Biometrics 65, 1021–1029
December 2009
DOI: 10.1111/j.1541-0420.2009.01200.x

Shrinkage-based Diagonal Discriminant Analysis and Its Applications in High-Dimensional Data

Herbert Pang,[1,*] Tiejun Tong,[2,] and Hongyu Zhao[3,***]**

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina 27705, U.S.A.

[2]Department of Applied Mathematics, University of Colorado, Boulder, Colorado 80309, U.S.A.

[3]Division of Biostatistics, Department of Epidemiology and Public Health, and Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, U.S.A.

[*]*email:* herbert.pang@duke.edu

[**]*email:* tiejun.tong@colorado.edu

[***]*email:* hongyu.zhao@yale.edu

Summary. High-dimensional data such as microarrays have brought us new statistical challenges. For example, using a large number of genes to classify samples based on a small number of microarrays remains a difficult problem. Diagonal discriminant analysis, support vector machines, and k-nearest neighbor have been suggested as among the best methods for small sample size situations, but none was found to be superior to others. In this article, we propose an improved diagonal discriminant approach through shrinkage and regularization of the variances. The performance of our new approach along with the existing methods is studied through simulations and applications to real data. These studies show that the proposed shrinkage-based and regularization diagonal discriminant methods have lower misclassification rates than existing methods in many cases.

Key words: Discriminant analysis; Large p small n; Microarray; Regularization; Shrinkage; Tumor classification.

1. Introduction

High throughput gene expression technologies have opened a whole new path in the development of novel ways to treat cancer and other diseases. The availability of these data has motivated the development of reliable biomarkers for disease diagnosis and prognosis, and the identification of drug targets for treatment. Many methods have been developed in the literature, such as diagonal linear discriminant analysis (DLDA; Dudoit, Fridlyand, and Speed, 2002), random forests (Breiman, 2001), support vector machines (SVM; Vapnik and Kotz, 2006), and penalized discriminant methods (Ghosh, 2003). These methods have been applied to many studies (e.g., Huang and Zheng, 2006; Moon et al., 2006; Montaner et al., 2006; Barrier et al., 2007). One particular disease area that has contributed most in shaping the development of microarray data collection and analysis is cancer. A well-known paper published by Golub et al. (1999) is a leukemia study using microarray data to identify cancer molecular subtypes. They used a weighted nearest-neighbor scoring method for discrimination between acute myeloid leukemia and acute lymphoblastic leukemia.

The three main statistical problems in cancer genomics research are (1) the identification of subclasses within a particular tumor type; (2) the classification of patients into known classes; and (3) the selection of biomarkers, i.e., genes that characterize a particular tumor subtype. In this article, we will mainly focus on (2), discriminant methods for classifying human tumors based on microarray data, which is unique in the sense that the number of samples is much smaller than the number of features (genes). The data available in public databases now contain mainly expression data ranging between 10,000 and 55,000 probes or probe sets for fewer than 100 samples. For some cancers, e.g., brain tumors, it is not uncommon to see fewer than 10 subjects per tumor group (e.g., Pomeroy et al., 2002; Dong et al., 2005). Therefore, there is a need to develop methods that have good performance when the sample size is small.

In 2002, Dudoit et al. performed a comprehensive comparison of various discriminant methods on different microarray data sets for different types of cancer. They compared nearest neighbors, classification trees, and linear and quadratic discriminant analysis, and found that nearest neighbors and DLDA had the smallest error rates. In more recent studies, SVM has been found to be one of the better classifiers (e.g., Lee et al., 2005; Shieh, Jiang, and Shih, 2006). Many researchers have pointed out that for high-dimensional data with small sample sizes, the naive Bayes classifier, sometimes known as DLDA and diagonal quadratic discriminant analysis (DQDA), has comparable or better performance than SVM, see for example Ye et al. (2004), Lee et al. (2005), and Shieh et al. (2006). Moreover, in situations where the sample size of each group is less than 10, it is clear that regularization and shrinkage techniques will enhance and improve estimation. The reason is that the commonly used estimators for

the class-specific variances or the pooled variance in DQDA or DLDA can become unstable and thus reduce the classification accuracy.

One solution to the challenge of having a small number of samples compared to the large number of genes in the microarray settings is to make use of shrinkage-based variance estimators. Tong and Wang (2007) derived a family of shrinkage estimators for gene-specific variances raised to a fixed power (nonzero) extending the idea from Cui et al. (2005) to a more general setting. These estimators borrow information across genes by shrinking each gene-specific variance estimator toward bias-corrected geometric mean of variance estimators for all genes. Their method has been applied to multiple testing problems by introducing an *F*-like statistic. To the best of our knowledge, their James–Stein shrinkage-based variance estimation has not been explored as a tool for improving discriminant analysis in microarray data analysis. This has given us the motivation to propose new shrinkage-based discriminant methods and to perform a comprehensive study on their performance.

In this article, we propose a new approach to improve DLDA and DQDA by applying shrinkage and regularized techniques to discrimination. We first improve upon the original DLDA and DQDA by performing shrinkage, which is in essence a method to borrow information across genes to improve estimation of the gene-specific variances by shrinking them toward a pooled variance. Secondly, we further improve the shrinkage-based DLDA and DQDA by using regularization, which is essentially a weighted version of the shrinkage-based DLDA and DQDA. Combining shrinkage-based variance in diagonal discriminant analysis and regularization results in a new classification scheme that shows improvement over the original DLDA, DQDA, SVM, and *k*-nearest neighbor (*k*-nn) in many scenarios, especially in small sample size settings.

2. Regularized Shrinkage-based Diagonal Discriminant Analysis

2.1 *Diagonal Discriminant Analysis*

The main purpose of discriminant analysis is to assign an unknown subject to one of K classes on the basis of a multivariate observation $\boldsymbol{x} = (x_1, \ldots, x_p)^T$, where p is the number of features. For simplicity of notation, the class labels y_i are defined to be integers ranging from 1 to K. We assume that there are n_k observations in class k with

$$\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,n_k} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_k, \Sigma_k), \quad k = 1, \ldots, K,$$

where $\boldsymbol{\mu}_k$ and Σ_k are the corresponding mean vector and covariance matrix of the p-dimensional multivariate normal distribution. The total number of observations is $n = n_1 + \cdots + n_K$.

