_{\boldsymbol{y}_c}) = O_p(e_n),$$

and

$$W_n(\hat{\delta}_{\boldsymbol{y}_c}) - W(\delta_{\boldsymbol{y}_c}) = O_p(e_n),$$

where $e_n = (c^2 \vee k)\sqrt{\frac{\log k}{n}}$.

This result is cleaner than Theorem 1, as the rate does not involve s_c and $\hat{s}_c$: they are simply replaced by the upper bound k. Accurate bounds for s_c and $\hat{s}_c$ are of interest for future exploration, but they are beyond the scope of this paper.

THEOREM 4. *Let* $\boldsymbol{y}_k^\lambda = \operatorname{argmin}_{\boldsymbol{y}:\boldsymbol{\mu}_d^T\boldsymbol{y}=1,\boldsymbol{y}\in M_k} R(\boldsymbol{y}) + \lambda\|\boldsymbol{y}\|_1$ *where* M_k *is the subspace in* R^p *with the last* $p-k$ *components being zero, and* $\boldsymbol{y}^0 = ((\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)^T/(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k), \mathbf{0}^T)^T$. *Then we have*

$$\|\boldsymbol{y}_k^\lambda - \boldsymbol{y}^0\|_2 \leq \frac{\lambda\sqrt{k}}{\lambda_{\min}(\boldsymbol{\Sigma}_k)}.$$

5.3. Continuous Piecewise Linear Solution Path

We use the word "linear" when referring to "affine", in line with the *status quo* in the statistical community. Continuous piecewise linear paths are of much interest to statisticians, as the property reduces the computational complexity of solutions and justifies the linear interpolations of solutions at discrete points. Previous well known investigations include Efron *et al.* (2004) and Rosset and Zhu (2007). Our setup differs from others mainly in that in addition to a complexity penalty, there is also an affine constraint. Our proof calls in point set topology, and is purely geometrical, in a spirit very different from the existing ones. In particular, we stress that the continuity property is intuitively correct, but it is far from a trivial consequence of the assumptions. The authors also believe that the claim holds true even if the $p-1$ dimensional affine subspace constraint is replaced by more generic ones, though the technicality of the proof must be more involved.

THEOREM 5. *Let* $\boldsymbol{\mu}_d \in \mathbb{R}^p$ *be a constant, and* $\boldsymbol{\Sigma}$ *be a positive definite matrix of dimension* $p \times p$. *Let*

$$\mathbf{w}_c = \operatorname*{argmin}_{\|\mathbf{w}\|_1 \leq c, \mathbf{w}^T\boldsymbol{\mu}_d = 1} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w},$$

then $\mathbf{w}_c$ *is a continuous piecewise linear function in* c.

PROPOSITION 1. $W(\delta_{\mathbf{w}_c})$ *is a Lipschitz function in* c.

PROOF. *Recall that*

$$W(\delta_{\mathbf{w}_c}) = 1 - \Phi\left(1/(R(\mathbf{w}_c))^{1/2}\right).$$

By Theorem 5 and the fact that composition of Lipschitz functions is again Lipschitz, the conclusion holds.

6. Numerical Investigation

In this section, several simulation and real data studies are conducted. We compare ROAD and its variants S-ROAD1 (Screening-based ROAD version 1), S-ROAD2 (Screening-based ROAD version 2) and D-ROAD (Diagonal ROAD) with NSC (Nearest Shrunken Centroid), SCRDA (Shrunken Centroids Regularized Discriminant Analysis), FAIR (Feature Annealed Independence Rule), NB (Naive Bayes), NFR (Naive Fisher Rule, which uses the generalized inverse of the sample covariance matrix), as well as the Oracle.

In all simulation studies, the number of variables is $p = 1000$, and the sample size of the training and testing data is $n = 300$ for each class. Each simulation is repeated 100 times to test the stability of the method. Without loss of generality, the mean vector of the first class $\boldsymbol{\mu}_1$ is set to be $\mathbf{0}$. We use five-fold cross-validation to choose the penalty parameter λ.

Fig. 2. Solution Path for ROAD (left panel) and D-ROAD (right panel). Equal correlation setting ($\rho = 0.5$), Sparse Signal ($s_0 = 10$) as in Section 6.1.

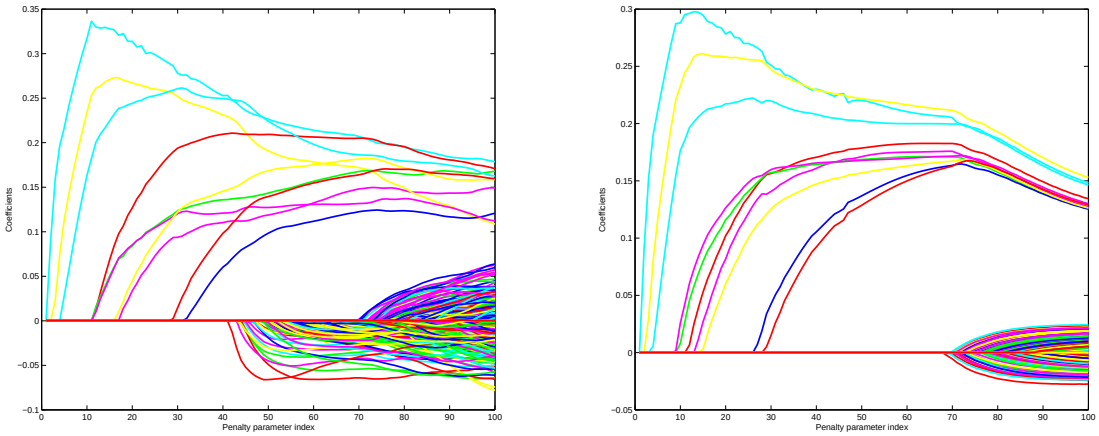


Table 1. Equal correlation setting, fixed signal: Median of the percentage for testing classification error and standard deviations (in parentheses). Signal all equal to 1. $s_0 = 10$.

ρ	ROAD	S-ROAD1	S-ROAD2	D-ROAD	SCRDA	NSC	FAIR	NB	Oracle
0	6.0(1.2)	6.0(1.1)	6.0(1.2)	5.7(1.1)	6.3(1.0)	5.9(1.0)	5.7(1.0)	11.2(1.4)	5.5(1.1)
0.1	6.3(2.5)	12.2(5.0)	8.8(2.4)	11.6(5.1)	10.3(1.4)	11.1(3.0)	12.4(1.4)	26.8(10.1)	5.0(0.9)
0.2	5.3(1.0)	16.0(6.3)	8.7(2.5)	16.1(7.5)	8.5(1.2)	14.5(4.3)	17.3(1.7)	34.8(11.6)	4.0(0.8)
0.3	4.2(0.9)	19.1(7.9)	7.8(2.6)	19.1(9.4)	6.6(1.1)	17.1(5.5)	20.8(1.7)	39.3(12.3)	3.2(0.7)
0.4	3.2(0.8)	22.8(9.4)	6.5(2.6)	22.2(9.9)	4.8(1.0)	20.5(6.1)	23.2(1.8)	41.6(11.3)	2.0(0.6)
0.5	2.0(0.6)	25.8(11.0)	4.8(1.4)	25.2(10.2)	2.9(0.7)	23.2(6.0)	25.3(1.6)	43.5(11.1)	1.3(0.5)
0.6	1.0(0.4)	18.3(12.4)	3.3(1.3)	28.1(10.3)	1.5(0.5)	25.8(5.7)	26.8(1.8)	44.4(12.1)	0.7(0.3)
0.7	0.3(0.2)	15.5(13.6)	1.7(1.0)	29.1(10.1)	0.5(0.3)	27.0(8.2)	28.2(2.0)	45.2(12.3)	0.2(0.2)
0.8	0.0(0.1)	5.0(14.0)	0.3(0.4)	29.5(9.9)	0.0(0.1)	28.3(8.7)	29.2(2.0)	46.2(10.3)	0.0(0.1)
0.9	0.0(0.0)	0.6(14.8)	0.0(0.1)	30.3(7.6)	0.0(0.2)	29.9(8.0)	30.2(1.9)	46.8(8.8)	0.0(0.0)

