Biostatistics (2005), **1**, 1, *pp.* 1–18
Printed in Great Britain

Regularized Discriminant Analysis and Its Application in Microarrays

By YAQIAN GUO

Department of Statistics, Stanford University
Stanford, CA 94305.
e-mail: yaqiang@stanford.edu.

and TREVOR HASTIE

Department of Statistics, Stanford University
Stanford, CA 94305
e-mail: hastie@stanford.edu.

and ROBERT TIBSHIRANI

Department of Health Research and Policy, Stanford University
Stanford, CA 94305
e-mail: tibs@stanford.edu

Summary

In this paper, we introduce a modified version of linear discriminant analysis, called "shrunken centroids regularized discriminant analysis" (SCRDA). This method generalizes the idea of "nearest shrunken centroids" (NSC) [Tibshirani *et al.*, 2003] into the classical discriminant analysis. The SCRDA method is specially designed for classification problems in high dimension low sample size situations, for example, microarray data. Through both simulated data and real life data, it is shown that this method performs very well in multivariate classification problems, often outperforms the PAM method and can be as competitive as the SVM classifiers. It is also suitable for feature elimination purpose and can be used as gene selection method. The open source R package for SCRDA is available and will be added to the R libraries in the near future.

Keywords: Classification, Discriminant analysis (DA), Microarray, Prediction analysis of microarrays (PAM), Regularization, Shrunken centriods.

1. Introduction

Discriminant analysis (DA) is widely used in classification problems. The traditional way of doing discriminant analysis is introduced by R. Fisher, known as the linear discriminant analysis (LDA). For the convenience of later discussion, we first describe the general setup of this method so that we can follow the notations used here throughout this paper.

Suppose there are G different populations, each assumed to have a multivariate normal distribution with common covariance matrix Σ $(p \times p)$ and different mean vectors μ_g $(p \times 1)$, for

$g = 1, \ldots, G$. Now we have a sample of size n, $x_1, x_2, \ldots, x_n$, randomly chosen from these populations and for the time being, we assume the group label of each observation is known. To be more explicit, let $x_{1,1}, \ldots, x_{1,n_1}$ be observations from population 1, $x_{2,1}, \ldots, x_{2,n_2}$ from population 2, and so on. Thus $n = n_1 + n_2 + \ldots + n_G$. Under our assumptions, we have

$$x_{g,i} \sim MVN(\mu_g, \Sigma), \quad 1 \leq g \leq G, 1 \leq i \leq n_g.$$

The idea of LDA is to classify observation $x_{g,i}$ to a population $\tilde{g}$ which minimizes $(x_{g,i} - \mu_{\tilde{g}})^T \Sigma^{-1} (x_{g,i} - \mu_{\tilde{g}})$, i.e.,

$$x_{g,i} \in \text{population } \left(\widetilde{g} = \text{argmin}_{g'} (x_{g,i} - \mu_{g'})^T \Sigma^{-1} (x_{g,i} - \mu_{g'})\right).$$

Under the above multivariate normal assumptions, this is equivalent to finding the population that maximizes the likelihood of the observation. More often, people have some prior knowledge as to the proportion of each population. For example, let π_g be the proportion of population g such that $\pi_1 + \ldots + \pi_G = 1$. Then, instead of maximizing the likelihood, we maximize the posterior probability the observation belongs to a particular group, i.e.,

$$x_{g,i} \in \text{population } \left(\widetilde{g} = \text{argmin}_{g'} \left[\frac{1}{2}(x_{g,i} - \mu_{g'})^T \Sigma^{-1} (x_{g,i} - \mu_{g'}) - \log \pi_{g'}\right]\right)$$

The linearity of this discriminant analysis method comes from the assumption of common covariance matrix, which simplifies the above criterion as

$$x_{g,i} \in \text{population } \left(\widetilde{g} = \text{argmax}_{g'} d_{g'}(x_{g,i})\right) \tag{1}$$

where

$$d_g(x) = x^T \Sigma^{-1} \mu_g - \frac{1}{2} \mu_g^T \Sigma^{-1} \mu_g + \log \pi_g$$

is the so-called discriminant function.

In reality, both μ_g and Σ are never known and therefore need to be estimated from the sample. Almost always, people take the maximum likelihood estimates for these parameters,

$$\widehat{\mu}_g = \bar{x}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{g,i}, \quad \widehat{\Sigma} = \frac{1}{n} (X - \bar{X})(X - \bar{X})^T,$$

where X is a $p \times n$ matrix with columns corresponding to the observations and $\bar{X}$ is a matrix of the same dimensions with each column corresponding to the sample mean vector of the population that column belongs to. Therefore a more practical form, the sample version discriminant function is usually used in the linear discriminant analysis,

$$\widehat{d}_g(x) = x^T \widehat{\Sigma}^{-1} \bar{x}_g - \frac{1}{2} \bar{x}_g^T \widehat{\Sigma}^{-1} \bar{x}_g + \log \pi_g. \tag{2}$$

When the assumption of common covariance matrix is not satisfied, people use an individual covariance matrix for each group and this leads to the so-called quadratic discriminant analysis (QDA) as the discriminating boundaries are quadratic curves. There is also an intermediate method between LDA and QDA, which is a regularized version of discriminant analysis (RDA) proposed by Friedman [1989]. However, the regularization used in that method is different from the one we will propose Here. A detailed source about the LDA, QDA and Friedman's

RDA methods can be found in the book by Hastie *et al.* [2001]. As we can see, the concept of discriminant analysis certainly embraces a broader scope. But in this paper, our main focus will be solely put on the LDA part and henceforth the term "discriminant analysis" will stand for the meaning of LDA unless otherwise emphasized.

This paper is arranged as follows. In Section 2, we will first discuss in details our version of regularization in discriminant analysis, its statistical properties and some computational issues (Section 2.1). Then we will introduce the SCRDA method based on this regularization (Section 2.2). In Section 3, we compare our SCRDA method against other classification approaches through several publicly available real life microarray data sets. We also discuss an important issue about how to choose the optimal parameter pairs (α, Δ) for our methods (Section 3.4). Section 4 is devoted to a simulation study, where we generate data sets under different scenarios to evaluate the performance of our SCRDA method. In Section 5, we briefly discuss the feature selection property of SCRDA method. Section 6 is the discussion.