Let π_k denote the prior probability of observing a class k member with $\pi_1 + \cdots + \pi_K = 1$. Under the normal distribution assumption, we assign a new subject $\boldsymbol{x}$ to class k, which minimizes the following discriminant score

$$D_k(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln|\Sigma_k| - 2\ln\pi_k, \quad (1)$$

i.e., we assign $\boldsymbol{x}$ to $\hat{k} = \operatorname{argmin}_k D_k(\boldsymbol{x})$. This is the so-called quadratic discriminant analysis (QDA) since the boundaries that separate the disjoint regions belonging to each class are quadratic. The first term on the right-hand side of equation (1) is known as the squared Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_k$. When the covariance matrices are all the same, i.e., $\Sigma_k = \Sigma$ for all k, the discriminant score can be simplified as

$$d_k(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) - 2\ln\pi_k. \quad (2)$$

This is referred to as the linear discriminant analysis (LDA). LDA assigns a new subject to $\hat{k} = \operatorname{argmin}_k d_k(\boldsymbol{x})$ which uses linear boundaries.

Note that both mean vectors $\boldsymbol{\mu}_k$ and covariance matrices Σ_k are unknown for microarray data, and need to be estimated from the training set. In practice, the most commonly used estimators are their maximum-likelihood estimates,

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_{k,i}, \quad \hat{\Sigma}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} (\boldsymbol{x}_{k,i} - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{k,i} - \hat{\boldsymbol{\mu}}_k)^T,$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{K} n_k \hat{\Sigma}_k.$$

The prior probabilities are usually estimated by the fraction of each class in the pooled training sample, i.e., $\hat{\pi}_k = n_k/n$.

The sample version rule for QDA is $\mathcal{C}(\boldsymbol{x}) = \operatorname{argmin}_k \hat{D}_k(\boldsymbol{x})$, where

$$\hat{D}_k(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k) + \ln|\hat{\Sigma}_k| - 2\ln\hat{\pi}_k.$$

Similarly, the sample version rule for LDA is $\mathcal{C}(\boldsymbol{x}) = \operatorname{argmin}_k \hat{d}_k(\boldsymbol{x})$, where

$$\hat{d}_k(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k)^T \hat{\Sigma}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k) - 2\ln\hat{\pi}_k.$$

These so-called "plug-in" estimates are straightforward to compute, but do not enjoy optimality properties (Anderson, 1958; Friedman, 1989). Classification rules based on QDA are known to require generally larger samples than those based on LDA (Wald and Kronmal, 1977) and are more sensitive to departures from basic model assumptions.

QDA requires that $\min\{n_1, \ldots, n_K\}$ is greater than or equal to p, the number of features, to ensure that the sample covariance matrices are nonsingular. LDA requires that $n \geqslant p$ to make Σ nonsingular. For high-dimensional data, especially for microarray data, it is common that the dimension is much larger than the sample size, i.e., $p \gg n$. This implies that traditional methods based on QDA and LDA cannot be applied to microarrays directly. Though we may use the generalized matrix inverse or use a regularized covariance matrix such as $\lambda\Sigma_k + (1-\lambda)I$ or $\lambda\Sigma + (1-\lambda)I$, such estimates are usually unstable due to the lack of observations.

To overcome the singularity problem, Dudoit et al. (2002) introduced two simplified discriminant rules by assuming independence between covariates and replacing off-diagonal elements of the sample covariance matrices with zeros. The first rule is called DQDA. Specifically, they estimate $\hat{\Sigma}_k = \operatorname{diag}(\hat{\sigma}_{k1}^2, \ldots, \hat{\sigma}_{kp}^2)$, and give the discriminant rule as $\mathcal{C}(\boldsymbol{x}) = \operatorname{argmin}_k \hat{D}_k^D(\boldsymbol{x})$, where

$$\hat{D}_k^D(\boldsymbol{x}) = \sum_{j=1}^{p} (x_j - \hat{\mu}_{kj})^2 \big/ \hat{\sigma}_{kj}^2 + \sum_{j=1}^{p} \ln\hat{\sigma}_{kj}^2 - 2\ln\hat{\pi}_k. \quad (3)$$

The second rule is called DLDA. Specifically, they assume a common diagonal covariance matrix and estimate $\hat{\Sigma} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$. The discriminant rule is then

$$\mathcal{C}(\boldsymbol{x}) = \underset{k}{\text{argmin}} \left(\sum_{j=1}^{p} (x_j - \hat{\mu}_{kj})^2 \big/ \hat{\sigma}_j^2 - 2 \ln \hat{\pi}_k \right). \tag{4}$$

Due to the small sample size, DLDA and DQDA, which ignore correlations between genes, performed remarkably well in practice and produced lower misclassification rates than more sophisticated classifiers (Dudoit et al., 2002). In addition, the scoring of DLDA and DQDA is easy to implement and not very sensitive to the number of predictor variables. DQDA and DLDA classifiers are sometimes called "naive Bayes" because they arise in a Bayesian setting. The reason why DLDA is a success is that even when the diagonal of the covariance matrices is substantially different, the drop in variance resulting from the use of the pooled estimate may lead to better performance, especially when the sample size is small.

For a more theoretical account on why the "naive Bayes" classifier that assumes independent covariates works well when $p > n$, see Bickel and Levina (2004). Specifically, they show in their Section 2 that under the worst-case scenario the "naive Bayes" classifier, which assumes independent covariates, greatly outperforms Fisher's linear discriminant function. They also demonstrate that under the assumption of known covariance matrix the "naive Bayes" rule is still at par with the original rule.

2.2 *Shrinkage-based Diagonal Discriminant Analysis*

Because n is typically much smaller than the number of features p for microarray data, the performance of DQDA or DLDA might not even be satisfactory due to the unreliable estimates of the sample variances. Therefore, we propose modifications to the original DQDA and DLDA to further improve their performance. This is achieved by developing several regularized discriminant rules to improve the variance estimation. For ease of notation, in what follows we focus on the derivation of the shrinkage-based DLDA only. The corresponding result for DQDA will be presented at the end of the section. Recall that for DLDA, the diagonal discriminant score is

$$\hat{d}_k^D(\boldsymbol{x}) = \sum_{j=1}^{p} (x_j - \hat{\mu}_{kj})^2 \big/ \hat{\sigma}_j^2 - 2 \ln \hat{\pi}_k,$$

where the first term on the right side is the so-called squared Mahalanobis distance.