6.1. Equal Correlation Setting, Sparse Fixed Signal

In this subsection, we consider the setting where $\Sigma_{i,i} = 1$ for all $i = 1, \cdots, p$ and $\Sigma_{i,j} = \rho$ for all $i, j = 1, \cdots, p$ and $i \neq j$, and take $\boldsymbol{\mu}_2$ to be a sparse vector: $\boldsymbol{\mu}_2 = (\mathbf{1}_{10}^T, \mathbf{0}_{990}^T)^T$, where $\mathbf{1}_d$ is a length d vector with all entries 1, $\mathbf{0}_d$ is a length d vector with all entries 0, where the sparsity size is $s_0 = 10$. Also, we fix $\gamma = 10$ in (14) for this simulation. Sensitivity of the performance due to the choice of γ will be investigated in the next subsection.

The solution paths for ROAD and D-ROAD of one realization are rendered in Figure 2. It is clear from the figure that, as the penalty parameter decreases (index increases), both ROAD and D-ROAD use more features. Also, the cutoff point for D-ROAD, where the number of features starts to increase dramatically, tends to come later than that for ROAD.

The simulation results for the pairwise correlations ranging from 0 to 0.9 are shown in Tables 1 and 2. We would like to mention that the results for NFR (Naive Fisher Rule) are not included in these (and the subsequent) tables because the test classification error is always around 50%, i.e., it is about the same as random guess. Also in the tables are the screening-based versions of the ROAD. S-ROAD1 refers to the

Table 2. Equal correlation setting, fixed signal: Median of number of nonzero coefficients and standard deviations (in parentheses). Signal all equal to 1. $s_0 = 10$.

ρ	ROAD	S-ROAD1	S-ROAD2	D-ROAD	SCRDA	NSC	FAIR
0	16.00(24.16)	10.00(1.31)	17.00(4.31)	29.50(58.54)	10.00(13.25)	10.00(44.86)	11.00(1.62)
0.1	117.50(30.50)	11.00(3.32)	21.00(4.15)	14.00(122.02)	1000.00(345.48)	35.50(117.32)	10.00(0.27)
0.2	130.50(33.33)	11.00(6.99)	22.00(8.98)	15.50(111.42)	1000.00(0.00)	95.00(120.17)	10.00(0.69)
0.3	136.50(36.16)	11.00(11.56)	22.00(10.38)	17.50(106.16)	1000.00(0.00)	103.50(117.52)	9.00(1.19)
0.4	135.00(34.43)	10.00(14.21)	22.00(17.07)	10.00(98.10)	1000.00(0.00)	70.00(131.65)	8.00(1.33)
0.5	138.50(38.17)	9.00(21.71)	22.00(21.56)	10.00(105.33)	1000.00(0.00)	65.00(137.97)	7.00(1.30)
0.6	148.00(49.74)	10.50(27.92)	22.00(31.88)	10.00(110.23)	1000.00(0.00)	38.00(141.91)	6.00(1.30)
0.7	170.50(52.29)	11.00(37.37)	22.00(41.76)	1.00(118.43)	1000.00(0.00)	27.50(140.10)	5.00(1.20)
0.8	203.00(27.72)	12.00(50.36)	24.00(59.23)	1.00(143.83)	1000.00(10.92)	15.00(157.98)	5.00(1.29)
0.9	151.50(8.02)	14.00(55.32)	28.00(50.45)	1.00(153.27)	1000.00(56.30)	14.00(225.38)	3.00(1.08)

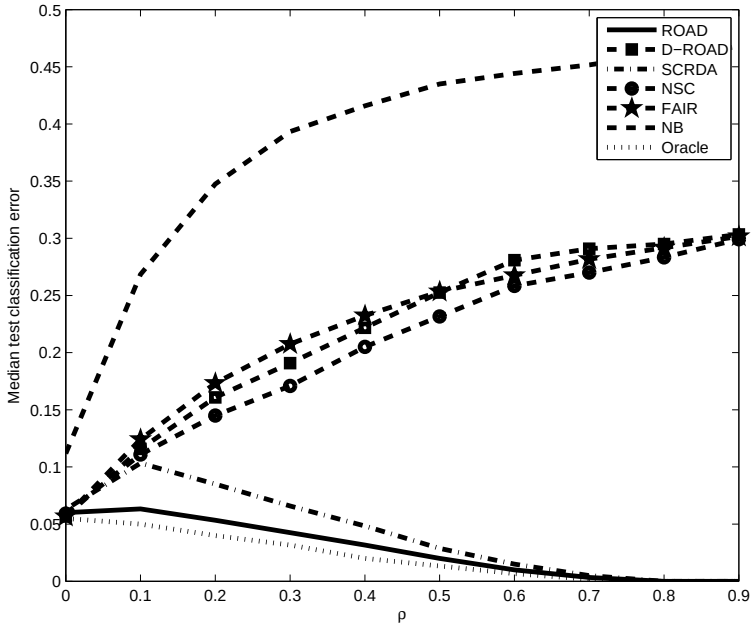
vanilla version where we first apply the two-sample t-test to select any features with the corresponding t-test statistic with absolute value larger than the maximum absolute t-test statistic value calculated on the permuted data. S-ROAD2 does the same except for each variable in S-ROAD1's pre-screened set, it adds an additional variable which is most correlated with that variable. Figure 3, a graphical summary of Table 1, presents the median test errors for different methods. We can see from Table 1 and Figure 3 that the oracle classification error decreases as ρ increases. This phenomenon is due to a similar reason to the two-dimensional showcase in the introduction. When ρ goes to 1, all the variables contribute in the same way to boost the classification power. ROAD performs reasonably close to the Oracle, while working independence based method such as D-ROAD, NSC, FAIR and NB fail when ρ is large. The huge discrepancy shows the advantage of employing the correlation structure. Since SCRDA also employ the correlation structure, it does not fail when ρ is large. However, ROAD still outperforms SCRDA in all the correlation settings. S-ROAD1 and S-ROAD2 both have misclassification rates similar to that of ROAD. It is worth to emphasize that the merits of the screening based ROADs mainly lie in the computation cost, which is reduced significantly by the pre-screening step.

The ROAD is a very robust estimator. It performs well even when all the variables are independent, in which case there could be a lot of noise for fitting the covariance matrix. Table 1 indicates that ROAD has almost the same performance as D-ROAD, NSC and FAIR under the independence assumption, i.e. $\rho = 0$. As ρ increases, the edge of ROAD becomes more substantial. In general, the ROAD is recommended on the grounds that even with pairwise correlation of about 0.1 (which is quite common in microarray data as well as financial data), the gain is substantial.

Another interesting observation is that the D-ROAD performs similarly to NSC and FAIR in terms of classification error. An intuitive explanation is that they are all "sparse" independence rules. NSC uses soft-thresholding on the standardized sample mean difference, and its equivalent LASSO derivation can be found in Wang and Zhu (2007). FAIR selects features with large marginal t-statistics in absolute values, while D-ROAD is another L1 penalized independence rule, whose implementation is different from NSC.

Table 2 summarizes the number of features selected by different classifiers. Note that ROAD mimics Fisher discriminant coordinate $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$, which has $p = 1000$ nonzero entries under our simulated model. Therefore, the large number of features selected by ROAD is not out of expectation.

Fig. 3. Median classification error as a function of ρ in the equi-correlation matrix. Sparse $\boldsymbol{\mu}_d$ as in Section 6.1.



