UW Biostatistics Working Paper Series

11-28-2005

Optimal Feature Selection for Nearest Centroid Classifiers, With Applications to Gene Expression Microarrays

Alan R. Dabney
University of Washington, adabney@u.washington.edu

John D. Storey
University of Washington, jstorey@u.washington.edu

Optimal Feature Selection for Nearest Centroid Classifiers, With Applications to Gene Expression Microarrays

Alan R. Dabney and John D. Storey

Department of Biostatistics

University of Washington, Seattle, WA 98195

adabney@u.washington.edu, jstorey@u.washington.edu

October 2005

Abstract

Nearest centroid classifiers have recently been successfully employed in high-dimensional applications. A necessary step when building a classifier for high-dimensional data is feature selection. Feature selection is typically carried out by computing univariate statistics for each feature individually, without consideration for how a subset of features performs as a whole. For subsets of a given size, we characterize the optimal choice of features, corresponding to those yielding the smallest misclassification rate. Furthermore, we propose an algorithm for estimating this optimal subset in practice. Finally, we investigate the applicability of shrinkage ideas to nearest centroid classifiers. We use gene-expression microarrays for our illustrative examples, demonstrating that our proposed algorithms can improve the performance of a nearest centroid classifier.

1 Introduction

Linear Discriminant Analysis (LDA) is a long-standing prediction method that has been well characterized when the number of features used for prediction is small (Mardia et al. 1979). The method has recently been shown to compare favorably with more complicated classifiers in high-dimensional applications, where there are thousands of potential features to employ, but only a subset are used (Dudoit et al. 2002, Tibshirani et al. 2002, Lee et al. 2005). In the LDA setting, each class is characterized by its vector of average feature values (i.e., class centroid). A new observation is evaluated by computing the scaled distance between its profile and each class centroid. The observation is then assigned to the class to which it is nearest, allowing LDA to be interpreted as a "nearest centroid classifier."

In high-dimensional applications, it is often desirable to build a classifier using only a subset of features due to the fact that (1) many of the features are not informative for classification and (2) the number of training samples available for building the classifier is substantially smaller than the number of possible features. In any case, it could be argued that a classifier built with a smaller number of features is preferable to an equally accurate classifier built with the complete set of features. Several approaches have been proposed for nearest centroid classifiers that rely on univariate statistics for feature selection (Golub et al. 1999, Hedenfalk et al. 2001, Dudoit et al. 2002, Tibshirani et al. 2002). These methods assess each feature individually by its ability to discriminate the classes. However, it has been noted that the features that best discriminate the classes individually are not necessarily the ones that work best *together* (Jaeger et al. 2003, Dabney 2005).

In the extremely simple case where features are uncorrelated and only two classes exist, it intuitively follows that the optimal set of features are those whose means are most different between

the two classes. However, this intuition does not easily carry over to the more complicated (and topical) case where features are correlated and/or there are more than two classes. For example, if we seek to classify among three classes, it has not been shown whether it is better to choose features that distinguish one class from the other two well or those that distinguish among all three classes well. The role of correlation between features is not currently well understood either.

In this paper, we show how to exactly choose the subset of features of a given size that minimizes the misclassification rate. This optimal feature set takes into account the joint behavior of the features in two ways. First, it explicitly incorporates information about correlation between features. Second, it assesses how a group of features as a whole is capable of distinguishing between multiple classes. Overall, it eliminates the need for heuristically-motivated feature selection criteria in high-dimensional settings, and shows one exactly what the optimal solution is. That is, our main theoretical result allows one to directly aim for the optimal choice. However, one must then *estimate* the optimal choice of features in practice, for which we propose a greedy algorithm and demonstrate its operating characteristics.