Denote $\nu = n - K$, $\hat{\sigma}_j^{2t} = (\hat{\sigma}_j^2)^t$, $\hat{\sigma}_{pool}^{2t} = \prod_{j=1}^{p} (\hat{\sigma}_j^2)^{t/p}$ and

$$h_{\nu,p}(t) = \left(\frac{\nu}{2}\right)^t \left(\frac{\Gamma(\nu/2)}{\Gamma(\nu/2 + t/p)} \right)^p,$$

where $\Gamma(\cdot)$ is the gamma function. Tong and Wang (2007) proposed the following family of shrinkage estimators for σ_j^{2t},

$$\tilde{\sigma}_j^{2t}(\alpha) = \left(h_{\nu,p}(t) \hat{\sigma}_{pool}^{2t}\right)^{\alpha} \left(h_{\nu,1}(t) \hat{\sigma}_j^{2t}\right)^{1-\alpha}, \tag{5}$$

where $h_{\nu,1}(t)\hat{\sigma}_j^{2t}$ is an unbiased estimator of σ_j^{2t}, and $h_{\nu,p}(t)\hat{\sigma}_{pool}^{2t}$ is an unbiased estimator of σ^{2t} when $\sigma_j^2 = \sigma^2$ for all j. The shrinkage parameter $\alpha \in [0, 1]$ controls the degree of shrinkage from the individual variance estimate toward the bias-corrected pooled estimate. There is no shrinkage when $\alpha = 0$, and all variance estimates are shrunken to the pooled estimate when $\alpha = 1$. Let $\boldsymbol{\sigma}^{2t} = (\sigma_1^{2t}, \ldots, \sigma_p^{2t})$, $\hat{\boldsymbol{\sigma}}^{2t} = (\hat{\sigma}_1^{2t}, \ldots, \hat{\sigma}_p^{2t})$, and $\Psi(\cdots) = \Gamma'(\cdots)/\Gamma(\cdots)$ the digamma function. Under the Stein loss function $L_{Stein}(\sigma^2, \tilde{\sigma}^2) = \tilde{\sigma}^2/\sigma^2 - \ln(\tilde{\sigma}^2/\sigma^2) - 1$, Tong and Wang (2007) proved that for any fixed p, ν, and $t > -\nu/2$, there exists a unique optimal shrinkage parameter α^* as the solution to $(\partial/\partial\alpha) R_{Stein}(\boldsymbol{\sigma}^{2t}, \tilde{\boldsymbol{\sigma}}^{2t}) = 0$, where the average risk is given by

$$\begin{aligned} R_{Stein}(\boldsymbol{\sigma}^{2t}, \tilde{\boldsymbol{\sigma}}^{2t}) = {} & \frac{h_{\nu,p}^{\alpha}(t) h_{\nu,1}^{1-\alpha}(t)}{h_{\nu,1}^{p-1}(\alpha t/p)\, h_{\nu,1}((1-\alpha+\alpha/p)t)} \left(\sigma_{pool}^2\right)^{\alpha t} \\ & \times \frac{1}{p} \sum_{j=1}^{p} (\sigma_j^2)^{-\alpha t} - \ln\left(h_{\nu,p}^{\alpha}(t) h_{\nu,1}^{1-\alpha}(t)\right) \\ & - t\Psi\left(\frac{\nu}{2}\right) + t \ln\left(\frac{\nu}{2}\right) - 1. \end{aligned}$$

In practice, α^* is unknown and needs to be estimated because $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_p^2)$ are unknown. For microarray data with at least four replicates for each class, a consistent estimator of α^* exists for both $\tilde{\sigma}_j^2(\alpha)$ and $\tilde{\sigma}_j^{-2}(\alpha)$. Otherwise, Tong and Wang (2007) suggested an alternative two-step procedure to estimate the optimal shrinkage parameter (see Section 3.3 of their paper for more details).

An important insight of Tong and Wang (2007) is that a better variance estimator does not necessarily lead to a more powerful test. In their paper, an F-test using the inverse of the variance is more powerful than using the reciprocal of an estimator. Note that a similar argument holds in discriminant analysis since the variances σ_j^2 appear in the denominator too. Therefore, for the estimation procedures that we propose, we consider using shrinkage estimators for σ_j^2 $(t = 1)$ or estimators for $1/\sigma_j^2$ $(t = -1)$. The formulas, as well as the implementation of our methods, can be developed analogously. In practice it turns out that this choice is not important, and results are very similar in every situation that we studied. Thus, for simplicity we focus in the remainder of the paper on $t = -1$. Results for $t = 1$ can be requested from the authors.

Specifically, the shrinkage-based discriminant rule is $\text{argmin}_k \tilde{d}_k^{-D}(\boldsymbol{x})$, where

$$\tilde{d}_k^{-D}(\boldsymbol{x}) = \sum_{j=1}^{p} (x_j - \hat{\mu}_{kj})^2 \tilde{\sigma}_j^{-2}(\hat{\alpha}^*) - 2 \ln \hat{\pi}_k, \tag{6}$$

where $\tilde{\sigma}_j^{-2}(\hat{\alpha}^*)$ is the estimate of $1/\sigma_j^2$. We will subsequently call this method SDLDA which is short for shrinkage-based DLDA.

Similarly, we can propose shrinkage-based DQDA by shrinking variances within each class k. Denote $\boldsymbol{\sigma}_k^{2t} = (\sigma_{k1}^{2t}, \ldots, \sigma_{kp}^{2t})$ and $\tilde{\boldsymbol{\sigma}}_k^{2t} = (\tilde{\sigma}_{k1}^{2t}, \ldots, \tilde{\sigma}_{kp}^{2t})$ for any $k = 1, \ldots, K$. Let

$$\alpha_k^* = \underset{\alpha \in [0,1]}{\text{argmin}}\ R_{Stein}\left(\boldsymbol{\sigma}_k^{2t}, \tilde{\boldsymbol{\sigma}}_k^{2t}\right).$$

Then the shrinkage-based discriminant rule is

$$\underset{k}{\text{argmin}} \left(\sum_{j=1}^{p} (x_j - \hat{\mu}_{kj})^2 \tilde{\sigma}_{kj}^{-2}\left(\hat{\alpha}_k^*\right) - \sum_{j=1}^{p} \ln \tilde{\sigma}_{kj}^{-2}\left(\hat{\alpha}_k^*\right) - 2 \ln \hat{\pi}_k \right).$$

Following our naming conventions, we refer to this as SDQDA.

2.3 *A Regularization Approach*

In this section, we propose a new Regularized Shrinkage-based Diagonal Discriminant Analysis (RSDDA) based on the shrinkage variances introduced in the previous section. Regularization techniques are not new in statistics. The estimation of parameters can be highly unstable when the number of parameters to be estimated outnumbers the sample size by a few folds. In our case, regularization attempts to improve the estimates by biasing them away from their shrinkage-based class values toward the shrinkage-based pooled values. It introduces biases, which is compensated for the reduction in the variances associated with the class-based estimate.