6.2. The Effect of γ

Under the settings of the previous subsection, we look into the variation of the ROAD performance as γ changes. In Table 3, the number of active variables varies; however, the median classification error remains about the same for a broad range of γ values. The reason is that the cross validation step chooses the "best" λ according to a specific γ. Therefore, the final performance remains almost unchanged. Since our primary concern is the classification error, we fix $\gamma = 10$ for simplicity in the subsequent simulations and in the real data analysis.

6.3. Block Diagonal Correlation Setting, Sparse Fixed Signal

In this subsection, we follow the same setup as in Section 6.1 except that the covariance matrix $\boldsymbol{\Sigma}$ is taken to be block diagonal. The first block is a 20×20 equi-correlated matrix and the second block is a $(p-20) \times (p-20)$ equi-correlated matrix, both with pairwise correlation ρ. In other words, $\Sigma_{i,i} = 1$ for all $i = 1, \cdots, p$, $\Sigma_{i,j} = \rho$ for all $i, j = 1, \cdots, 20$ and $i \neq j$, $\Sigma_{i,j} = \rho$ for all $i, j = 21, \cdots, p$ and $i \neq j$, and the rest elements are zeros. As before, we examine the performances of various estimators when ρ varies. The percentage for testing error and the number of selected features in the estimators are shown in Tables 4 and 5, respectively.

In this block-diagonal setting, we have observed similar results to those in Section 6.1: ROAD and S-ROAD2 perform significantly better than the other methods. One interesting phenomenon is that S-ROAD1 does not perform well when ρ is large. The reason is that the current true model has 20 important features, and by looking only at marginal contribution, S-ROAD1 misses some important variables, as shown in Table 4. Indeed, because those features have no expressed mean differences, it does not fully take advantage of highly correlated features. In contrast, S-ROAD2 is able to pick up all the important

Table 3. Equal correlation setting; signals all equal to 1; $s_0 = 10$. Results for different γ.

		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
Median classification error (in percentage)	$\text{ROAD}_{\gamma=0.01}$	5.8(1.2)	2.7(0.6)	0.2(0.2)
	$\text{ROAD}_{\gamma=0.1}$	6.0(1.2)	2.0(0.6)	0.2(0.1)
	$\text{ROAD}_{\gamma=1}$	6.0(1.3)	2.0(0.6)	0.0(0.1)
	$\text{ROAD}_{\gamma=10}$	6.0(1.2)	2.0(0.6)	0.0(0.0)
	$\text{ROAD}_{\gamma=100}$	6.2(1.2)	2.3(0.6)	0.0(0.1)
		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
Median number of nonzeros	$\text{ROAD}_{\gamma=0.01}$	14.0(19.2)	129.5(42.5)	657.0(179.6)
	$\text{ROAD}_{\gamma=0.1}$	14.0(19.6)	137.0(37.6)	773.5(103.2)
	$\text{ROAD}_{\gamma=1}$	16.5(22.9)	139.0(37.9)	514.0(39.7)
	$\text{ROAD}_{\gamma=10}$	16.0(24.2)	138.5(38.2)	151.5(8.0)
	$\text{ROAD}_{\gamma=100}$	22.0(16.1)	114.5(9.4)	94.0(9.6)

Table 4. Block diagonal correlation setting, sparse fixed signal: Median of the percentage for testing classification error and standard deviations (in parentheses). Signal all equal to 1. $s_0 = 10$.

ρ	ROAD	S-ROAD1	S-ROAD2	D-ROAD	SCRDA	NSC	FAIR	NB	Oracle
0	6.0(1.2)	6.0(1.1)	6.0(1.2)	5.7(1.1)	6.0(0.1)	5.5(0.3)	5.7(1.0)	11.2(1.4)	5.5(1.1)
0.1	10.8(3.6)	13.0(4.8)	10.3(3.0)	12.8(4.4)	13.0(0.3)	12.5(0.8)	12.7(1.5)	25.7(7.6)	8.8(1.2)
0.2	10.7(4.1)	18.0(5.7)	9.7(3.6)	17.7(5.9)	14.2(1.1)	17.2(0.4)	17.7(1.6)	34.4(7.9)	8.8(1.2)
0.3	9.5(3.8)	23.2(5.5)	8.8(4.0)	23.2(5.6)	12.7(0.9)	20.0(0.8)	20.4(1.6)	38.3(7.5)	7.7(1.0)
0.4	8.0(3.3)	29.7(4.2)	7.5(4.2)	29.3(4.1)	11.0(1.2)	23.8(1.3)	23.2(1.8)	41.0(6.9)	6.6(1.1)
0.5	6.2(2.6)	30.1(3.9)	5.7(0.9)	30.0(3.1)	8.7(0.4)	26.2(1.7)	25.1(1.7)	42.2(6.6)	5.0(1.0)
0.6	4.2(0.9)	30.3(4.2)	4.0(0.8)	30.3(2.2)	6.4(0.1)	26.5(1.2)	26.8(1.8)	43.6(7.0)	3.5(0.7)
0.7	2.3(0.7)	30.0(6.4)	2.2(0.7)	30.6(2.1)	2.5(0.7)	28.1(3.2)	28.2(2.0)	44.2(6.5)	1.8(0.6)
0.8	0.8(0.4)	29.8(9.8)	0.7(0.4)	30.6(2.1)	0.6(0.4)	29.2(1.6)	29.2(2.0)	44.8(5.7)	0.7(0.3)
0.9	0.0(0.1)	29.8(12.8)	0.0(0.1)	30.6(1.9)	0.2(0.2)	29.2(1.2)	30.2(1.9)	45.2(4.9)	0.0(0.1)

Table 5. Block diagonal correlation setting, fixed signal: Median of number of nonzero coefficients and standard deviations (in parentheses). Signal all equal to 1. $s_0 = 10$.

ρ	ROAD	S-ROAD1	S-ROAD2	D-ROAD	SCRDA	NSC	FAIR
0	16.00(24.16)	10.00(1.31)	17.00(4.31)	29.50(58.54)	10.00(1.15)	10.00(1.73)	11.00(1.62)
0.1	48.50(35.99)	10.00(2.73)	20.00(3.77)	14.00(26.73)	33.00(17.79)	65.00(38.84)	18.00(2.67)
0.2	48.00(31.48)	10.00(4.59)	20.00(5.84)	10.00(18.23)	38.00(117.54)	10.00(16.17)	18.00(2.77)
0.3	47.50(42.75)	9.00(5.28)	20.00(6.03)	10.00(11.80)	208.00(103.94)	10.00(13.58)	18.00(3.91)
0.4	40.50(32.42)	1.00(4.82)	20.00(10.08)	1.00(9.25)	27.00(90.95)	33.00(14.22)	17.00(5.43)
0.5	40.50(33.23)	1.00(4.88)	20.00(10.10)	1.00(8.51)	24.00(76.79)	10.00(1.15)	7.00(5.98)
0.6	39.50(30.03)	1.00(3.74)	20.00(14.53)	1.00(5.92)	127.50(6.36)	6.50(2.12)	6.00(5.98)
0.7	40.00(41.35)	1.00(4.71)	20.00(8.07)	1.00(2.49)	94.50(2.12)	9.50(0.71)	5.00(5.52)
0.8	55.00(58.67)	1.00(6.20)	20.00(18.32)	1.00(0.93)	58.00(2.83)	6.00(5.66)	5.00(4.84)
0.9	120.00(30.66)	1.00(21.29)	20.00(30.46)	1.00(0.35)	20.00(0.00)	8.00(2.83)	3.00(3.81)

Table 6. Block-Diagonal Negative Correlation Setting, Sparse Fixed Signal: Median error (in percentage) and number of nonzero coefficients with standard deviations in parentheses.