2. Shrunken Centroids RDA

2.1. Regularization in discriminant analysis

LDA is straightforward in the cases where the number of observations is greater than the dimensionality of each observation, i.e., $n > p$. In addition to being easy to apply, it also has nice properties, like robustness to deviations from model assumptions. However, it becomes a serious challenge to use this method in the microarray analysis settings, where $p >> n$ is always the case. There are two major concerns here. First, the sample covariance matrix estimate is singular and cannot be inverted. Although we may use the generalized inverse instead, the estimate will be very unstable due to lack of observations. Actually, the performance of LDA in high dimensional situation is far from optimal [Dipillo, 1976, 1977]. Second, high dimensionality makes direct matrix operation formidable, hence hindering the applicability of this method. Therefore we will make some adaptations of the original LDA to overcome these problems. First, to resolve the singularity problem, instead of using $\widehat{\Sigma}$ directly, we use

$$\widetilde{\Sigma} = \alpha\widehat{\Sigma} + (1-\alpha)I_p \tag{3}$$

for some α, $0 \leq \alpha \leq 1$. Some other forms of regularization on $\widehat{\Sigma}$ can be

$$\widetilde{\Sigma} = \lambda\widehat{\Sigma} + I_p \tag{4}$$

or

$$\widetilde{\Sigma} = \widehat{\Sigma} + \lambda I_p \tag{5}$$

for some λ, $\lambda \in [0, \infty)$. It is easy to see that if we ignore the prior constant, the three forms of regularization are equivalent in terms of discriminant function. In this paper, the form (3) is our main focus and has been used in all of our computational results. But we may use all these three forms interchangeably at our own convenience when discussing some theoretical results without making much distinction between them.

The formulations above are not something entirely new and actually have been frequently seen in situations, such as the ridge regression [Hoerl and Kennard, 1970], where the correlation between predictors is high. By introducing a slightly biased covariance estimate, not only do we resolve the singularity problem, we also stabilize the sample covariance estimate. For example, the discriminant function (7) below is the main formula that we will be using in this paper. As

we can see (Figure 1 and 2), using regularization (3) both stabilizes the variance and reduces the bias of the discriminant function. And as a result, the prediction accuracy is improved. In Figure 1, the independent covariance structure is used to generate the data. From the plot we can see that the optimal regularization parameter α does tend to 0. On the other hand, in Figure 2, an auto-regressive covariance structure is used and the optimal α now lies in between 0 and 1.

Another more sensible version of regularization probably is to modify the sample correlation matrix $\widehat{R} = \widehat{D}^{-1/2}\widehat{\Sigma}\widehat{D}^{-1/2}$ in the same way,

$$\widetilde{R} = \alpha\widehat{R} + (1-\alpha)I_p, \tag{6}$$

where $\widehat{D}$ is the diagonal matrix taking the diagonal elements of $\widehat{\Sigma}$. Then we compute the regularized sample covariance matrix by $\widetilde{\Sigma} = \widehat{D}^{1/2}\widetilde{R}\widehat{D}^{1/2}$. In this paper, we will consider both cases and their performance will be compared. Now, having introduced the regularized covariance matrix, we can define the corresponding regularized discriminant function as,

$$\widetilde{d}_g(x) = x^T\widetilde{\Sigma}^{-1}\bar{x}_g - \frac{1}{2}\bar{x}_g^T\widetilde{\Sigma}^{-1}\bar{x}_g + \log\pi_g, \tag{7}$$

where the $\widetilde{\Sigma}$ can be from either (3) or (6).

Our next goal is to facilitate the computation of this new discriminant score. We have addressed the issue that direct matrix manipulation is impractical in microarray settings. But if we employ the singular value decomposition (SVD) trick to compute the matrix inversion, we can get around this trouble. This enables a very efficient way of computing the discriminant function and reduces the computation complexity from the order of $O(p^3)$ to $O(pn^2)$, which will be a significant saving when $p >> n$. For more details about the algorithm, please refer to Hastie *et al.* [2001].

2.2. Shrunken centroids RDA (SCRDA)

In this section, we will define our new method "shrunken centroids RDA" based on the regularized discriminant function (7) in the previous section. The idea of this method is similar to the "nearest shrunken centroids" (NSC) method [Tibshirani *et al.*, 2003], which we will describe briefly first. In microarray analysis, a widely accepted assumption is that most genes do not have differential expression level among different sample classes. In reality, the differences we observe are mostly due to random fluctuation. The NSC method removes the noisy information from such fluctuation by setting a soft threshold. This will effectively eliminate most non-contributing genes and leave only those truly significant ones for further analysis. In the NSC method, the group centroids of each gene are shrunken individually. This is based on the assumption that genes are independent of each other, which however, for most of the time is not totally valid. Notice that after shrinking the group centroids of a particular gene g, they compute the following gene-specific score for an observation x^*,

$$d_{g,k}(x_g^*) = \frac{(x_g^* - \bar{x}'_{g,k})^2}{2s_g^2} = \frac{(x_g^*)^2}{2s_g^2} - \frac{x_g^*\bar{x}'_{g,k}}{s_g^2} + \frac{(\bar{x}'_{g,k})^2}{2s_g^2}, \tag{8}$$

where x_g^* is the g-th component of the $p \times 1$ vector x^*, $\bar{x}'_{g,k}$ is the shrunken centroid of group k for gene g and s_g is the pooled standard deviation of gene g. Then x^* is classified to group k if k minimizes the sum of the scores over all genes (If prior information is available, a term $\log\pi_k$