In this setting, λ is the parameter that controls the strength of biasing toward the pooled parameter. For a given value of λ, the increase in bias will depend on how closely the pooled value represents that of the population. Depending on whether the pooled value is a good measure or not, one would adjust and employ a small or high degree of regularization.

Denote $\hat{D}_k^t = \text{diag}(\hat{\sigma}_{k1}^{2t}, \ldots, \hat{\sigma}_{kp}^{2t})$, $\hat{V}_k^t = \text{diag}(\ln \hat{\sigma}_{k1}^{2t}, \ldots, \ln \hat{\sigma}_{kp}^{2t})$, and $\mathbf{1} = (1, \ldots, 1)^T$. For DQDA, we have

$$\begin{aligned}\hat{D}_k^D(\boldsymbol{x}) &= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k)\hat{D}_k^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k) + \mathbf{1}^T \hat{V}_k \mathbf{1} - 2\ln \hat{\pi}_k \\ &= L_{k1} + L_{k2} - 2\ln \hat{\pi}_k,\end{aligned}$$

where $L_{k1} = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k)\hat{D}_k^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k)$ and $L_{k2} = \mathbf{1}^T \hat{V}_k \mathbf{1}$. To improve the estimates of the discriminant score, $\hat{D}_k^D(\boldsymbol{x})$, is then equivalent to improve the estimators L_{k1} and L_{k2}, given that the prior probabilities $\hat{\pi}_k$s stay the same.

Consider the following regularized covariance matrix for $\hat{V}_k$,

$$\tilde{V}_k^t(\alpha_k) = (1 - \alpha_k)\hat{V}_k^t + \alpha_k \ln\big(h_{\nu_k,p}(t)\hat{\sigma}_{k,\,pool}^{2t}\big) I_p, \qquad (7)$$

where $\nu_k = n_k - 1$, $\tilde{V}_k^t(\alpha_k) = \text{diag}(\ln \tilde{\sigma}_{k1}^{2t}, \ldots, \ln \tilde{\sigma}_{kp}^{2t})$, $\hat{V}_k^t = \text{diag}(\ln \hat{\sigma}_{k1}^{2t}, \ldots, \ln \hat{\sigma}_{kp}^{2t})$, $\hat{\sigma}_{k,\,pool}^{2t} = \prod_{j=1}^{p} (\hat{\sigma}_{kj}^{2t})^{1/p}$, and I_p is the identity matrix of size p. Then it is easy to see that

$$\tilde{\sigma}_{kj}^{-2}(\alpha_k) = \big(h_{\nu_k,p}(-1)\hat{\sigma}_{k,\,pool}^{-2}\big)^{\alpha_k} \big(h_{\nu_k,1}(-1)\hat{\sigma}_{kj}^{-2}\big)^{1-\alpha_k},$$

which reduces to (5), if we estimate $1/\sigma_{kj}^2$ directly by $\tilde{\sigma}_{kj}^{-2}$, with $t = -1$. As in Friedman (1989), we now propose a regularized discrimination, called RSDDA, by taking the following regularization for the matrix $\tilde{V}_k^t$. Specifically, we estimate

$$\breve{V}_k^t(\lambda, \boldsymbol{\alpha}) = (1 - \lambda)\tilde{V}_k^t(\alpha_k) + \lambda \tilde{V}^t(\alpha_{pool}), \quad \text{for } t > -\nu/2, \quad (8)$$

where $\boldsymbol{\alpha} = (\alpha_k, \alpha_{pool})$, $\tilde{V}^t(\alpha_{pool}) = (1 - \alpha_{pool})\hat{V}^t + \alpha_{pool} \times \ln(h_{\nu,p}(t)\hat{\sigma}_{pool}^{2t}) I_p$ with $\hat{V}^t = \text{diag}(\ln \hat{\sigma}_1^{2t}, \ldots, \ln \hat{\sigma}_p^{2t})$. The regularization parameter λ takes value within [0, 1], with $\lambda = 0$ giving rise to SDQDA and $\lambda = 1$ to SDLDA. It is also interesting to see that (8) is equivalent to estimating σ_{kj}^{2t} by

$$\begin{aligned}\breve{\sigma}_{kj}^{2t}(\lambda, \boldsymbol{\alpha}) &= \big\{\tilde{\sigma}_{kj}^{2t}(\alpha_k)\big\}^{1-\lambda} \big\{\tilde{\sigma}_j^{2t}(\alpha_{pool})\big\}^{\lambda} \\ &= \Big\{\big(h_{\nu_k,p}(t)\hat{\sigma}_{k,pool}^{2t}\big)^{\alpha_k} \big(h_{\nu_k,1}(t)\hat{\sigma}_{kj}^{2t}\big)^{1-\alpha_k}\Big\}^{1-\lambda} \\ &\quad \times \Big\{\big(h_{\nu,p}(t)\hat{\sigma}_{pool}^{2t}\big)^{\alpha_{pool}} \big(h_{\nu,1}(t)\hat{\sigma}_j^{2t}\big)^{1-\alpha_{pool}}\Big\}^{\lambda}.\end{aligned}$$

As mentioned in Section 3.2, in what follows we focus only on $t = -1$ and we refer to RSDDA as RSDDA ($t = -1$). And correspondingly, we replace $\hat{D}_k^t$ in L_{k1} by $\breve{D}_k^t = \text{diag}(\breve{\sigma}_{k1}^{-2}(\lambda, \boldsymbol{\alpha}), \ldots, \breve{\sigma}_{kp}^{-2}(\lambda, \boldsymbol{\alpha}))$.

RSDDA provides a fairly rich class of regularization alternatives. The following four special cases define well-known classification rules:

i) ($\alpha_1 = 0, \ldots, \alpha_K = 0$, $\alpha_{pool} = 0$, $\lambda = 0$) represents DQDA;
ii) ($\alpha_1 = 1, \ldots, \alpha_K = 1$, $\alpha_{pool} = 1$, $\lambda = 0$) represents weighted nearest-means classifier;
iii) ($\alpha_1 = 0, \ldots, \alpha_K = 0$, $\alpha_{pool} = 0$, $\lambda = 1$) represents DLDA;
iv) ($\alpha_1 = 1, \ldots, \alpha_K = 1$, $\alpha_{pool} = 1$, $\lambda = 1$) represents the nearest-means classifier.