	ROAD	S-ROAD1	S-ROAD2	D-ROAD	SCRDA	NSC	FAIR	NB	Oracle
error	7.3(3.4)	16.0(5.2)	12.7(3.4)	17.8(8.0)	18.5(1.1)	20.8(0.6)	24.8(2.1)	33.5(2.1)	3.2(0.7)
nonzero	168.00(47.59)	10.00(2.40)	20.00(3.58)	15.50(15.32)	24.00(0.58)	41.00(17.90)	59.00(4.27)	–	–

Table 7. Random correlation setting, double exponential signal: Median error (in percentage) and number of nonzero coefficients with standard deviations in parentheses.

	ROAD	S-ROAD1	S-ROAD2	D-ROAD	SCRDA	NSC	FAIR	NB	Oracle
error	2.0(0.6)	11.0(5.2)	5.8(3.9)	17.0(2.2)	5.2(1.1)	16.2(1.3)	17.0(1.6)	46.2(2.4)	1.3(0.5)
nonzero	83.00(39.54)	4.00(8.13)	9.00(10.69)	1.00(3.89)	1000.00(0.00)	4.00(0.58)	1.00(0.17)	–	–

variables, takes advantage of correlation structure, and leads to a sparser model than the vanilla ROAD. In view of the results from this simulation setting and the previous one, we recommend S-ROAD2 over S-ROAD1.

6.4. Block-Diagonal Negative Correlation Setting, Sparse Fixed Signal

In this subsection, we again follow a similar setup as in Section 6.1. Here, the covariance matrix $\mathbf{\Sigma}$ is taken to be block diagonal with each block size equals to 10. Each block is an equi-correlated matrix with pairwise correlation $\rho = -0.1$. In other words, $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Sigma}_0, \cdots, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0$ is a 10×10 equi-correlated matrix with correlation -0.1. Here, $\boldsymbol{\mu}_2 = 0.5 \times (\mathbf{1}_5^T, \mathbf{0}_5^T, \mathbf{1}_5^T, \mathbf{0}_{985}^T)^T$ and the sparsity size is $s_0 = 10$. As before, we examine the performances of various estimators when ρ varies. The percentage for testing error and the number of selected features in the estimators are shown in Table 6.

6.5. Random Correlation Setting, Double Exponential Signal

To evaluate the stability of the ROAD, we take a random matrix $\mathbf{\Sigma}$ as the correlation structure, and use a signal $\boldsymbol{\mu}$ whose nonzero entries come from a double exponential distribution. A random covariance matrix $\mathbf{\Sigma}$ is generated as follows:

(i) For a given integer m (here we take $m = 10$), generate a $p \times m$ matrix $\mathbf{\Omega}$ where $\Omega_{i,j} \sim \mathrm{Unif}(-1, 1)$. Then the matrix $\mathbf{\Omega}\mathbf{\Omega}^T$ is positive semi-definite.

(ii) Denote $c_{\mathbf{\Omega}} = \min_i(\mathbf{\Omega}\mathbf{\Omega}^T)_{ii}$. Let $\mathbf{\Xi} = \mathbf{\Omega}\mathbf{\Omega}^T + c_{\mathbf{\Omega}}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. It is clear that $\mathbf{\Xi}$ is positive definite.

(iii) Normalize the matrix $\mathbf{\Xi}$ to get $\mathbf{\Sigma}$ whose diagonal elements are unity.

For the signal, we take $\boldsymbol{\mu}$ to be a sparse vector with sparsity size $s = 10$, and the nonzero elements are generated from the double exponential distribution with density function

$$f(x) = \exp(-2|x|).$$

Table 7 summaries the results. It shows that even under random correlation setting and random signals, our procedure ROAD still outperforms other competing classification rules such as SCRDA, NSC and FAIR in terms of the classification error.

6.6. Real Data

Though the ROAD seems to perform best in a broad spectrum of idealized experiments, it has to be tested against reality. We now evaluate the performance of our newly proposed estimator on three popular

Table 8. Classification error and number of selected genes by various methods of leukemia data. Training and testing samples are of sizes 38 and 34, respectively.

	ROAD	S-ROAD1	S-ROAD2	SCRDA	FAIR	NSC	NB
Training Error	0	0	0	1	1	1	0
Testing Error	1	3	1	2	1	3	5
No. of selected genes	40	49	66	264	11	24	7129

Table 9. Classification error and number of selected genes by various methods of lung cancer data. Training and testing samples are of sizes 32 and 149, respectively.

	ROAD	S-ROAD1	S-ROAD2	SCRDA	FAIR	NSC	NB
Training Error	1	1	1	0	0	0	6
Testing Error	1	4	1	3	7	10	36
No. of selected genes	52	56	54	2410	31	38	12533

gene expression data sets: "Leukemia" (Golub *et al.*, 1999), "Lung Cancer" (Gordon *et al.*, 2002), and "Neuroblastoma data set" (Oberthuer *et al.*, 2006). The first two data sets come with predetermined, separate training and test sets of data vectors. The Leukemia data set contains $p = 7,129$ genes for $n_1 = 27$ acute lymphoblastic leukemia (ALL) and $n_2 = 11$ acute myeloid leukemia (AML) vectors in the training set. The test set includes 20 ALL and 14 AML vectors. The Lung Cancer data set contains $p = 12,533$ genes for $n_1 = 16$ adenocarcinoma (ADCA) and $n_2 = 16$ mesothelioma training vectors, along with 134 ADCA and 15 mesothelioma test vectors. The Neuroblastoma data set, obtained via the MicroArray Quality Control phase-II (MAQC-II) project, consists of gene expression profiles for $p = 10,707$ genes from 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004. We analyzed the gene expression data with the 3-year event-free survival (3-year EFS), which indicates whether a patient survived 3 years after the diagnosis of neuroblastoma. There are 239 subjects with the 3-year EFS information available (49 positives and 190 negatives). We randomly select 83 subjects (19 positives and 64 negatives, which are about one third of the total subjects) as the training set and the rest as the test set. The readers can find more details about the data sets in the original papers.

Following Dudoit *et al.* (2002) and Fan and Fan (2008), we standardized each sample to zero mean and unit variance. The classification results for ROAD, S-ROAD1, S-ROAD2, SCRDA, FAIR, NSC and NB are shown in Tables 8, 9 and 10. For the leukemia and lung cancer data, ROAD performs the best in terms of classification error. For the neuroblastoma data, NB performs best, however, it makes use of all 10,707 genes, which is not very desirable. In contrast, ROAD has a competitive performance in terms of classification error and it only selects 33 genes. Although SCRDA has a close performance, the number of selected variables varies a lot for the three data set (264, 2410, 1). Overall, ROAD is a robust classification tool for high-dimensional data.

Table 10. Classification error and number of selected genes by various methods of neuroblastoma data. Training and testing samples are of sizes 83 and 163, respectively.

	ROAD	S-ROAD1	S-ROAD2	SCRDA	FAIR	NSC	NB
Training Error	3	22	14	16	15	16	14
Testing Error	33	47	37	37	44	35	32
No. of selected genes	33	1	9	1	18	41	10707

7. Discussion

With a simple two-class gaussian model, we explored the bright side of using correlation structure for high dimensional classification. Targeting directly on the classification error, ROAD employs un-regularized pooled sample covariance matrix and sample mean difference vector without suffering from curse of dimensionality and noise accumulation. The sparsity of chosen features is evident in simulations and real data analysis; however, we have not discovered intuitively good conditions on $\mathbf{\Sigma}$ and $\boldsymbol{\mu}_d$, such that a certain desirable sparsity pattern of $\hat{\mathbf{w}}_c$ follows. We resolve a part of the problem by introducing screening-based variants of ROAD, but the precise control of the sparsity size is worth for further investigation. Furthermore, we can explore the conditions for the model selection consistency.