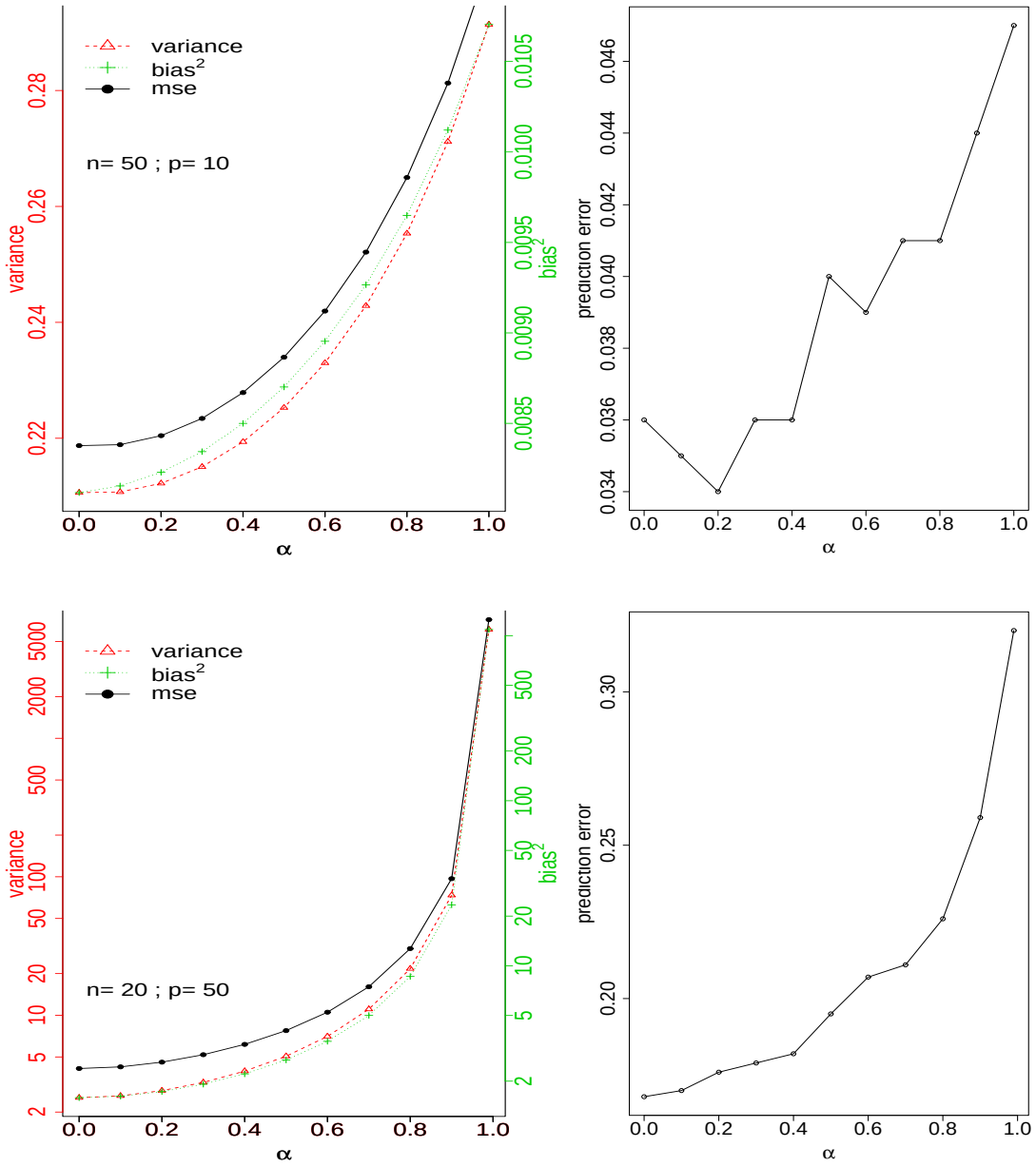

Fig. 1. The two plots on the left show the variability and bias of the discriminant function (7) as a function of the regularization parameter α in (3) for different sample sizes and dimensions. The two plots on the right show the prediction error of the discriminant function for the corresponding conditions on the left. The data points are generated from a p-dimensional multivariate normal distribution with an independent covariance structure.