In addition, keeping all the αs equal to 0 and varying λ gives the down-weighted nearest-means classifier, with no weight at $\lambda = 1$. While keeping $\lambda = 0$ and varying α_k leads to SDQDA and keeping $\lambda = 1$ and varying α_{pool} leads to SDLDA.

Note that the values of αs and λ are not likely to be known in advance, and usually need to be estimated from the training set. In practice, there are two possible choices for estimating α and λ:

Approach 1. Estimate α_k and α_{pool} by α_k^* and α_{pool}^*, respectively as in Tong and Wang (2007), and use a cross-validation or bootstrapping method to estimate λ within [0, 1] through a grid search.

Approach 2. Use a cross-validation or bootstrapping method to choose the $K + 2$ parameters for the K-class discrimination problem, $(\alpha_1, \ldots, \alpha_K, \alpha_{pool}, \lambda)$ through a grid search in the (K+2)-dimensional space $[0, 1]^{K+2}$. Due to high-dimensional search, this is difficult given the computational load.

In this article, we take Approach 1 since it is computationally less expensive.

3. Simulation Studies

3.1 *Simulation Design*

In this section we describe the design of our simulations to assess the performance of the proposed shrinkage-based diagonal discriminant rules. We consider both misclassification rates and the accuracy of the proposed shrinkage-based discriminant scores.

First, we examine the misclassification rates. We will investigate how our new methods work in a simulation study. Three different simulation setups were devised to investigate the behavior of the proposed SDQDA, SDLDA, and RSDDA in a controlled manner. We chose SVM and k-nn, two well-known classification schemes for comparison. Four different kernels for SVM were chosen in our analysis: radial basis, linear, polynomial, and sigmoid. For k-nn, we also tried $k = 1$, 3, and 5. Grid search was performed to identify the degree of regularization, our λ of equation (12) was equally spaced between 0 and 1 with a 0.01 step size.

In Setup (A), we consider two classes of multivariate normal distributions: $N_p(\boldsymbol{\mu_1}, \Sigma)$ and $N_p(\boldsymbol{\mu_2}, \Sigma)$. All components of $\boldsymbol{\mu}_1$ are 0 and for $\boldsymbol{\mu}_2$ are 0.5. The covariance is of independent structure with $\Sigma = I_p$. We simulated data with three different dimensions of p : $p = 30$, 50, 100 and 300. Setup (B)

is basically the same as Setup (A) except that this time $\boldsymbol{\mu}_2$ is equal to 1, i.e., the two classes have better separations.

Misclassification rates were calculated as follows: for each simulation, a training set of size n was generated using the setups described above, and a validation set of size $2n$ was generated with the identical setup in order to assess the error rate. The mean error rates for each method were obtained by running 500 simulations and taking an average over them. For each setup, we generated training sets of $n = 4$, 5, 8, 10, and 15 for the respective validation set of size $2n$.

Setup (C) considers a more realistic covariance matrix structure. Let us consider the case like (A) where $\boldsymbol{\mu}_1$ are 0 and $\boldsymbol{\mu}_2$ are 0.5. This time the covariance matrix Σ is a block-diagonal matrix of size 2500 by 2500 with each of the 50 by 50 diagonal blocks Σ_ρ alternating in sign, and the rest of the matrix is zero, where

$$\Sigma = \begin{pmatrix} \Sigma_\rho & 0 & \dots & \dots & \dots & \vdots \\ 0 & \Sigma_{-\rho} & 0 & \ddots & \ddots & \vdots \\ \vdots & 0 & \Sigma_\rho & 0 & \ddots & \vdots \\ \vdots & \ddots & 0 & \Sigma_{-\rho} & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 & \Sigma_\rho & \vdots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}_{2500\times 2500},$$

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \dots & \rho^{48} & \rho^{49} \\ \rho & 1 & \ddots & \dots & \rho^{48} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{48} & \dots & \ddots & 1 & \rho \\ \rho^{49} & \rho^{48} & \dots & \rho & 1 \end{pmatrix}_{50\times 50}.$$

This is a similar setup as the one in Guo, Hastie, and Tibshirani (2007) except that we took a smaller covariance matrix to ease the memory and computational load. Since having 50 genes in a pathway is reasonable, the block matrix of size 50×50 was chosen. Matrices with an autocorrelation of $|\rho| = 0.5$, 0.7, 0.8, and 0.9 were chosen for Setup (C). Misclassification rates were calculated the same way as in the first two setups, except that we first performed a gene selection procedure. The selection of the top genes of size 30, 50, and 100 was done according to the ratio of between-group to within group sums of squares (Dudoit et al., 2002). Specifically, the ratio for gene j is:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{*j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2}. \tag{9}$$

Second, we examine the accuracy of the estimation. We compare the mean squared errors (MSE) from the simulated data for $\hat{d}_k^D$, $\tilde{d}_k^D$, and $\tilde{d}_k^{-D}$ for both DLDA and DQDA using the setups we have described above.

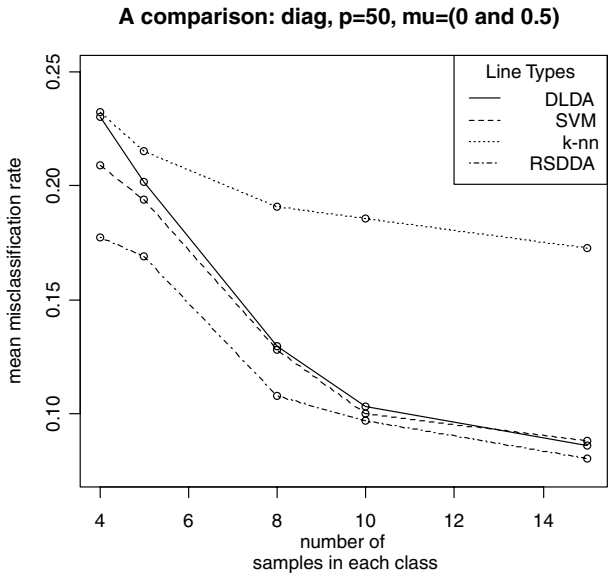

Figure 1. A comparison between the original DLDA, SVM, k-nn, and the newly proposed RSDDA for simulation setup (A). This plot investigates the effect of number of samples in each class on the improvement RSDDA over other methods. It shows a general pattern for other setups that RSDDA does better than all the other existing methods. The difference is more evident for smaller sample sizes, i.e., less than 8.