Finally, we investigate the application of shrinkage ideas to the classification problem. Shrinkage of multivariate estimates is known to improve their overall accuracy. It has also been suggested that shrinkage can improve nearest centroid classifiers (Tibshirani et al. 2002). However, the shrinkage approach taken in Tibshirani et al. (2002) tends to actually increase misclassification error (Dabney 2005), leaving the question of shrinkage in classifiers open. We show that shrinkage can, in principle, improve classification. Furthermore, we propose a novel shrinkage procedure for nearest centroid classifiers.

Nearest centroid classifiers have been shown to perform well with gene-expression microarrays (Dudoit et al. 2002, Lee et al. 2005, Tibshirani et al. 2002, Dabney 2005), and we illustrate our

findings in this setting. We compare our proposed method with existing nearest centroid classifiers on four previously-published microarray datasets, demonstrating that improvements in prediction accuracy can be attained by estimating the optimal feature-selection criteria (Figure 2). Our estimated optimal-feature selection algorithm can be further improved by employing our shrinkage procedure (Figure 2).

2 Linear Discriminant Analysis (LDA)

The problem LDA addresses is to classify unknown samples into one of K classes. To build a classifier, we obtain n_k training samples per class, $k = 1, 2, \ldots, K$, with m features per sample. For each training sample, we observe class membership Y and profile $\boldsymbol{X}$. For simplicity, we will represent the classes by the numbers $1, 2, \ldots, K$. Note that each profile is a vector of length m. We assume that profiles from class k are distributed as $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, the multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}$. Call $L(\cdot\,;\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ the corresponding probability density function. Finally, let π_k be the prior probability that an unknown sample comes from class k, $k = 1, 2, \ldots, K$.

Bayes' Theorem states that the probability that an observed sample comes from class k is proportional to the product of the class density and prior probability:

$$\Pr(Y = k \mid \boldsymbol{X} = \boldsymbol{x}) \propto L(\boldsymbol{x}\,;\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \times \pi_k. \qquad \text{(a)}$$

We call (a) the posterior probability that sample $\boldsymbol{x}$ comes from class k. LDA assigns the sample

to the class with the largest posterior probability:

$$\hat{y} = \text{argmax}_k \left\{ L(\boldsymbol{x}\,;\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \times \pi_k \right\}. \tag{b}$$

This can be shown to be the rule that minimizes misclassification error (Mardia et al. 1979).

We can rewrite (b) as

$$\hat{y} = \text{argmin}_k \left\{ (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) - 2\log(\pi_k) \right\}. \tag{c}$$

Thus, a sample is assigned to the class to which it is nearest, as measured by the metric $\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}} - 2\log(\pi)$, where $\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}} = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is the square of the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$.

3 Optimal Subsets in Theory

Misclassification rates. A misclassification occurs when a sample is assigned to the incorrect class. The probability of making a classification error is:

$$\Pr(\text{error}) = \sum_{j=1}^{K} \left[\Pr(\hat{Y} \neq j \mid Y = j) \times \pi_j \right].$$

We can derive misclassification rates using the LDA rule (c). In particular, we can calculate misclassification rates for any subset of features. An optimal subset can be found by simply assigning misclassification rates to all possible subsets of a given size and choosing the one with the lowest error rate. Lemma 1 characterizes misclassification rates for a set of centroids, Theorem 1 de-

scribes how to find the optimal subset of a given size, and the Remark shows a simplified result corresponding to equal prior probabilities.

Lemma 1 *Suppose that a sample from class k is distributed according to* $N_m(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, 2, \ldots, K$. *Let* π_k *be the prior probability that a sample comes from class* k, $k = 1, 2, \ldots, K$. *The misclassification rate of a nearest-centroid (LDA) classifier is*

$$\Pr(\text{error}) = \sum_{j=1}^{K} \left\{ \left[1 - \phi\Big(\min_{i \neq j} \left\{ \frac{\|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}}^2 + 2\log(\frac{\pi_j}{\pi_i})}{2\|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}}} \right\} \Big) \right] \times \pi_j \right\}, \tag{d}$$

where ϕ *is the cdf of the standard normal distribution, and* $\|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}}^2 = (\boldsymbol{\mu_j} - \boldsymbol{\mu_i})^T \Sigma^{-1} (\boldsymbol{\mu_j} - \boldsymbol{\mu_i})$ *is the square of the Mahalanobis distance between* $\boldsymbol{\mu_j}$ *and* $\boldsymbol{\mu_i}$.