In this paper, we have restricted ourselves to the linear rules. They can be easily extended to nonlinear discriminants via transformations such as low order polynomials or spline basis functions. One may also use the popular "kernel tricks" in the machine learning community. See, for example, Hastie *et al.* (2009) for more details. After the features are transformed, we can hit the ROAD. One essential technical challenge of the current paper is rooted in a stochastic linear constraint. The precise role of this constraint has not been completely pinned down. In the following, a preliminary proposal is provided for extending ROAD to multi-class settings.

7.1. Extension to Multi-Class

In this section, we outline an extension of ROAD to multi-class classification problems. Suppose that there are K classes, and for $j = 1, \cdots, K$, the jth class has mean $\boldsymbol{\mu}_j$ and common covariance $\mathbf{\Sigma}$. Denote the overall mean of features by $\boldsymbol{\mu}_a = K^{-1} \sum_{j=1}^{K} \boldsymbol{\mu}_j$. Fisher's reduced rank approach to multi-class classification is a minimum distance classifier in some lower dimensional projection space. The first step is to find $s \leq K-1$ discriminant coordinates $(\mathbf{w}_1^*, \cdots, \mathbf{w}_s^*)$ that separate the population centroids $\{\boldsymbol{\mu}_j\}_{j=1}^K$ the most in the projected space $\mathcal{S} = \text{span}\{\mathbf{w}_1^*, \cdots, \mathbf{w}_s^*\}$. Then the population centroids $\boldsymbol{\mu}_j$'s and new observation $\mathbf{X}$ are both projected onto $\mathcal{S}$. The observation $\mathbf{X}$ will be assigned to the class whose projected centroid is closest to the projection of $\mathbf{X}$ onto $\mathcal{S}$. Note that it is usually not necessary to compute all $K-1$ discriminant coordinates whose span is that of all K population centroids; the process can stop as long as the projected population centroids are well spread out in $\mathcal{S}$.

We adopt the above procedure for multi-class classification. However, the large-p-small-n scenario demands regularization in selecting discriminant coordinates. Indeed, in the Fisher's proposal the first discriminant coordinate $\mathbf{w}_1^*$ is the solution of

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{B} \mathbf{w}}{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}, \tag{18}$$

where $\mathbf{B} = \mathbf{\Psi}^T \mathbf{\Psi}$, and the jth column of $\mathbf{\Psi}^T$ is $(\boldsymbol{\mu}_j - \boldsymbol{\mu}_a)$. Note that a multiple of $\mathbf{B}$ is the between-class variance matrix. The second discriminant coordinate $\mathbf{w}_2^*$ is the maximizer of $\mathbf{w}^T \mathbf{B} \mathbf{w} / (\mathbf{w}^T \mathbf{\Sigma} \mathbf{w})$ with constraint $\mathbf{w}_1^{*T} \mathbf{\Sigma} \mathbf{w} = 0$, and the subsequent discriminant coordinates are determined analogously.

Since solving (18) is the same as looking for the eigenvector of $\mathbf{\Sigma}^{-1/2} \mathbf{B} \mathbf{\Sigma}^{-1/2}$ corresponding to the largest eigenvalue, diverging spectrum and noise accumulation have to be considered when we work on the sample. To address these issues, we regularize $\mathbf{w}$ as in the binary case,

$$\min_{\|\mathbf{w}\|_1 \leq c, \mathbf{w}^T \mathbf{B} \mathbf{w} = 1} \mathbf{w}^T \mathbf{\Sigma} \mathbf{w}, \tag{19}$$

whose solution is the first regularized discriminant coordinate $\bar{\mathbf{w}}_1^*$. Here, equation (19) is related to the null space method in (Krzanowski *et al.*, 1995). The second regularized discriminant coordinate is obtained

by solving (19) with additional constraint $\bar{\mathbf{w}}_1^{*T}\mathbf{\Sigma}\mathbf{w} = 0$. Other regularized discriminant coordinates can be found similarly. With these s ($\leq K-1$) regularized discriminant coordinates, the classifier is now based on the minimum distance to the projected centroids in the s-dimensional space spanned by $\{\bar{\mathbf{w}}_j^*\}_{j=1}^s$.

The implementation and theoretical properties for multi-class ROAD are interesting topics for future research.

Acknowledgements

The authors thank the Editor, the Associate Editor and two referees, whose comments have greatly improved the scope and presentation of the paper. The financial support from NSF grant DMS-0704337 and NIH Grant R01-GM072611 is greatly acknowledged.

A. Proofs

A.1. Proof of Theorem 1

We now show first part of the theorem. Let $f_0(\mathbf{w}) = \mathbf{w}^T\boldsymbol{\mu}_d/(\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})^{1/2}$, $f_1(\mathbf{w}) = \mathbf{w}^T\hat{\boldsymbol{\mu}}_d/(\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})^{1/2}$, and $f_2(\mathbf{w}) = \mathbf{w}^T\hat{\boldsymbol{\mu}}_d/(\mathbf{w}^T\hat{\mathbf{\Sigma}}\mathbf{w})^{1/2}$. Then, it follows easily that

$$|f_0(\mathbf{w}_c) - f_2(\hat{\mathbf{w}}_c)| \leq \Lambda_1 + \Lambda_2,$$

where $\Lambda_1 = |f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)})|$ and $\Lambda_2 = |f_1(\mathbf{w}_c^{(1)}) - f_2(\hat{\mathbf{w}}_c)|$. We now bound both terms separately in the following two steps.

Step 1(bound Λ_1): For any $\mathbf{w}$, we have

$$\begin{aligned}
|f_0(\mathbf{w}) - f_1(\mathbf{w})| &\leq |\frac{\mathbf{w}^T\mu_d}{(\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})^{1/2}} - \frac{\mathbf{w}^T\hat{\boldsymbol{\mu}}_d}{(\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})^{1/2}}| \\
&\leq \frac{\|\mathbf{w}\|_1\|\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|_\infty}{\|\mathbf{w}\|_2\lambda_{\min}^{1/2}(\mathbf{\Sigma})} \\
&\leq \sqrt{\|\mathbf{w}\|_0}\frac{\|\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|_\infty}{\sigma_0} \\
&= \sqrt{\|\mathbf{w}\|_0}O_p(a_n). \qquad (20)
\end{aligned}$$

Since $\mathbf{w}_c^{(1)}$ maximizes $f_1(\cdot)$, it follows that

$$\begin{aligned}
f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)}) &= f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c) + [f_1(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)})] \\
&\leq f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c), \qquad (21)
\end{aligned}$$

and similarly noticing w_c maximizing $f_0(\cdot)$, we have

$$\begin{aligned}
f_1(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c) &= f_1(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c^{(1)}) + [f_0(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c)] \\
&\leq f_1(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c^{(1)}). \qquad (22)
\end{aligned}$$

Combining the results of (21) and (22) and using (20), we conclude that

$$\Lambda_1 = |f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)})| = O_p\left((s_c \vee s_c^{(1)})a_n\right).$$

By the Lipschitz property of Φ,

$$|\Phi(f_1(\mathbf{w}_c^{(1)})) - \Phi(f_0(\mathbf{w}_c))| = O_p\left((s_c \vee s_c^{(1)})a_n\right).$$

Step 2(bound Λ_2): Note that $\mathbf{w}_c^{(1)}$ and $\hat{\mathbf{w}}_c$ both are in the set $\{\mathbf{w} : \mathbf{w}^T\boldsymbol{\mu}_d = 1, \|\mathbf{w}\|_1 \le 1\}$. Therefore, by definition of minimizers, we have

$$\mathbf{w}_c^{(1)^T}\boldsymbol{\Sigma}\mathbf{w}_c^{(1)} - \hat{\mathbf{w}}_c^T\boldsymbol{\Sigma}\hat{\mathbf{w}}_c \le 0, \text{ and } \hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)^T}\hat{\boldsymbol{\Sigma}}\mathbf{w}_c^{(1)} \le 0.$$