3.2 *Simulation Results*

There is no clear pattern as to which of the variants of SVM and k-nn performed better than others. Therefore, we have decided to keep the defaults in the following figures and tables. See Figure 1 for a comparison between the original DLDA, SVM, k-nn, and our RSDDA for Setup (A) with $p = 50$. In this simple setup, we see that RSDDA performs better than both SVM and k-nn. RSDDA shows mean misclassification errors that are comparable to the shrinkage-based method, see Table 1 for sample size 5. The original DLDA is close to SVM and k-nn in a majority of the scenarios. But the shrinkage-based and regularization discrimination methods are better and in a league of their own. This result is more evident when the ratio of the number of features to the sample size gets larger.

Table 1

A comparison of mean misclassification rates between the original DQDA, DLDA, SVM, k-nn, and the newly proposed classifiers for simulation Setup (A) and 10 samples (five samples in each group)

Method	30 genes	50 genes	100 genes
DQDA	0.341	0.324	0.271
DLDA	0.242	0.202	0.116
SVM	0.247	0.194	0.089
k-nn ($k = 3$)	0.278	0.215	0.136
SDLDA	0.226	0.171	0.070
RSDDA	0.229	0.169	0.070

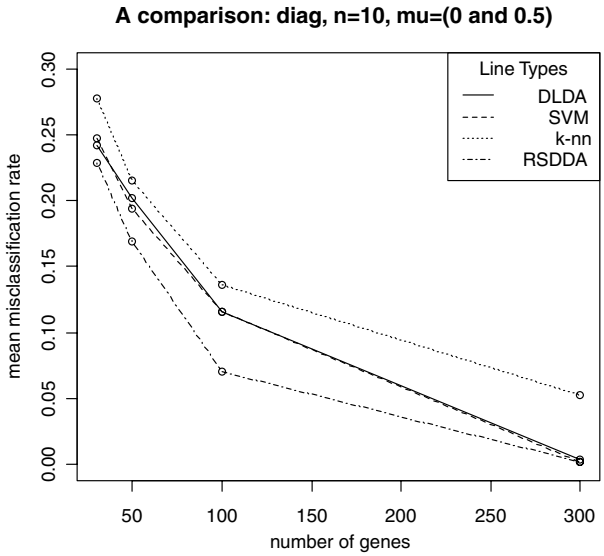

Figure 2. A comparison between the original DLDA, SVM, k-nn, and the newly proposed RSDDA for simulation setup (A). This plot investigates the effect of number of genes in each class on the improvement RSDDA over other methods. It shows a general pattern for other setups that RSDDA does better than all the other existing methods.

Figure 2 illustrates the effect of the number of genes on misclassification rates for DLDA, SVM, k-nn, and RSDDA for 10 samples (5 per group). RSDDA once again shows improvement over existing methods for less than 100 genes. For 300 genes, all the methods except k-nn have misclassification rates close to zero.

In Setup (B), we see a similar trend that RSDDA $\geqslant$ SDLDA $\geqslant$ DQDA/DLDA; although, this time, SVM and k-nn's performance is only slightly worse than SDLDA. Due to the high degree of separation between the means of the two groups, we only observe a tiny improvement over the traditional methods, see Web Table 3.

For Setup (C), we see that RSDDA outperforms all other methods under this setup that more closely resembles real microarray data, see Table 2 for the case p = 50 and $|\rho| = 0.5$. SDLDA follows RSDDA closely and only does slightly worse. Although SVM and k-nn outperform both DQDA and DLDA in the majority of cases, they do worse than the newly proposed SDLDA and RSDDA for $|\rho| \leqslant 0.7$, except for one case when SVM slightly edges out RSDDA. SVM performs best when $|\rho|$ is more than 0.8 with RSDDA and k-nn just behind in those cases. For the cases of $p = 30$ and $p = 100$, we see similar results. Overall, DLDA performs better than DQDA. SDLDA improves upon the original DLDA and it is in turn slightly inferior to the regularized method.

Table 2
A comparison of mean misclassification rates between the original DQDA, DLDA, SVM, k-nn, and the newly proposed RSDDA for simulation Setup (C), $\rho = 0.5$, 10 samples (5 samples in each group)

Method	30 genes	50 genes	100 genes
DQDA	0.236	0.177	0.137
DLDA	0.139	0.089	0.035
SVM	0.131	0.074	0.027
k-nn ($k = 3$)	0.131	0.069	0.025
RSDDA	0.126	0.063	0.020

As for the accuracy of the estimated discriminant scores, we observe that for Setups (A) and (B), the shrinkage-based estimates have smaller MSEs in almost all cases than the original DQDA and DLDA with SDQDA doing slightly better. Due to page constraints, we have left more simulation studies regarding the misclassification rates comparison and the accuracy of the discriminant scores estimation to the Supplementary Materials.

4. Applications to Microarray Data

In this section, we investigate the performance of the RSDDA method on real microarray data sets. First, we compare our methods with existing ones by subsetting a microarray data with a large sample size to simulate small sample size training data. Data from four different studies with small sample sizes ($\leqslant 10$) in each class are chosen. Two of the microarray data sets contain binary outcomes and two have more than two outcomes. The details of these data sets are discussed below.

We vary the number of top genes chosen, from 10, 50, to 200 for each data set. The top genes are selected from the training set for each cross-validation cut using the ratio of between-group to within-group sums of squares as described in Section 3.1. In addition, for each data set, we standardize the expression data, i.e., the observations (arrays) have mean 0 and variance 1 across genes as described in Dudoit et al. (2002). A grid search as done in the simulation is used to tune the regularization parameter.

4.1 *Subsetting Analysis on Microarray Data Set*

A large Multiple Myeloma microarray data set (Zhan et al., 2007) with 351 patients in the Therapy 2 group and 208 patients in the Therapy 3 group is used to conduct a subsetting analysis. One hundred simulations are done and for each simulation we take a random sample of five or eight patients from each group. A test set of size 203 or 200 from each group is then used to assess the error rate. This is performed for the top 10, 50, and 200 genes. This allows us to see how well our newly proposed methods performed for small sample size microarray data compared with existing methods given a large test set.

For the simulations based on subsetting the large microarray data set, see Table 3. Consistent with what we have found in the previous simulations, RSDDA and SDLDA outperform the original DDA methods, SVM, and k-nn for the top 10, 50 and 200 in both settings with five samples and eight samples per group. This is still true for other variants of SVM and k-nn. We investigate the significance of improvements over existing DLDA methods using paired t-test across the 100 independent runs. The p-values of the paired t-test to test for a difference between the misclassification rates of DLDA and RSDDA are all significant except one at the 0.05 level, see Web Table 5.