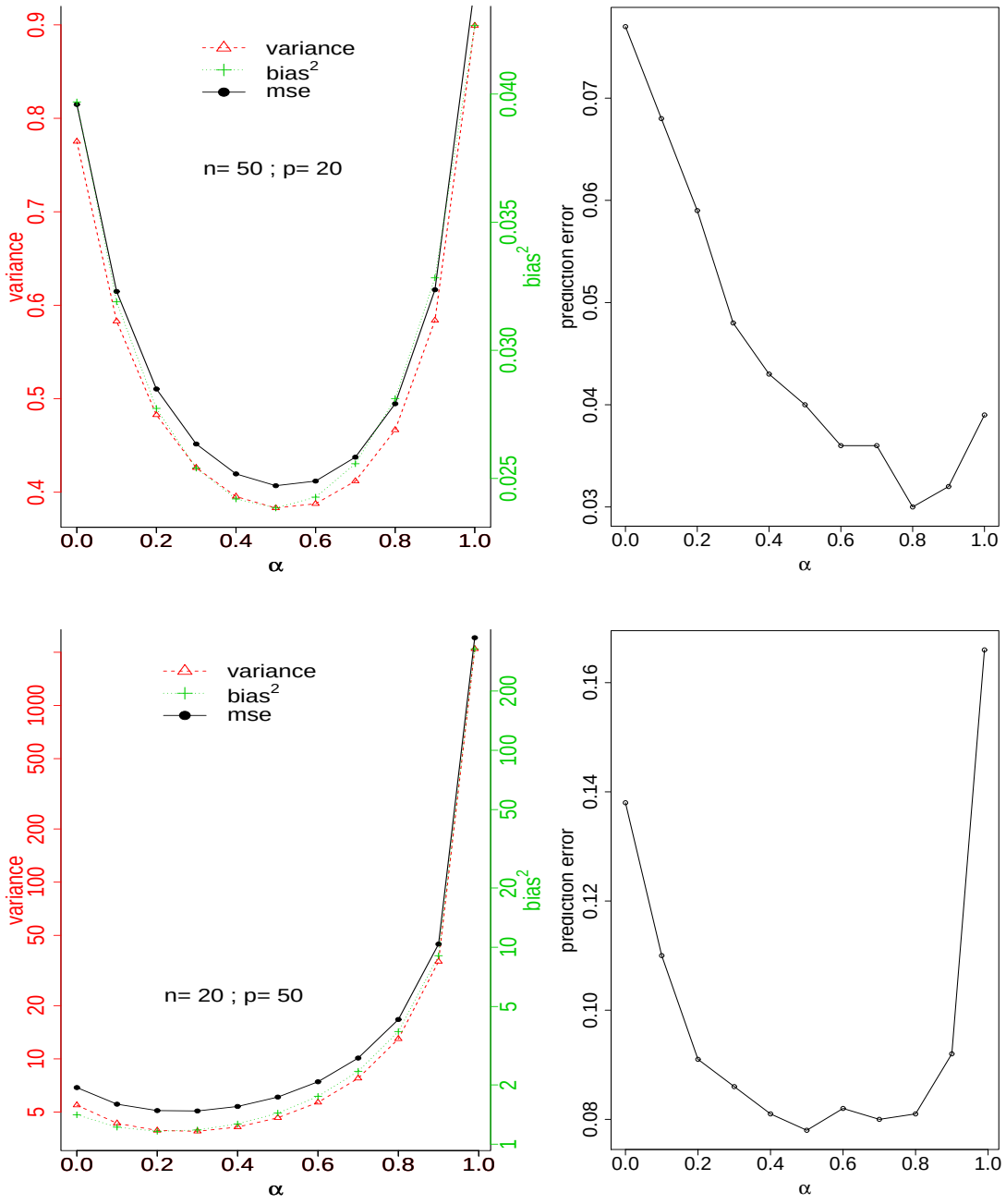

Fig. 2. The data are generated from a p-dimensional multivariate normal distribution with an auto-regressive correlation matrix. The auto-correlation is $\rho = 0{\cdot}6$

should be included.), i.e.,

$$x^* \in \text{group}\left(k = \text{argmin}_{k'} \sum_{g=1}^{p} d_{g,k'}(x_g^*) - \log \pi_{k'}\right)$$

which is also equivalent to

$$x^* \in \text{group}\left(k = \text{argmin}_{k'} (x^* - \bar{x}'_{k'})^T \widehat{D}^{-1} (x^* - \bar{x}'_{k'}) - \log \pi_{k'}\right),$$

given $\widehat{D} = \text{diag}(s_1^2, \ldots, s_p^2)$. This is similar to the discriminant function (7) except that we replace $\widetilde{\Sigma}$ with the diagonal matrix $\widehat{D}$ and the centroid vector $\bar{x}_g$ with the shrunken centroid vector $\bar{x}'_{k'}$. Therefore, a direct modification in the regularized discriminant function (7) to incorporate the idea of the NSC method is to shrink the centroids in (7) before calculating the discriminant score, i.e.,

$$\bar{x}' = \text{sgn}(\bar{x})(|\bar{x}| - \Delta)_+ . \tag{9}$$

However, in addition to shrinking the centroids directly, there are also two other possibilities. One is to shrink $\bar{x}^* = \widetilde{\Sigma}^{-1}\bar{x}$, i.e.,

$$\bar{x}^{*\prime} = \text{sgn}(\bar{x}^*)\left(|\bar{x}^*| - \Delta\right)_+ , \tag{10}$$

and the other is to shrink $\bar{x}_* = \widetilde{\Sigma}^{-1/2}\bar{x}$, i.e.,

$$\bar{x}'_* = \text{sgn}(\bar{x}_*)\left(|\bar{x}_*| - \Delta\right)_+ . \tag{11}$$

Although, it has been shown in our numerical analysis that all three shrinking methods have very good classification performance, only (10) will be the main focus of this paper as it also has the feature elimination property, which we will discuss later. Hence we will refer to the discriminant analysis resulted from (10) as SCRDA without differentiating whether $\widetilde{\Sigma}$ comes from (3) or (6). We will say more specifically which method is actually used when such distinction is necessary.

3. Comparison Based on Real Microarray Data

In this section, we first compare our new method with the penalized logistic regression and SVM methods (via univariate ranking and recursive feature elimination) proposed by Zhu and Hastie [2004] using the *Tamayo* and *Golub* data sets (Section 3.1 and 3.2). PAM, as the sibling method of SCRDA, is also included for comparison. Then, we will do an extended comparison of the performance of SCRDA, PAM and SVM based on 7 other public microarray data sets.

3.1. Tamayo data

The *Tamayo* data set [Ramaswamy *et al.*, 2001; Zhu and Hastie, 2004] is divided into a training subset, which contains 144 samples and a test subset of 54 samples. They consist of totally 14 different types of cancers and the number of genes in each array is 16063. Since there are two tuning parameters in the SCRDA method, i.e., the regularization parameter α and the shrinkage parameter Δ, we choose the optimal pairs (α, Δ) for $\alpha \in [0, 1)$ and $\Delta \in [0, \infty)$ using cross-validation on the training samples. And then we calculate the test error based on the tuning

parameter pairs we chose and compare it with the results from Zhu and Hastie [2004]. The result is summarized in Table 1. Based on how the covariance matrix is regularized in (10), two different forms of SCRDA are considered. In the table, we use "SCRDA" to denote the one from (3) and "SCRDA^r" for the case otherwise. We can see that SCRDA clearly dominates PAM and SCRDA^r. It even slightly outperforms the last 4 methods in the table. Meanwhile it also does a fairly good job on selecting informative gene subset.

Table 1. *Tamayo Data. The last four rows are excerpted from Zhu and Hastie [2004] for comparison. The SCRDA and SCRDA^r methods correspond to the situations where $\widetilde{\Sigma}$ in (7) comes from (3) and (6) respectively.*

Methods	8-fold CV Error	Ave. TE[b]	# of genes selected	Min. TE^{1}[c]	Min. TE^{2}[d]
SCRDA	24/144	8/54	1450	8/54	7/54
Hard[a] SCRDA	21/144	12/54	1317	12/54	9/54
SCRDA^r	27/144	15/54	16063	15/54	12/54
Hard SCRDA^r	28/144	13/54	16063	13/54	12/54
(Hard SCRDA^r)	*30/144*	*17/54*	*3285*	*17/54*	*12/54*
PAM	54/144	19/54	1596	NA	19/54
SVM UR	19/144	12/54	617	NA	8/54
PLR UR	20/144	10/54	617	NA	7/54
SVM RFE	21/144	11/54	1950	NA	11/54
PLR RFE	20/144	11/54	360	NA	7/54

[a]Hard SCRDA means the hard thresholding instead of the soft one is used.