Consequently,

$$\begin{aligned}
\mathbf{w}_c^{(1)^T}\boldsymbol{\Sigma}\mathbf{w}_c^{(1)} - \hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c &= [\mathbf{w}_c^{(1)^T}\boldsymbol{\Sigma}\mathbf{w}_c^{(1)} - \hat{\mathbf{w}}_c^T\boldsymbol{\Sigma}\hat{\mathbf{w}}_c] + \hat{\mathbf{w}}_c^T\boldsymbol{\Sigma}\hat{\mathbf{w}}_c - \hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c \\
&\le \hat{\mathbf{w}}_c^T(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\hat{\mathbf{w}}_c \\
&\le \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty \|\hat{\mathbf{w}}_c\|_1^2 \\
&\le c^2\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty \\
&= O_p(a_n c^2). \qquad (23)
\end{aligned}$$

By the same argument, we also have

$$\begin{aligned}
\hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)^T}\boldsymbol{\Sigma}\mathbf{w}_c^{(1)} &= [\hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)^T}\hat{\boldsymbol{\Sigma}}\mathbf{w}_c^{(1)}] + \mathbf{w}_c^{(1)^T}\hat{\boldsymbol{\Sigma}}\mathbf{w}_c^{(1)} - \mathbf{w}_c^{(1)^T}\boldsymbol{\Sigma}\mathbf{w}_c^{(1)} \\
&\le \mathbf{w}_c^{(1)^T}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{w}_c^{(1)} \\
&\le c^2\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty \\
&= O_p(a_n c^2). \qquad (24)
\end{aligned}$$

Combination of (23) and (24) leads to

$$|\hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)^T}\boldsymbol{\Sigma}\mathbf{w}_c^{(1)}| = O_p(a_n c^2).$$

Let $g(x) = \Phi(x^{-1/2})$. The function g is Lipschitz on $(0, \infty)$, as $g'(x)$ is bounded on $(0, \infty)$. Hence, $|\Phi(f_2(\hat{\mathbf{w}}_c)) - \Phi(f_0(\mathbf{w}_c^{(1)}))| = O_p(a_n c^2)$. Thus,

$$\begin{aligned}
|W_n(\hat{\delta}_{\mathbf{w}_c}, \boldsymbol{\theta}) - W(\delta_{\mathbf{w}_c}, \boldsymbol{\theta})| &\le |\Phi(f_2(\hat{\mathbf{w}}_c)) - \Phi(f_0(\mathbf{w}_c^{(1)}))| + |\Phi(f_1(\hat{\mathbf{w}}_c^{(1)})) - \Phi(f_0(\mathbf{w}_c))| \\
&= O_p\left((s_c \vee s_c^{(1)})a_n\right) + O_p(a_n c^2) \\
&= O_p(b_n).
\end{aligned}$$

We now prove the second result of the Theorem. Since $|\hat{\mathbf{w}}_c^T\boldsymbol{\Sigma}\hat{\mathbf{w}}_c - \hat{\mathbf{w}}_c^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}}_c| = O_p(a_n c^2)$, we have

$$|\Phi(f_1(\hat{\mathbf{w}}_c)) - \Phi(f_2(\hat{\mathbf{w}}_c))| = O_p(a_n c^2). \qquad (25)$$

By (20), (25), and the first part of the Theorem, we have

$$\begin{aligned}
&|W(\hat{\delta}_{\mathbf{w}_c}, \boldsymbol{\theta}) - W(\delta_{\mathbf{w}_c}, \boldsymbol{\theta})| \\
=&|\Phi(f_0(\hat{\mathbf{w}}_c)) - \Phi(f_0(\mathbf{w}_c))| \\
\le&|\Phi(f_0(\hat{\mathbf{w}}_c)) - \Phi(f_1(\hat{\mathbf{w}}_c))| + |\Phi(f_1(\hat{\mathbf{w}}_c)) - \Phi(f_2(\hat{\mathbf{w}}_c))| + |\Phi(f_2(\hat{\mathbf{w}}_c)) - \Phi(f_0(\mathbf{w}_c))| \\
=&O_p(\hat{s}_c a_n) + O_p(a_n c^2) + O_p(b_n) \\
=&O_p(d_n).
\end{aligned}$$

This completes the proof of Theorem.

A.2. Proof of Theorem 2

Let $\mathbf{w}^\lambda = \mathbf{w}_\infty + \boldsymbol{\gamma}^\lambda$. Then, from the definition of $\mathbf{w}^\lambda$, we have

$$\begin{aligned}\boldsymbol{\gamma}^\lambda &= \underset{\boldsymbol{\mu}_d^T\mathbf{w}_\infty+\boldsymbol{\mu}_d^T\boldsymbol{\gamma}=1}{\operatorname{argmin}} R(\mathbf{w}_\infty+\boldsymbol{\gamma}) + \lambda\|\mathbf{w}_\infty+\boldsymbol{\gamma}\|_1 \\ &= \underset{\boldsymbol{\mu}_d^T\boldsymbol{\gamma}=0}{\operatorname{argmin}} f(\boldsymbol{\gamma}),\end{aligned} \tag{26}$$

where $f(\boldsymbol{\gamma}) = R(\boldsymbol{\gamma}) + \lambda\Sigma_{k\in K^c}|\boldsymbol{\gamma}_k| + \lambda\Sigma_{k\in K}\left(|\mathbf{w}_\infty^k+\boldsymbol{\gamma}_k| - |\mathbf{w}_\infty^k|\right)$. In the last statement, we used the fact that

$$\mathbf{w}_\infty^T\boldsymbol{\Sigma}\boldsymbol{\gamma} = \boldsymbol{\mu}_d^T\boldsymbol{\gamma}/(\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d) = 0.$$

We write $\boldsymbol{\gamma}$ for $\boldsymbol{\gamma}^\lambda$ for short in what follows.

By (26), we have $f(\boldsymbol{\gamma}) \le f(\mathbf{0}) = 0$. This implies that

$$R(\boldsymbol{\gamma}) \quad\le\quad \lambda\Sigma_{k\in K}\left(|\mathbf{w}_\infty^k| - |\mathbf{w}_\infty^k+\boldsymbol{\gamma}_k|\right) \le \lambda\Sigma_{k\in K}|\boldsymbol{\gamma}_k| \le \lambda\sqrt{s}\|\boldsymbol{\gamma}\|_2.$$

On the other hand, $R(\boldsymbol{\gamma}) \ge \lambda_{\min}(\boldsymbol{\Sigma})\|\boldsymbol{\gamma}\|_2^2$. Bringing the upper and lower bound of $R(\boldsymbol{\gamma})$ together, we conclude that

$$\|\boldsymbol{\gamma}\|_2 \le \frac{\lambda\sqrt{s}}{\lambda_{\min}(\boldsymbol{\Sigma})}.$$

The proof is now complete.

A.3. Proof of Theorem 5

By the positive definiteness of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ are also positive definite. Let $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\mathbf{w}$, then the transformation $\boldsymbol{v} \mapsto \mathbf{w}$ is linear. Define

$$\boldsymbol{v}_c = \underset{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}\|_1\le c, \boldsymbol{v}^T\bar{\boldsymbol{\mu}}_d=1}{\operatorname{argmin}} \boldsymbol{v}^T\boldsymbol{v},$$

where $\bar{\boldsymbol{\mu}}_d = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_d$. It is enough to show that $\boldsymbol{v}_c$ is piecewise linear in c.