Table 3

Mean misclassification rates for a simulation of small sample size data using a large data set. In each of the 100 *simulations, five or eight samples per group are used as the training set and the rest, around* 200 *samples, are used as the test set*

Method	5 samples per group	8 samples per group
Top 10		
DQDA	0.3437	0.2206
DLDA	0.3114	0.2004
SVM	0.3168	0.2110
k-nn ($k = 3$)	0.3105	0.2096
SDLDA	0.3023	0.1956
RSDDA	0.3046	0.1965
Top 50		
DQDA	0.3092	0.2009
DLDA	0.2585	0.1637
SVM	0.2726	0.1795
k-nn ($k = 3$)	0.2979	0.1913
SDLDA	0.2477	0.1595
RSDDA	0.2555	0.1613
Top 200		
DQDA	0.2967	0.2070
DLDA	0.2481	0.1794
SVM	0.2767	0.1891
k-nn ($k = 3$)	0.2967	0.1885
SDLDA	0.2405	0.1756
RSDDA	0.2488	0.1760

4.2 *Binary Outcome Data Sets*

Apart from the subsetting analysis in the previous section, we evaluate our methods using several other real data sets. In order to assess the performance of the different methods, we randomly divide the data into training sets and validation sets. Approximately 60% of the samples are assigned to the training set. The rest, about 40%, is used as a validation set to assess the error rate. This process is repeated 100 times. For every training set, gene selection is performed as outlined in equation (13) in the simulations section.

In this section, we consider two data sets having binary outcomes. Huttmann et al. (2006) is a leukemia study and Dong et al. (2005) is a brain tumor study. Both studies use the Affymetrix HGU-133a chips. The Huttmann et al. (2006) data set contains 22,215 probe sets. It is a study consisting of 16 B-cell chronic lymphocytic leukemia patients, half of which (i.e., eight subjects) have good prognosis and the other half of poor prognosis. For the top 10, 50, and 200 genes, we see that RSDDA dominates (Table 4). When we consider the top 200 genes, k-nn performs worst among all the methods. The Dong et al. (2005) data set is a balanced design study with nine specific tumor cells, pseudopalisading cells, and nine controls, common tumor cells, in human glioblastoma. It is also the same Affymetrix chipset as the Huttmann et al. i.e., containing 22,215 probe sets. For Dong et al. (2005), we see that RSDDA outperforms all of the other methods across different numbers of top genes chosen. The results are summarized in Table 4. Not only does it beat SVM and k-nn, we also see that RSDDA is better than shrinkage-based DLDA which is in turn better than the original DQDA or DLDA. Note that SVM is outperformed by the shrinkage-based methods for top 10, 50, and 200 genes. In the case of top 50 genes, k-nn beats the original DLDA slightly, but performs worse than RSDDA. Overall, RSDDA has smaller misclassification rates than all the other methods.

Table 4

Mean misclassification rates for two binary outcome data sets

Method	Huttmann (2006) 2 classes	Dong (2005) 2 classes
Top 10		
DQDA	0.243	0.198
DLDA	0.227	0.142
SVM	0.248	0.198
k-nn ($k = 3$)	0.235	0.170
SDLDA	0.219	0.142
RSDDA	0.208	0.135
Top 50		
DQDA	0.225	0.167
DLDA	0.192	0.127
SVM	0.218	0.157
k-nn ($k = 3$)	0.252	0.112
SDLDA	0.180	0.117
RSDDA	0.155	0.090
Top 200		
DQDA	0.197	0.132
DLDA	0.185	0.122
SVM	0.185	0.098
k-nn ($k = 3$)	0.253	0.208
SDLDA	0.178	0.115
RSDDA	0.137	0.077

4.3 *Multiple Class Data Sets*

To show how our methods perform on data sets with more than two classes, we consider Pomeroy et al. (2002) and Ross et al. (2000). These data sets contain four classes and eight classes, respectively, see Table 5.

Pomeroy et al. (2002) studied the central nervous system. The number of probe sets in the array is smaller as it is one of the earlier Affymetrix chipsets, Hu6800. It contains four classes, 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumors, and eight primitive neuroectodermal tumors. We can see that the RSDDA method once again beats all of the other methods across the different numbers of genes selected. SVM and k-nn perform poorly when the top 50 and 200 genes are selected.

For the NCI60 data set by Ross et al. (2000), we have eight classes of different tumors from 59 samples with 9703 genes. These are distributed as follows: 7 breast, 6 central nervous system, 7 colon, 8 leukemia, 8 melanoma, 9 nonsmall cell lung carcinoma, 6 ovarian, and 8 renal. This data set contains missing values and it is processed as Dudoit et al. suggested using the nearest neighbor method with $k = 5$ to impute the missing values. We see that RSDDA is similar to the performance of DLDA in this case, both beating SVM and k-nn when the top 50 and 200 genes are chosen. SVM comes pretty close to RSDDA in the top 10 genes scenario.

Table 5
Mean misclassification rates for two multiclass data sets

Method	Pomeroy (2002) 4 classes	Ross (2000) 8 classes
Top 10		
DQDA	0.431	0.507
DLDA	0.383	0.439
SVM	0.399	0.433
k-nn ($k = 3$)	0.399	0.468
SDLDA	0.382	0.432
RSDDA	0.381	0.432
Top 50		
DQDA	0.307	0.399
DLDA	0.236	0.238
SVM	0.276	0.263
k-nn ($k = 3$)	0.309	0.308
SDLDA	0.229	0.239
RSDDA	0.231	0.239
Top 200		
DQDA	0.278	0.399
DLDA	0.239	0.247
SVM	0.288	0.268
k-nn ($k = 3$)	0.258	0.296
SDLDA	0.234	0.242
RSDDA	0.235	0.242

5. Discussion

Microarrays have become a standard tool for biomedical studies. However, the analysis of microarray data still presents statistical challenges as the number of genes is much larger than the number of samples, especially with an ever increasing number of genomic features that can be put on an array, e.g., tiling arrays. Due to cost and in some cases rare diseases, it is not uncommon to see studies with fewer than 10 patients per group. Some researchers have taken the direction of grouping studies together known as meta-analysis (Cahan et al., 2007; Fishel, Kaufman, and Ruppin, 2007). However, this can be difficult as different labs utilize nonmatching gene chips and it is not an easy task to find a good way to combine them. More importantly, the patients from different studies may differ from each other in many aspects. In this article, we have presented novel approaches to performing discriminant analysis from microarray experiments. Our methods bring together shrinkage and regularization to the original diagonal discriminant analyses (DLDA and DQDA), which are known to do well in many discrimination problems.