[b]For SCRDA, "Ave TE" is calculated as the average of the test errors based on the optimal pairs. For the last 5 methods, "Ave TE" just means test error.

[c]"Min. TE^{1}" is the minimal test error one can get using the optimal (α, Δ) pairs; "Min. TE^{2}" is the minimal test error one can get over the whole parameter space.

[d] A method in parentheses means for that method, if we would like to sacrifice a little cross-validation error, then the number of genes selected can be greatly reduced than the row right above it.

3.2. Golub data

The *Golub* data [Golub *et al.*, 1999; Zhu and Hastie, 2004] consists of 38 training samples and 34 test samples from two cancer classes. The number of genes on each array is 7129. As there are only two groups to predict, this data set is much easier to analyze than the *Tamayo* data. The classification performance is generally impressive for most methods such that the difference among them is almost negligible. The result is summarized in Table 2.

3.3. Other real data sets

In this section, we further investigate the classification performance of the SCRDA method. We are particularly interested in how it compares to PAM and SVM. But we don't do the feature

Table 2. *Golub data. This is similar to Table 1.*

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	0/38	1/34	46	1/34	1/34
Hard SCRDA	2/38	3/34	123	3/34	1/34
SCRDA^r	3/38	6/34	1234	6/34	1/34
Hard SCRDA^r	1/38	4/34	92	4/34	1/34
PAM	2/38	2/34	149	1/34	1/34
SVM UR	2/38	3/34	22	NA	NA
PLR UR	2/38	3/34	17	NA	NA
SVM RFE	2/38	5/34	99	NA	NA
PLR RFE	2/38	1/34	26	NA	NA

selection for SVM as in Zhu and Hastie [2004]. Instead, we just focus on its classification performance on these data sets. The first six data sets we use here are all from a cancer research paper by Dettling [2004] and we call them `brain`, `colon`, `leukemia`, `prostate`, `SRBCT` and `lymphoma` following the naming there. The detailed information about these data sets is summarized in Table 3. Also following the rule in that paper, We divide each of these data sets into the training and test subsets with a ratio 2:1. Notice that the `leukemia` data is actually just the *Golub* data. As the training to test ratio is different there (about 1:1), we still include it here. The last data set we use, called `brown`, is also a cancer data set. Similar to the *Tamayo* data above, it contains large number of samples ($n = 348$) and classes ($G = 15$, 14 cancer types and 1 normal type). The number of genes on the arrays is however much smaller ($p = 4718$) than the *Tamayo* data ($p = 16013$).

Table 3. *Summary of seven cancer microarray data sets.*

Name	# of Samples (n)	# of Genes (p)	# of Classes (G)
Brain	42	5597	5
Colon	62	2000	2
Leukemia	72	3571	2
Prostate	102	6033	2
SRBCT	63	2308	4
Lymphoma	62	4026	3
Brown	348	4718	15

Based on the criteria like sample size, number of classes and genes, the first six data sets are considered relatively "easier" and as shown in the tables 4 - 9, the classification performance of the SCRDA method is comparable with the PAM and SVM methods. The differences among all the methods in these tables are rather slim to the negligible extent. The true advantage of SCRDA method comes in the last data set `brown` (Table 10), where both the number of samples and the number of classes far exceed the complexity of the previous six ones. As we can see, if we don't impose any restriction on the number of genes remained, then the SCRDA method

uniformly dominates both PAM and SVM by a large margin. On the other hand, even if we would like to keep the remaining gene set small, at the expense of some increase in the cross-validation error and the test error, the SCRDA method is still much better than PAM. Notice that the SVM method seems to over-train the training data and therefore the error rate on the test data is much higher than the cross-validation error. We also observed a similar phenomenon for SVM in the *Tamayo* data. But this is not the case for the SCRDA method and the PAM method, for which the cross-validation error rate and the test error rate are always about the same, implying more reliability in the analysis result from these methods.

Table 4. *Brain data* ($n = 42,\ p = 5597,\ G = 5$).

Methods	3-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	5/28	5/14	183	4/14	2/14
Hard SCRDA	6/28	4/14	106	4/14	2/14
SCRDA^r	7/28	3/14	794	2/14	0/14
Hard SCRDA^r	8/28	3/14	385	2/14	0/14
PAM	5/28	4/14	60	4/14	2/14
SVM	5/28	3/14	NA	3/14	3/14

Table 5. *Colon data* ($n = 62,\ p = 2000,\ G = 2$).

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	3/40	3/22	143	3/22	3/22
Hard SCRDA	3/40	5/22	44	4/22	2/22
SCRDA^r	3/40	4/22	11	4/22	3/22
Hard SCRDA^r	3/40	5/22	4	4/22	2/22
PAM	3/40	3/22	19	3/22	3/22
SVM	5/40	6/22	NA	6/22	6/22

Table 6. *Leukemia data* ($n = 72,\ p = 3571,\ G = 2$).