Let $\Omega_c = \{\boldsymbol{v} : \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}\|_1 \le c\}$ and $S = \{\boldsymbol{v} : \boldsymbol{v}^T\bar{\boldsymbol{\mu}}_d = 1\}$. When c is small, the solution set is $\emptyset$; when c is large, the constraint Ω_c is inactive. Denote by "a" the smallest "c" such that $\Omega_c \bigcap S \neq \emptyset$, and by "b" the smallest such that $\boldsymbol{v}_c$ are the same for all $c \ge b$. Hence we are interested in $c \in [a, b]$, when changes in c actually affects the solution.

Let P be the projection of the origin O onto the hyperplane S in the p dimensional space. Let

$$\mathcal{F}_c = \{S_{1,c}^0, \cdots, S_{j_0,c}^0; S_{1,c}^1, \cdots, S_{j_1,c}^1; \cdots; S_{1,c}^{p-1}, \cdots S_{j_{p-1},c}^{p-1}\},$$

where $S_{j,c}^i$ denotes an i-dimensional face of Ω_c, i.e., $S_{j,c}^0$ represents a vertex, $S_{j,c}^1$ an edge, and $S_{j,c}^{p-1}$ a facet. It is clear that $\mathcal{F}_c$ is a finite set.

Define a mapping $\varphi : [a, b] \to \mathbb{Z}\times\mathbb{Z}$, where $\varphi(c) = (i, j)$ such that i) $\boldsymbol{v}_c \in S_{j,c}^i$ and ii) i is minimal. By definition, this mapping is single valued.

For any $c_0 \in (a, b]$, denote $D_{c_0} = \{(i, j)|\forall\epsilon > 0, \exists c \in [c_0-\epsilon, c_0)$ s.t. $\varphi(c) = (i, j)\}$. The set D_{c_0} is non-empty because the collection $\{(i, j) \in \mathbb{Z}\times\mathbb{Z}|S_{j,c}^i \in \mathcal{F}_c\}$ is finite. Then the theorem follows from compactness of $[a, b]$ and Lemma 2, Remark 4 and Lemma 3.

LEMMA 1. *$\forall c_0 \in (a, b]$, $\exists\epsilon > 0$ such that $\forall(i, j) \in D_{c_0}$ and $\forall c \in (c_0-\epsilon, c_0)$, $P_{j,c}^i \in S_{j,c}^{io}\cap S$, where $P_{j,c}^i$ is the projection of P onto $S\cap\widetilde{S_{j,c}^i}$, and $\widetilde{S_{j,c}^i}$ denotes the i-dimensional affine space in which $S_{j,c}^i$ embeds, and $S_{j,c}^{io}$ is the interior of $S_{j,c}^i$, where the topology is the natural subspace topology restricted to $\widetilde{S_{j,c}^i}$.*

PROOF. Fix $c_0 \in (a, b]$. For any $(i, j) \in D_{c_0}$ and $\bar{\epsilon} > 0$, by the definition of D_{c_0}, there exists $c' \in [c_0 - \bar{\epsilon}, c_0)$ such that $\varphi(c') = (i, j)$. The minimality of i in the definition for φ implies that $\boldsymbol{v}_{c'} = P^i_{j,c'} \in S^{i\circ}_{j,c'}$, which in the interior of $S^i_{j,c'}$. Therefore, $P^i_{j,c'} \in S^{i\circ}_{j,c'} \cap S$. By arbitrariness of $\bar{\epsilon}$, $\exists(c_n) \nearrow c_0$ such that $P^i_{j,c_n} \in S^{i\circ}_{j,c_n} \cap S$ for all n.

It can also be shown that $\{c | P^i_{j,c} \in S^{i\circ}_{j,c} \cap S\}$ is connected: let $P^i_{j,c'_1} \in S^{i\circ}_{j,c'_1} \cap S$, $P^i_{j,c'_2} \in S^{i\circ}_{j,c'_2} \cap S$, $c'_1 < c'_2$. For any $c'_3 \in (c'_1, c'_2)$, P^i_{j,c'_3} is on the line segment with endpoints P^i_{j,c'_1} and P^i_{j,c'_2} because $\widetilde{S^i_{j,c}}$ are parallel affine subspace in $\mathbb{R}^p$. Let $S^i_{j,cone} := \cup_{c\geq 0} S^{i\circ}_{j,c}$, then it is a cone. Since $P^i_{j,c'_1} \in S^i_{j,cone}$ and $P^i_{j,c'_2} \in S^i_{j,cone}$, we have $P^i_{j,c'_3} \in S^i_{j,cone}$. Then, $P^i_{j,c'_3} \in S^i_{j,cone} \cap S \cap \widetilde{S^i_{j,c'_3}} = S^{i\circ}_{j,c'_3} \cap S$. Hence, $\exists \epsilon_{ij} > 0$ such that for all $c \in [c_0 - \epsilon_{ij}, c_0)$, $P^i_{j,c} \in S^{i\circ}_{j,c}$. Take $\epsilon = \min_{(i,j)\in D_{c_0}} \epsilon_{ij}$, the claim follows.

LEMMA 2. *$\forall c_0 \in (a, b]$, D_{c_0} is a singleton, and $\exists \epsilon' > 0$ such that $\boldsymbol{v}_c$ is linear in c on $(c_0 - \epsilon', c_0)$.*

PROOF. Fix $c_0 \in (a, b]$. We claim that for some $(i, j) \in D_{c_0}$, there exists positive $\epsilon' (\leq \epsilon$ that validates Lemma 1) such that for any $c \in (c_0 - \epsilon', c_0)$, $\boldsymbol{v}_c = P^i_{j,c}$. Assume that the claim is not correct, then pick any $(i, j) \in D_{c_0}$, there exists a sequence $\{c_k\}$ ($c_k \neq c_{k'}$ if $k \neq k'$) converging to c_0 from the left s.t. $\boldsymbol{v}_{c_k} \neq P^i_{j,c_k}$. Without loss of generality, take $\{c_k\} \subset (c_0 - \epsilon, c_0)$. Lemma 1 implies that $P^i_{j,c_k} \in S^{i\circ}_{j,c_k} \cap S$. If $\boldsymbol{v}_{c_k} \in S^i_{j,c_k}$, we would have $\boldsymbol{v}_{c_k} = P^i_{j,c_k}$. Hence $\boldsymbol{v}_{c_k} \notin S^i_{j,c_k}$. By finiteness of the index pairs in $\mathcal{F}_c$, there exists $(i', j') \neq (i, j)$ such that $\varphi(c) = (i', j')$ for $c \in \{c_{k_l}\}$, where $\{c_{k_l}\}$ is some subsequence of $\{c_k\}$. This implies $(i', j') \in D_{c_0}$, which together with Lemma 1 implies $\boldsymbol{v}_c = P^{i'}_{j',c}$ for $c \in \{c_{k_l}\}$. Therefore

$$\|P^{i'}_{j',c} - P\|_2 < \|P^i_{j,c} - P\|_2$$

for $c \in \{c_{k_l}\}$.

On the other hand, because $(i, j) \in D_{c_0}$, there exist infinitely many $c' \in (c_0 - \epsilon, c_0)$ such that $\|P^{i'}_{j',c'} - P\|_2 \geq \|P^i_{j,c'} - P\|_2$. Therefore,

$$g(c) = \|P - P^i_{j,c}\|_2^2 - \|P - P^{i'}_{j',c}\|_2^2$$

changes signs infinitely many times on $(c_0 - \epsilon, c_0)$. This leads to a contradiction because $P^i_{j,c}$ and $P^{i'}_{j',c}$ are both linear functions of c. Hence, the conclusion holds.

To show that D_{c_0} is a singleton, suppose it has two distinct elements (i, j) and (i', j'). We have shown that $\boldsymbol{v}_c = P^i_{j,c}$ and $\boldsymbol{v}_c = P^{i'}_{j',c}$ for all c in a left neighborhood of c_0 (not including c_0). Also we have $P^i_{j,c} \in S^{i\circ}_{j,c}$ and $P^{i'}_{j',c} \in S^{i'\circ}_{j',c}$ by Lemma 1. This can be true only when $S^{i\circ}_{j,c} \subset S^{i'\circ}_{j',c}$ (or vice versa), but then $i < i'$, contradicting with minimality in definition of D_{c_0}.