This allows us to state the following theorem showing how to determine the optimal subset of features of size m_0 for LDA classification.

Theorem 1 *(LDA Optimal Subset Selection) Under the setting described in Lemma 1, the subset of features of size* $m_0 \leq m$ *that minimizes the misclassification rate is the one with the lowest value of equation* (d).

Remark 1 *If* $\pi_1 = \pi_2 = \ldots = \pi_K$, *then*

$$\Pr(\text{error}) = \frac{1}{K} \sum_{j=1}^{K} \left[1 - \phi\Big(\min_{i \neq j} \left\{ \frac{1}{2} \|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}} \right\} \Big) \right].$$

A subset of features can be evaluated using (d), where only the subset of features are included in the centroids. Note that equation (d) can be interpreted as measuring the collective distance between all of the class centroids. In general, the misclassification rate will be small when all of the class

Table 1: Simulated class means with 10 features and 3 classes.

Feature	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\mu_3}$	Score
1	1.25	0.00	0.00	0.35
2	1.50	0.00	0.00	0.50
3	2.00	0.00	0.00	0.89
4	-3.00	0.00	0.00	2.00
5	0.00	1.00	0.00	0.22
6	0.00	1.10	0.00	0.27
7	0.00	-0.90	0.00	0.18
8	0.00	0.00	0.75	0.12
9	0.00	0.00	0.85	0.16
10	0.00	0.00	-0.65	0.09

centroids are far away from each other. Note, however, that the score in (d) is actually a complicated combination of the pairwise differences between the centroids and the class priors. Furthermore, correlations between features are explicitly incorporated through the distance functions $\|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}}$, $i \neq j$. Further intuition into (d) can be attained by considering the following simple example.

A Simple Example. The data in Table 1 represent a simulated example with 10 features and 3 classes. The population means of each class are shown in columns two through four; we assume that each feature has variance 0.5 and that all features are uncorrelated. Suppose that we wish to select the five features that correspond to the lowest misclassification error. The final column of the table lists univariate scores for each feature, where we have used the average squared difference from the feature mean as the score. A high value for a feature on this score indicates large overall differences between this feature's class means. The five largest univariate scores correspond to features 1, 2, 3, 4, and 6.

An alternative approach to using univariate scores to select features is to consider all $\binom{10}{5} = 252$ possible quintuplets and choose the set with the lowest overall misclassification rate. Note that, to do this, we must be able to assign misclassification probabilities to arbitrary feature subsets.

This highlights the novelty and usefulness of the multivariate score (d). Using (d), we find that the set of features chosen by the univariate scores has an overall misclassification rate of 15%. Similarly, we find that the optimal set in this example contains features 4, 5, 6, 7, and 9, with an associated error rate of 6%. The most obvious difference between this subset and that chosen by univariate scores is the exclusion of features 1, 2, and 3. Apparently, class one can be sufficiently characterized by feature 4. The other features do not contain sufficient *additional* information to merit their selection.

Correlation between features. An important aspect of the optimal feature-selection procedure is its explicit incorporation of correlation between features. It is not necessarily clear what effect correlation between features should have on a classifier. Some have argued that it is inefficient to select correlated features (Jaeger et al. 2003). Bickel & Levina (2004) show that estimating correlations as zero can lead to better prediction when the number of features is large relative to the number of samples. Meanwhile, recent optimality results in the context of multiple hypothesis testing (Storey 2005, Storey et al. 2005) suggest that correlated features may be beneficial in distinguishing groups. Intuitively, many weakly informative, correlated genes might be expected to collectively be highly informative.