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	1/49	2/23	7	2/23	0/23
Hard SCRDA	1/49	2/23	9	2/23	0/23
SCRDA^r	2/49	1/23	39	1/23	0/23
Hard SCRDA^r	1/49	2/23	30	2/23	0/23
PAM	2/49	2/23	8	2/23	0/23
SVM	1/49	0/23	NA	0/23	0/23

Table 7. *Prostate data ($n = 102$, $p = 6033$, $G = 2$).*

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	3/68	2/34	40	2/34	1/34
Hard SCRDA	2/68	2/34	34	2/34	1/34
SCRDA^r	4/68	3/34	487	3/34	1/34
Hard SCRDA^r	5/68	2/34	27	1/34	1/34
PAM	6/68	4/34	4	4/34	4/34
SVM	7/68	3/34	NA	3/34	3/34

Table 8. *SRBCT data ($n = 63$, $p = 2308$, $G = 4$).*

Methods	5-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	2/42	1/21	38	1/21	0/21
Hard SCRDA	2/42	0/21	44	0/21	0/21
SCRDA^r	2/42	0/21	458	0/21	0/21
Hard SCRDA^r	2/42	1/21	386	1/21	0/21
PAM	1/42	0/21	8	0/21	0/21
SVM	1/42	0/21	NA	0/21	0/21

Table 9. *Lymphoma data ($n = 62$, $p = 4026$, $G = 3$).*

Methods	6-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	3/41	1/21	76	0/21	0/21
Hard SCRDA	1/41	0/21	34	0/21	0/21
SCRDA^r	3/41	3/21	88	2/21	0/21
Hard SCRDA^r	1/41	0/21	60	0/21	0/21
PAM	3/41	1/21	1558	1/21	0/21
SVM	0/41	0/21	NA	0/21	0/21

We summarize all the results in Table 11. As we can see, the SCRDA method is better than PAM and has a comparable performance with SVM.

3.4. Choosing the optimal tuning parameters (α, Δ) in SCRDA

Although the idea of choosing the optimal tuning parameter pairs (α, Δ) based on cross-validation is easy to understand, in practice, how to choose the correct ones can be confusing. It is usually subject to one's experience and preference. The main problem is that there are many possible tuning parameter pairs giving the same cross-validation error rate. Yet, the test error

Table 10. *Brown data ($n = 349$, $p = 4718$, $G = 15$).*

Methods	4-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	24/262	6/87	4718	5/87	5/87
(SCRDA)	*48/262*	*16/87*	*2714*	*14/87*	*5/87*
Hard SCRDA	23/262	6/87	4718	5/87	5/87
(Hard SCRDA)	*45/262*	*18/87*	*2187*	*18/87*	*5/87*
SCRDA^r	22/262	5/87	4718	5/87	5/87
(SCRDA^r)	*48/262*	*26/87*	*2137*	*26/87*	*5/87*
Hard SCRDA^r	21/262	6/87	4718	6/87	4/87
(Hard SCRDA^r)	*48/262*	*23/87*	*2423*	*22/87*	*4/87*
PAM	51/262	17/87	4718	17/87	17/87
(PAM)	*77/262*	*24/87*	*1970*	*24/87*	*17/87*
SVM	25/262	19/87	NA	19/87	19/87

As in Table 1 for *Tamayo* data set, a method in parentheses means we sacrifice the error rate to reduce the number of genes selected as compared to the line that is right above it.

Table 11. *Average test error rate and average ranking of different methods*

	Brain	Colon	Leuk-emia	Pro-state	SRBCT	Lym-phoma	Brown	average ranking
SCRDA	35.7%	13.6%	8.7%	5.9%	4.8%	4.8%	6.9%	5.43
Hard SCRDA	28.6%	22.7%	8.7%	5.9%	0%	0%	6.9%	4.57
SCRDA^r-1	21.4%	18.2%	4.3%	8.8%	0%	14.3%	5.7%	4.42
Hard SCRDA^r	21.4%	22.7%	8.7%	5.9%	4.8%	0%	6.9%	4.71
PAM	28.6%	13.6%	8.7%	11.8%	0%	4.8%	19.5%	6.07
SVM	21.4%	27.3%	0%	8.8%	0%	0%	21.8%	5.14

The average ranking in the last column is the based on the ranks of all the methods in the table for each data set. It only provides a rough sense about how well each method performs rather than an accurate evaluation index.

rate based on them may vary differently. Therefore, how to choose the best parameter pairs is an essential issue in evaluating the performance of the SCRDA method. Therefore, we suggest several rules for choosing parameters based on our experience.

First let's take a look at how the classification errors and the number of genes remained are distributed across the varying scopes of the tuning parameters (α, Δ). Figure 3 shows the cross-validation error and test error given by SCRDA for the *Tamayo* data. α is chosen to lie between 0 and 1 while Δ between 0 and 3. Figure 4 shows the number of genes remained for the same range of the parameters.

The most significant pattern we can observe in Figure 4 is the decreasing gradient approx-

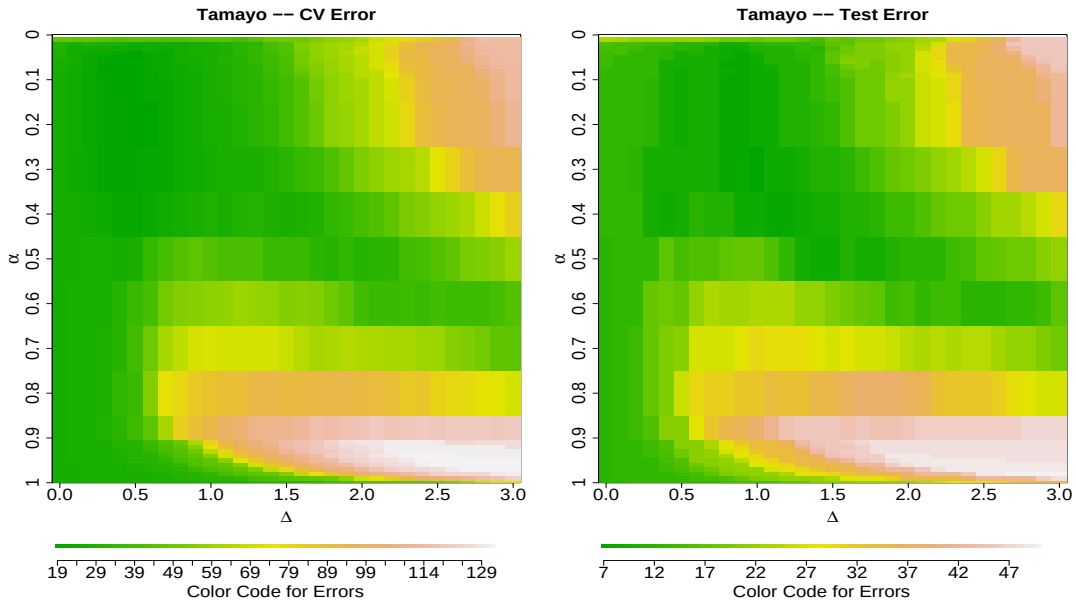

Fig. 3. Heat map of the distribution of the cross-validation error (left) and test error (right) of SCRDA for the *Tamayo* data. In each plot, the x- and y- axes correspond to Δ and α respectively. The color bar at the bottom shows the magnitude of the errors.