REMARK 4. *Similarly, $\forall c_0 \in [a, b)$, $\exists \epsilon' > 0$ such that $\boldsymbol{v}_c$ is linear in c on $(c_0, c_0 + \epsilon')$.*

LEMMA 3. *$\boldsymbol{v}_c$ is a continuous function of c on $[a, b]$.*

PROOF. The continuity follows from two parts i) and ii).

i) $\forall c_0 \in [a, b)$, $\exists \epsilon > 0$ such that $\boldsymbol{v}_c$ is continuous on $[c_0, c_0 + \epsilon)$. Indeed, let

$$h(c) = \min_{\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{v}\|_1 \leq c, \boldsymbol{v}^T \bar{\boldsymbol{\mu}}_d = 1} \boldsymbol{v}^T \boldsymbol{v}.$$

We know that the mapping $c \mapsto \boldsymbol{v}_c (= P^i_{j,c})$ is linear and hence continuous on $(c_0, c_0 + \epsilon)$ for some small $\epsilon > 0$. It only remains to show that the mapping is right continuous at c_0. Notice here $h(c) = \|P^i_{j,c}\|_2^2$ for $c \in (c_0, c_0 + \epsilon)$. Let $L = \lim_{c\downarrow c_0} P^i_{j,c}$. It is clear that $L \in S^i_{j,c_0}$. Because $L \in \Omega_{c_0} \cap S$, $h(c_0) \leq \|L\|_2^2$. This

inequality has to take the equal sign because $h(\cdot)$ is monotone decreasing, and $h(c) = \|P^i_{j,c}\|_2^2 \to \|L\|_2^2$ as c approaches c_0 from the right. Because $\boldsymbol{v}_{c_0}$ is unique, $\boldsymbol{v}_{c_0} = L = \lim_{c\downarrow c_0} P^i_{j,c} = \lim_{c\downarrow c_0} \boldsymbol{v}_c$.

ii) $\forall c_0 \in (a,b]$, $\exists \epsilon > 0$ such that $\boldsymbol{v}_c$ is continuous on $(c_0-\epsilon, c_0]$. Again, it remains to show that there is no jump at c_0. Let $(i_{c_0}, j_{c_0}) = \varphi(c_0)$. Clearly $P^{i_{c_0}}_{j_{c_0},c_0} \in S^{i_{c_0}\circ}_{j_{c_0},c_0}$. Introduce a notion of parallelism of affine subspaces in $\mathbb{R}^p$. We denote $\widetilde{S^{i_{c_0}}_{j_{c_0},c}} \parallel S$, if only by translation, $\widetilde{S^{i_{c_0}}_{j_{c_0},c}}$ becomes a subset of S (or vice versa in other situations); use the notation $\widetilde{S^{i_{c_0}}_{j_{c_0},c}} \nparallel S$ otherwise.

If $\widetilde{S^{i_{c_0}}_{j_{c_0},c}} \nparallel S$, for c in some left neighborhood of c_0, $P^{i_{c_0}}_{j_{c_0},c}$ exists and $P^{i_{c_0}}_{j_{c_0},c} \in S^{i_{c_0}\circ}_{j_{c_0},c}$. Note $P^{i_{c_0}}_{j_{c_0},c} \in \Omega_c \cap S$, and $\|P^{i_{c_0}}_{j_{c_0},c}\|_2 \to \|P^{i_{c_0}}_{j_{c_0},c_0}\|_2$ as c approaches c_0 from the left. Since $h(\cdot)$ is monotone decreasing, obviously $h(c) \to \|P^{i_{c_0}}_{j_{c_0},c_0}\|_2^2 = h(c_0)$. This shows the left continuity of h at c_0. Suppose $D_{c_0} = \{(i,j)\}$, then we know on a left neighborhood of c_0 (not including c_0), $\boldsymbol{v}_c = P^i_{j,c}$. Let $E = \lim_{c\uparrow c_0} P^i_{j,c}$, then $E \in \Omega_{c_0} \cap S$. Note that $\|P^{i_{c_0}}_{j_{c_0},c}\|_2 \geq \|P^i_{j,c}\|_2$ for all c in c_0's left neighborhood, so we have $\|P^{i_{c_0}}_{j_{c_0},c_0}\|_2 \geq \|E\|_2$. On the other hand, $\|P^{i_{c_0}}_{j_{c_0},c_0}\|_2 \leq \|E\|_2$ by the definition of $P^{i_{c_0}}_{j_{c_0},c_0}$. Also, consider the uniqueness of distance minimizing point in $\Omega_{c_0} \cap S$ to origin O, $E = P^{i_{c_0}}_{j_{c_0},c_0}$, and hence $\boldsymbol{v}_c$ has left continuity at c_0.

If $\widetilde{S^{i_{c_0}}_{j_{c_0},c}} \parallel S$, $\exists Q \in \Omega_{c_0-\epsilon/2} \cap S$ such that $Q \neq P^{i_{c_0}}_{j_{c_0},c_0}$. When c goes from $c_0 - \epsilon/2$ to c_0, there exists a point $Q_c \in \Omega_c \cap S$ moving on the line segment from Q to $P^{i_{c_0}}_{j_{c_0},c_0}$. Therefore, $h(\cdot)$ is left continuous at c_0. Replace $P^{i_{c_0}}_{j_{c_0},c}$ by Q_c in the previous paragraph, the left continuity of $\boldsymbol{v}_c$ at c_0 follows from the same argument.

References

Ackermann, M. and Strimmer, K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563–570.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006) Prediction by supervised principal components. *J. Amer. Statist. Assoc.*, **101**, 119–137.

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

Boulesteix, A.-L. (2004) PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol.*, **3**, Art. 33, 32 pp. (electronic).

Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge University Press.

Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, **5**, 232–253.

Domingos, P. and Pazzani, M. (1997) On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, **29**, 103–130.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97**, 77–87.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.

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

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1600.

Fan, J. and Lv, J. (2008) Sure independence screening for ultra-high dimensional feature space(with discussion). *J. R. Statist. Soc. B*, **70**, 849–911.

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101–148.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**, 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62**, 4963–4967.

Guo, Y., Hastie, T. and Tibshirani, R. (2005) Regularized discriminant analysis and its application in microarrays. *Biostatistics*, **1**, 1–18.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition).* Springer-Verlag Inc.

Huang, X., P. W. (2003) Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19**, 2072–2978.

Krzanowski, W., Jonathan, P., McCarthy, W. and Thomas, M. (1995) Discriminat analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, **44**, 101115.

Lewis, D. D. (1998) Naive (bayes) at forty: The independence assumption in information retrieval. 4–15. Springer Verlag.

Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**, 316–342. With discussion and a rejoinder by the author.

Nguyen, D. V. and Rocke, D. M. (2002) Tumor classification by partial least squares using microarray gene expression data . *Bioinformatics*, **18**, 39–50.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006) Customized oligonucleotide microarray gene expression based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, **24**, 5070–5078.

Rosset, S. and Zhu, J. (2007) Piecewise linear regularized solution paths. *Ann. Statist.*, **35**, 1012–1030.

Ruszczynski, A. (2006) *Nonlinear Optimization*. Princeton University Press.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011) Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, **39**, to appear.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

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

Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, **109**, 475–494.

Vapnik, V. N. (1995) *The nature of statistical learning theory*. New York: Springer-Verlag.

Wang, S. and Zhu, J. (2007) Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, **23**, 972–979.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zhao, D. S. and Li, Y. (2010) Principled sure independence screening for cox models with ultra-high-dimensional covariates. Manuscript.

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

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 768–768.

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

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