We briefly investigate the effect of different correlation patterns in the example of Table 1. Let $\mathbf{\Sigma}$ be the 10×10 covariance matrix. In Table 2, we refer to the 4×4 diagonal block corresponding to features $1-4$ as "Block 1." The diagonal blocks corresponding to features $5-7$ and $8-10$ are similarly referred to as "Block 2" and "Block 3." "Block 1$\sim$2" refers to the off-diagonal block relating features $1-4$ to features $5-6$, *etc.*. In all cases, we include a common pairwise correlation of 0.9; no qualitative differences were found when considering negative correlation.

Table 2: The effect of covariance on the optimal feature-selection procedure.

Covariance	Selected Features	$\widehat{\Pr}$(error)
None	4, 5, 6, 7, 9	6.2%
Block 1	4, 5, 6, 7, 9	6.2%
Block 2	2, 3, 4, 6, 7	0.2%
Block 3	3, 4, 8, 9, 10	0.4%
Block 1 $\sim$ 2	4, 5, 6, 7, 9	6.3%
Block 1 $\sim$ 3	4, 5, 6, 7, 9	6.6%
Block 2 $\sim$ 3	4, 5, 6, 8, 9	4.2%

In this example, correlation between features in Blocks 2 and 3 has the largest effect, on both the optimal set of features and the probability of misclassification. The features in Blocks 2 and 3 are individually less informative than those in Block 1. Note, in particular, that when the features in Block 3 are correlated, all three components are selected, whereas only one is selected in the absence of correlation. These results suggest that correlated features can be useful together, particularly when the correlated features contain relatively little information individually. More generally, there are many possible scenarios in which correlation could play a role. The main point is that the feature-selection procedure (d) automatically identifies the optimal combination of features, even in the presence of correlation.

4 Optimal Subsets in Practice

Often, there will be many more than 10 features from which to choose. With gene expression microarrays, for example, there are thousands of genes under consideration. Thus, in practice, it may be impossible to perform an exhaustive search for a best subset. Even if we could perform an exhaustive search, we would still need to estimate the class centroids, and this may not lead to the correct solution. We propose a greedy algorithm for estimating the optimal subset of given size

when an exhaustive search is not possible. We compare our algorithm with a more conventional univariate scoring algorithm for choosing subsets on both simulated (Figure 1) and real (Figure 2) datasets.

Univariate Scoring Procedures. The novelty of our proposed method can better be understood by first considering some of the available methods for choosing features, which are all based on univariate scoring of features. An early suggestion in the context of gene expression microarrays used F-statistics to evaluate each feature (here, a gene) individually on the basis of comparisons of between-class variation and within-class variation (Dudoit et al. 2002):

$$F_i = \frac{BSS}{WSS} = \frac{\sum_{j=1}^{n}\sum_{k=1}^{K}\mathbf{I}(y_j = k)(\hat{\mu}_{ik} - \hat{\mu}_i)^2}{\sum_{j=1}^{n}\sum_{k=1}^{K}\mathbf{I}(y_j = k)(x_{ij} - \hat{\mu}_{ik})^2},$$

$i = 1, 2, \ldots, m$. To form a classifier using only $\tilde{m}$ features, class centroids are formed with only the features corresponding to the largest $\tilde{m}$ F-statistics. All other features are discarded.

Another example (again in the context of microarrays) is Prediction Analysis of Microarrays (PAM) (Tibshirani et al. 2002). Instead of F-statistics, PAM uses the statistics

$$d_{ik} = \frac{\hat{\mu}_{ik} - \hat{\mu}_i}{w_k(s_i + s_0)}$$

to select genes, where s_i is the pooled standard deviation for feature i, $w_k = (1/n_k - 1/n)^{1/2}$ makes $w_k \times s_i$ equal to the standard error of the numerator, and s_0 is a fudge factor intended to guard against very large statistics for very small standard errors. Without s_0, d_{ik} is a t-statistic comparing the mean of feature i in class k to the overall mean of feature i. Hence, d_{ik} measures the difference between feature i in class k and feature i in all classes combined. PAM then shrinks the d_{ik}'s

toward zero, eliminating the features that do not provide sufficient discriminatory information. The Classification to Nearest Centroids (ClaNC) method (Dabney 2005) uses simple t-statistics without fudge factors or shrinkage and outperforms PAM on both simulated and real datasets.