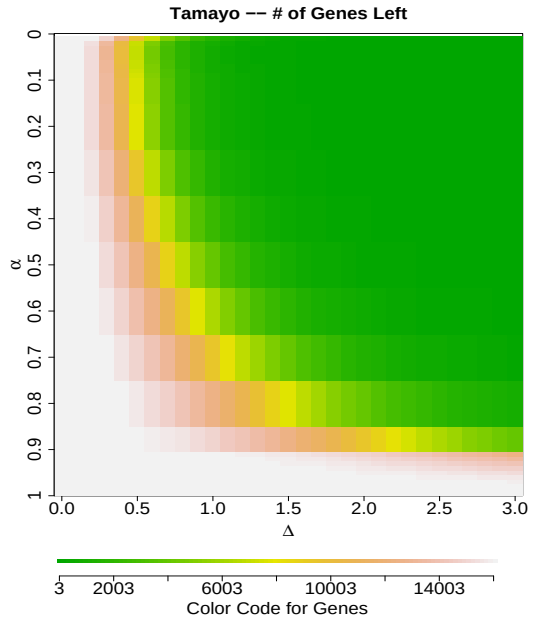

Fig. 4. Heat map of the distribution of the number of genes by SCRDA remained for the *Tamayo* data.

imately along the 45 degree diagonal line, i.e., the number of genes remained decreases as Δ increases or α decreases, which makes sense by intuition and has been consistently observed for all the real data and simulation data we have worked on. On the other hand, the distribution of the classification errors (both CV and test) as in Figure 3 doesn't have such a clearly consistent pattern. They may change dramatically from one data set to another. Further, as it is not possible to establish a unified correspondence between the distribution map of the classification error and the number of genes remained, we need to consider two distribution maps at the same time to achieve improved classification performance using a reasonably reduced subset of genes. We hence suggest the following "Min-Min" rule to identify the optimal parameter pair (α, Δ),

1. First, find all the pairs (α, Δ) that correspond to the minimal cross-validation error from training samples.

2. Select the pair or (pairs) that use the minimal number of genes.

If there is more than one optimal pair, it is recommended to calculate the averaged test error based on all the pairs chosen as we did in this paper.

4. Simulation Study

In this part, we investigate the performance of the SCRDA method under 3 different simulation situations. Particularly, we are interested in comparing its performance with PAM.

4.1. Two-group independence structure

The first setup is as follows. There are two classes. For each class, we generate $n = 100$ independent samples and each of them has $p = 10,000$ independent variables. For the first class, all 10,000 variables are generated from a standard normal distribution $N(0, 1)$. For the second class, the first 100 variables are from $N(1/2, 1)$ while the rest 9,900 are from a standard normal distribution $N(0, 1)$. We also generate 500 test samples from each class.

This is not a situation where we see much advantage of the SCRDA method over PAM (Table 12). In this setup, all methods produce basically the same result. PAM seems to be even slightly better than the SCRDA method. However, it is hard to declare PAM as a clear winner in this case. There are two reasons. First, the number of classes is only two, the simplest case in all classification problems. As people are aware of, it is much easier for most classification methods to work well in the 2-group classification problems and often it is hard to really observe the advantages of one over the another. Second, the data is generated from the covariance structure of identity matrix. This suits exactly the assumption in PAM to make it work well. As we can see in the next two examples, when the number of classes increases, especially when data is more correlated, the SCRDA method will start to show true advantage over PAM.

4.2. Multi-group independence structure

The second simulation setup is slightly more complicated. We generate a multiple group $(G = 14)$ classification scenario. Again there are $p = 10000$ genes and $n = 200$ training samples. The group label of each sample is generated with equal probabilities from 14 group labels. If a sample is from group g, then the $p = 10000$ genes jointly have a multivariate normal distribution $MVN(\mu_g, I_p)$, where μ_g is a p-vector with $l = 20$ positions equal to $1/2$ and all other positions

Table 12. *Setup I — two groups with independent structure.*

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	8/200	30/1000	240	29/1000	29/1000
Hard SCRDA	11/200	35/1000	186	33/1000	24/1000
SCRDA^r	11/200	33/1000	229	32/1000	29/1000
Hard SCRDA^r	13/200	27/1000	110	27/1000	27/1000
PAM	10/200	24/1000	209	24/1000	22/1000

equal to 0. The $l = 20$ positions for group g is randomly chosen from the $p = 10000$ positions. Using the same method, we also generate $m = 1000$ test samples. This time, we start to observe big differences among these methods (Table 13). Particularly, the SCRDA method starts to outperform PAM.

Table 13. *Setup II — multiple groups with independent structure.*

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	31/200	154/1000	785	154/1000	150/1000
Hard SCRDA	48/200	259/1000	169	259/1000	191/1000
SCRDA^r	41/200	207/1000	1398	207/1000	197/1000
Hard SCRDA^r	67/200	323/1000	140	323/1000	269/1000
PAM	36/200	179/1000	769	179/1000	166/1000

4.3. Two-group dependence structure

In this case, we produce a scenario that more resembles the real microarray data. The simulation structure is as follows. We again consider a two-group classification problem as in setup I. We generate 200 training samples and 1000 test samples, with half from each class. There are $p = 10000$ genes. We divide the 10000 genes into $k = 100$ blocks, each containing 100 genes. We assume that the genes in different blocks are independent of each other yet genes within the same blocks are correlated with each other with some covariance structure, which is taken to be the autoregressive structure here,

$$\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho^{98} & \rho^{99} \\ \rho & 1 & \ddots & \ddots & \rho^{98} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{98} & \ddots & \ddots & \ddots & \rho \\ \rho^{99} & \rho^{98} & \dots & \rho & 1 \end{pmatrix}. \tag{12}$$

For simplicity, we assume that $|\rho| = 0{\cdot}9$ to be same in all blocks and for 50 blocks ρ is positive and the other 50 ρ is negative. To generate the microarray expression data, we first generate each

gene's expression from a standard normal distribution $N(0,1)$ and then multiply the expression matrix by the square root of above covariance matrix. Now for each sample, the transformed gene profile has $MVN(0,\Sigma)$ distribution. We then add 1/2 to only the first 200 genes of samples from the second class.