From the simulated and real data studies, we conclude that RSDDA is a promising classifier for small sample size classification; it performs better than SVM and k-nn in many situations. It improves upon the original DQDA and DLDA through our shrinkage-based methods, SDQDA and SDLDA. The regularization and shrinkage-based approaches introduced in this study have the potential to increase the power of discriminant analysis for which sample sizes are small and there are a large number of features or genes in the microarray setting. We have described the estimation procedure in Section 2 using shrinkage estimator for $1/\sigma^2$ ($t = -1$), but this can also be estimated with σ^2 ($t = 1$). Since the two are very similar in every instance studied, we presented the results for $t = -1$ only.

Of course, it is difficult to predict what the real situation might be for a particular data set, but RSDDA appears to be a good choice for small sample sizes. We suggest using RSDDA unless it becomes computationally infeasible. The good performance of SDLDA on its own though is an indication that RSDDA can do better than the original DQDA and DLDA. We recommend using RSDDA when DQDA and DLDA perform unsatisfactorily as well as for situations where SVM or k-nn is only slightly better than or comparable to DQDA or DLDA.

There are many interesting problems that remain to be addressed, for example, the theoretical justification on why the shrinkage-based method would improve discrimination. This problem can also be extended to gene selection purposes. Moreover, simulations on unbalanced, multiclass, and nonnormal data might be needed to further explore the properties of the newly proposed RSDDA methods. We also see that RSDDA did quite well for a small number of features, e.g., $p = 10$ in real data set. This implies that there is also an opportunity to apply this method in the pathway-based context (Pang et al., 2006). Overall, the new RSDDA method can substantially improve classification accuracy in small sample size situations. RSDDA is not difficult to implement and the corresponding `R` code can be found at the URL specified in the Supplementary Materials.

6. Supplementary Materials

Detailed results of our simulations and real data analysis are given in the Supplementary Materials, referenced in Section 3, and Web Table referenced in Sections 3 and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`. Our software code in `R` is available from the following website `http://bioinformatics.med.yale.edu/rsdda/rsdda.htm`.

Acknowledgements

This work was supported in part by the National Institutes of Health (NIH) grants R01 GM59507, N01 HV28286, P30 DA018343, and U24 NS051869. The majority of the computation was done through the Yale University Biomedical High Performance Computing Center that is supported by the NIH grant RR19895. We thank Matthew Holford for proofreading the paper. We also thank the associate editor and two referees for their comments and suggestions which helped improve the presentation of our work substantially.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. New York: John Wiley.

Barrier, A., Roser, F., Boelle, P., Franc, B., Tse, C., Brault, D., Lacaine, F., Houry, S., Callard, P., Penna, C., Debuire, B., Flahault, A., Dudoit, S., and Lemoine, A. (2007). Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling. *Oncogene* **26,** 2642–2648.

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

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Cahan, P., Rovegno, F., Mooney, D., Newman, J. C., St Laurent, G., and McCaffrey, T. A. (2007). Meta-analysis of microarray results:

Challenges, opportunities, and recommendations for standardization. *Gene* **401,** 12–18.

Cui, X., Hwang, J. T., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6,** 59–75.

Dong, S., Nutt, C. L., Betensky, R. A., Stemmer-Rachamimov, A. O., Denko, N. C., Ligon, K. L., Rowitch, D. H., and Louis, D. N. (2005). Histology-based expression profiling yields novel prognostic markers in human glioblastoma. *Journal of Neuropathology and Experimental Neurology* **64,** 948–955.

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

Fishel, I., Kaufman, A., and Ruppin, E. (2007). Meta-analysis of gene expression data: A predictor-based approach. *Bioinformatics* **23,** 1599–1606.

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

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics* **59,** 992–1000.

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.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8,** 86–100.

Huang, D. S. and Zheng, C. H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22,** 1855–1862.

Huttmann, A., Klein-Hitpass, L., Thomale, J., Deenen, R., Carpinteiro, A., Nückel, H., Ebeling, P., Führer, A., Edelmann, J., Sellmann, L., Dührsen, U., and Dürig, J. (2006). Gene expression signatures separate B-cell chronic lymphocytic leukaemia prognostic subgroups defined by ZAP-70 and CD38 expression status. *Leukemia* **20,** 1774–1782.

Lee, J. W., Lee, J. B., Park, M., and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis* **48,** 869–885.

Montaner, D., Tarraga, J., Huerta-Cepas, J., Burguet, J., Vaquerizas, J., Conde, L., Minguez, P., Vera, J., Mukherjee, S., Valls, J., Pujana, M., Alloza, E., Herrero, J., Al-Shahrour, F., and Dopazom, J. (2006). Next station in microarray data analysis: GEPAS. *Nucleic Acids Research* **34,** W486–W491.

Moon, H., Ahn, H., Kodell, R. L., Lin, C. J., Baek, S., and Chen, J. J. (2006). Classification methods for the development of genomic signatures from high-dimensional data. *Genome Biology* **7,** R121.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E., and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics* **22,** 2028–2036.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., and Golub, T. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415,** 436–442.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C., Rees, C., Spellman, P., Iyer, V., Jeffrey, S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J., Lashkari, D., Shalon, D., Myers, T., Weinstein, J., Botstein, D., and Brown, P. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24,** 227–235.

Shieh, G. S., Jiang, Y. C., and Shih, Y. S. (2006). Comparison of support vector machines to other classifiers using gene expression data. *Communications in Statistics: Simulation and Computation* **35,** 241–256.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association* **102,** 113–122.

Vapnik, V. and Kotz, S. (2006). *Estimation of Dependences Based on Empirical Data*. New York: Springer.

Wald, P. M. and Kronmal, R. A. (1977). Discriminant functions when covariates are unequal and sample sizes are moderate. *Biometrics* **33,** 479–484.

Ye, J., Li, T., Xiong, T. and Janardan, R. (2004). Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1,** 181–190.

Zhan, F., Barlogie, B., Arzoumanian, V., Huang, Y., Williams, D., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Zangari, M., Dhodapkar, M., Shaughnessy, J. Jr. (2007). Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis. *Blood* **109,** 1692–1700.

Received October 2007. *Revised October* 2008.
Accepted October 2008.