Algorithm for Estimating the Optimal Solution. When it is not feasible to perform an exhaustive search for the best subset of features of a given size, we propose the following greedy algorithm. Let $\boldsymbol{s}$ be the indices of a subset of features, and let $\boldsymbol{\mu^s}$ denote a centroid indexed by $\boldsymbol{s}$. Define the estimated scoring function to be

$$\widehat{\Pr(\text{error}\,;\,\boldsymbol{s})} = \sum_{j=1}^{K} \left\{ \left[1 - \phi\Big(\min_{i \neq j} \left\{ \frac{\|\hat{\boldsymbol{\mu}}_{\boldsymbol{j}}^{\boldsymbol{s}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{s}}\|_{\hat{\boldsymbol{\Sigma}}}^2 + 2\log(\frac{\pi_j}{\pi_i})}{2\|\hat{\boldsymbol{\mu}}_{\boldsymbol{j}}^{\boldsymbol{s}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{s}}\|_{\hat{\boldsymbol{\Sigma}}}} \right\} \Big) \right] \times \pi_j \right\}. \tag{e}$$

1. Evaluate all m features individually using (e), and select the first feature to be the one with the lowest score. This leaves $m-1$ features from which to choose.

2. Combine each remaining feature with the one already chosen and compute the $m-1$ scores corresponding to each invidual *pair*.

3. Choose the second feature to be the one that produces the lowest score when combined with the first.

4. Continue this process until the desired number of features have been selected.

Note that $\hat{\boldsymbol{\Sigma}}$ in (e) is a matrix of estimated covariances. As we saw with the examples in Tables 1 and 2, classification accuracy can be improved by incorporating correlation between features. However, in high-dimensional settings, it may be impractical and/or unnecessary (Bickel & Levina 2004) to estimate large covariance matrices for large numbers of features. This is one motivation behind shrinking covariances to zero when building classifiers for microarrays. In the microarray

examples below, the classifiers we present do not include estimated covariances. There was no gain in accuracy in these examples when estimating covariances by maximum likelihood (results not shown). This could either mean covariances are truly irrelevant in these examples, or that our covariance estimates are imprecise. In future work, we plan to investigate the utility of shrinkage estimates of the covariance matrix; it is desirable to include such information, since it is fully characterized by Theorem 1.

Illustration on Simulated Example. For each of 50 simulations, we generated random observations from the setting described in Table 1, assuming a common pairwise correlation of 0.5 within each of Blocks 1, 2, and 3, as described in Table 2. In each simulation, 15 training samples (five per class) and 15 test samples (again, five per class) were generated. We then compared the misclassification rates (computed on the test data) after selecting subsets of size five in different ways (on the training data), with the results shown as boxplots in Figure 1. Our standard of comparison was the classifier built using the optimal subset found above, labeled "Optimal." We also carried out an exhaustive search for the optimal subset in each simulation; these results are labeled "Estimated Optimal." "Greedy Approximation" refers to the algorithm described above. Finally, the "Univariate" results correspond to choosing subsets based on univariate scores. In this comparison, we used the F-statistic approach of Dudoit et al. (2002).

Figure 1 about here.

The true optimal subset of size five produces the most accurate classifier, as expected. The exhaustive search apparently did not identify the true optimal subset, although its choice was superior to that made by the univariate scoring procedure. The greedy algorithm estimates the optimal feature subset well in this example.