This simulation setup does have sound basis in real microarray data. It is common knowledge that genes are networked in groups. Although it is true that weak connections between groups may exist, independence between groups is usually a reasonable assumption. Also, within each group, genes are either positively or negatively correlated and due to their relative distance in the regulatory pathway, the further apart two genes, the less correlation between them. These are exactly the reasons why we use the above simulation model. From the results in Table 14, we can clearly see that the SCRDA method outperforms PAM by a significant margin. Considering this is only a two-group classification problem mimicking the real microarray data, we should expect the difference will be more significant when the number of classes is large.

Table 14. *Setup III — two groups with dependent structure.*

Methods	10-fold CV Error	Ave. TE	# of genes selected	Min. TE[1]	Min. TE[2]
SCRDA	25/200	108/1000	282	107/1000	94/1000
Hard SCRDA	21/200	96/1000	167	92/1000	86/1000
SCRDA^r	25/200	123/1000	283	123/1000	113/1000
Hard SCRDA^r	25/200	125/1000	116	125/1000	111/1000
PAM	36/200	170/1000	749	170/1000	167/1000

5. Feature Selection by the SCRDA Method

Remember that the discriminant function (7) is linear in X with coefficients vector $\widetilde{\Sigma}^{-1}\bar{x}_g$. Now if the i-th element of the coefficients vector is 0 for all $1 \le g \le G$, then it means gene i doesn't contribute to our classification purpose and hence can be eliminated. Therefore, SCRDA potentially can be used for the gene selection purpose. For example, as shown in Table 15, the number of genes that are truly differentially expressed is 100, 280 and 200 respectively in the 3 simulation setups above. Correspondingly, SCRDA picks out 240, 785 and 282 genes in each setup. Among these genes, 82, 204 and 138 are truly differentially expressed respectively. The detection rate is at least 70% in all situations. However, the false positive rate is also high, especially when the number of classes is large. For now, there is not a good way to adjust this high false positive rate. Therefore, SCRDA can be conservatively used as gene selection method.

Table 15. *Feature selection by SCRDA.*

	Setup I	Setup II	Setup III
True Positive	82/100	204/280	138/200
Total Positive	240	785	282

6. Discussion

Through extensive comparison using both real microarray data sets and simulated data sets, we have shown that the SCRDA method can be a promising classification tool. Particularly, it is consistently better than its sibling method, PAM in many problems. This new method is also very competitive to some other methods, e.g., SVM.

This new method is not only empirically better in terms of classification performance, it also has some interesting theoretical implications. For example, it can be shown that the regularized discriminant function (7) is equivalent to the penalized log-likelihood method and in some special cases, our new method SCRDA can be related to another recently proposed new method called "elastic net" [Zou and Hastie, 2005]. These results are interesting since not only do they give different perspectives of statistical methods, they also provide new computational approaches. For example, an alternative method for estimating the shrunken regularized centroids other than the way we have proposed in this paper is to solve the solution of the mixed L^1-L^2 penalty function. This has been made possible as the problem will convert to the usual LASSO [Tibshirani, 1996] solution. And with the emergence of the new algorithm LARS [Efron *et al.*, 2004], efficient numerical solution is also available. More details about this discussion can be found in the technical report of this paper.

As mentioned before, choosing the optimal parameter pairs for the SCRDA method is not as straightforward as in PAM and in some cases, the process can be somewhat tedious. The guideline we gave in Section 3.4 works generally well, at least for all the data examples provided in this paper. However, it may require some experience with the SCRDA method to get the best result. Also, the computation in the SCRDA method is not as fast as in PAM due to two reasons. First, we have two parameters (α, Δ) to optimize over a 2-dimensional grid rather than the 1-dimensional one in PAM. Second, although the SVD algorithm is very efficient, the computation still involves large matrix manipulation in practice, while only vector operations are involved in PAM. On the other hand, as shown in this paper, PAM doesn't always do a worse job than the SCRDA method. In some situations, e.g., when the number of classes is small or the covariance structure is nearly diagonal, PAM is both accurate in prediction and computationally efficient. Therefore, we recommend using the SCRDA method only when PAM cannot perform well in classification. In general, the result from both PAM and SCRDA is reliable. This has been reflected by the numerical examples in this paper. Both method don't seem to over-train the training samples and therefore, the resulting test error is almost always about the same as the cross-validation error.

Also, the SCRDA method can be used directly for gene selection proposes. As we have seen in Section 5, the selection process of SCRDA is rather conservative, tending to include many more genes unnecessary. But overall speaking, it is not generally worse than PAM. And since it includes most of the genes that are truly differentially expressed, it is a safer way of not excluding the ones we really should detect.

References

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, **20**(18), 3583–3593.

Dipillo, P. (1976). The application of bias to discriminant analysis. *Communication in Statistics — Theory and Methodology*, **A5**, 843–854.

Dipillo, P. (1977). Further application of bias discriminant analysis. *Communication in Statistics — Theory and Methodology*, **A6**, 933–943.

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

Friedman, J. (1989). Regularized discriminant analysis. *Journal of American Statistical Association*, **84**, 165–175.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., and Caligiuri, M. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–536.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., and Golub, T. (2001). Multiclass cancer diagnosis using tumor gene expression signature. *PNAS*, **98**, 15149–15154.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, **58**(1), 267–288.

Tibshirani, R., Hastie, T., Narashimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids with applications to dna microarrays. *Statistical Science*, **18**, 104–117.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**(3), 427–443.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B*, **67**(2), 301–320.

[*Received*]