Illustration on Real Examples. We now illustrate our methods on four previously published gene-expression microarray experiments. In each analysis, any missing values were imputed using K-nearest-neighbors (Troyanskaya et al. 2001) with $K = 10$. We compare the methods on the basis of error rates from five-fold cross-validation. We avoid gene-selection bias by completely rebuilding classifiers to identical specifications in each cross-validation iteration (Ambroise & McLachlan 2002). Cross-validated error rates are nearly unbiased, being slightly conservative (Ambroise & McLachlan 2002, Hastie et al. 2001), and they are thus sufficient for comparing classifiers.

The first example involves small round blue cell tumors (SRBCT) of childhood (Khan et al. 2001). Expression measurements were made on 2,307 genes in 83 SRBCT samples. The tumors were classified as Burkitt lymphoma, Ewing sarcoma, neuroblastoma, or rhabdomyosarcoma. There are 11, 29, 18, and 25 samples in each respective class. In the second example, expression measurements were made on 4,026 genes in 58 lymphoma patients (Alizadeh et al. 2000). The tumors were classified as diffuse large B-cell lymphoma and leukemia, follicular lymphoma, and chronic lymphocytic leukemia. There are 42, 6, and 10 samples in each respective class. The third example involves the cell lines used in the National Cancer Institute's screen for anti-cancer drugs (Ross et al. 2000, Scherf et al. 2000). Expression measurements were made on 6,830 genes in 60 cell tumors. There are representative cell lines for each of lung cancer, prostate cancer, CNS, colon cancer, leukemia, melanoma, NSCLC, ovarian cancer, renal cancer, and one unknown sample. We filtered out 988 genes for which 20% or more of the tumors had missing values. We also excluded samples from prostate cancer (due to there being only two samples) and the one unknown sample. There are 9, 6, 7, 6, 8, 7, 6, and 8 samples in each remaining respective class. The fourth example involves acute myeloid leukemia (AML) and acute lymphblastic leukemia (ALL) (Golub et al. 1999). The public version of the training data used in the original analysis include expression measurements

on 3,857 genes in 38 leukemia patients. There are 11 and 27 samples in each respective class.

Figure 2 about here.

Figure 2 shows misclassification rates for a variety of subset sizes. The estimated optimal subsets using the algorithm described above compare favorably with both PAM and ClaNC. That is, for any given number of active features, classifiers built with estimated optimal subsets have error rates that tend to be smaller than (or equal to) those for PAM or ClaNC.

Shrinkage. As mentioned earlier, the LDA classification rule minimizes the misclassification rate, because the LDA rule equals the Bayes' rule under its assumptions (Mardia et al. 1979). For simplicity in motivating shrinkage, assume all features are independent with variance one, and that each class has equal prior probability. Then the Bayes' rule is to classify a new sample to the class for which $\sum_{i=1}^{m}(x_i^* - \mu_{ik})^2$ is smallest. However, we must estimate the centroids $\boldsymbol{\mu}_k$ in practice, using $\hat{\boldsymbol{\mu}}_k$ in their place. Suppose $\boldsymbol{x^*}$ comes from class k_0. Then, expanding the squared distance between $\boldsymbol{x^*}$ to class k_0 and taking expectations, we have

$$\begin{aligned} \mathrm{E}\sum_{i=1}^{m}(x_i^* - \hat{\mu}_{ik_0})^2 &= \mathrm{E}\sum_{i=1}^{m}(x_i^* - \mu_{ik_0} + \mu_{ik_0} - \hat{\mu}_{ik_0})^2 \\ &= \mathrm{E}\sum_{i=1}^{m}(x_i^* - \mu_{ik_0})^2 + \mathrm{E}\sum_{i=1}^{m}(\mu_{ik_0} - \hat{\mu}_{ik_0})^2, \end{aligned} \tag{f}$$

or the Bayes' rule plus the mean squared error (MSE) of the centroid estimate.

Reducing the MSE of $\hat{\boldsymbol{\mu}}_{\boldsymbol{k_0}}$ will bring us closer to the Bayes' rule. According to Stein's Paradox of statistics (Stein 1956), we can reduce the MSE of $\hat{\boldsymbol{\mu}}_{\boldsymbol{k_0}}$ by shrinking towards the overall mean $\sum_{i=1}^{m}\hat{\mu}_{ik}/m$ (or any other constant). In our setting, this suggests shrinking each centroid across its m components. The PAM method (Tibshirani et al. 2002) incorporates shrinkage ideas. However, PAM shrinks each feature across classes. Since this makes the estimated class centroids more similar

to one another, the shrinkage employed by PAM appears to actually increase misclassification error (Dabney 2005). We now investigate whether shrinking centroids across features can improve classification.

While there are many possible approaches to shrinking the centroids, we take the following simple approach. Let $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}$ be the m-vector with each component equal to $\hat{\mu}_k^o = \sum_{i=1}^m \hat{\mu}_{ik}/m$, $k = 1, 2, \ldots, K$. We consider shrunken centroids of the form

$$\tilde{\boldsymbol{\mu}}_k = \omega_k \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}} + (1 - \omega_k)\hat{\boldsymbol{\mu}}_k, \tag{g}$$

$k = 1, 2, \ldots, K$. We choose ω_k so that $\mathrm{E}(\|\boldsymbol{\mu}_{\boldsymbol{k}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}\|_{\boldsymbol{\Sigma}}^2)$ is minimized, $k = 1, 2, \ldots, K$. Theorem 2 presents a general result, and the Corollary translates this result to the ideal case where all genes are independent.

Theorem 2 *Let the maximum likelihood estimates of the class centroids be $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}$ be the m-vector with each component equal to the estimated overall mean for class k, $k = 1, 2, \ldots, K$. Among all estimators of the form $\tilde{\boldsymbol{\mu}}_k = \omega_k \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}} + (1 - \omega_k)\hat{\boldsymbol{\mu}}_k$, the error measure $E(\|\boldsymbol{\mu}_{\boldsymbol{k}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}\|_{\boldsymbol{\Sigma}}^2)$ is minimized by choosing*

$$\hat{\omega}_k = \frac{V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k) - C_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k,\, \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}})}{(\bar{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) + V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k) + V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}) - 2C_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k,\, \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}})},$$

where $\bar{\boldsymbol{\mu}}_k$ is the m-vector with each component equal to $\bar{\mu}_k = \frac{1}{m}\sum_{i=1}^m \mu_{ik}$, $V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k) = E(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)$, $V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}) = E(\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}} - \bar{\boldsymbol{\mu}}_k)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}} - \bar{\boldsymbol{\mu}}_k)$, and $C_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}) = E(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}} - \bar{\boldsymbol{\mu}}_k)$, $k = 1, 2, \ldots, K$.

Corollary 1 *Assuming all genes are independent,*

$$\hat{\omega}_k = \frac{m-1}{n_k \sum_{i=1}^{m} \frac{(\bar{\mu}_k - \mu_{ik})^2}{\sigma_i^2} + m - 2 + (\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\sigma_i^2})(\frac{1}{m}\sum_{i=1}^{m}\sigma_i^2)},$$

$k = 1, 2, \ldots, K$.

In practice, we use plug-in estimates for the unknown parameters.

Figure 2 also illustrates the performance of classifiers with estimated optimal subsets and shrunken centroids. The shrinkage was carried out on the complete centroids, prior to the feature-selection step. Another alternative would be to shrink the centroids of each considered feature subset individually. No qualitative differences were found when employing this shrinkage approach (results not shown). Improvements are evident over the other classifiers under consideration. This suggests that shrinkage can be successfully applied to nearest centroid classifiers and is worth further exploration.

5 Proofs of Theorems

Proof of Lemma 1. We begin with the simple two-class case. A classification error is made if we either predict class two for a sample that truly comes from class one or predict class one for a sample that truly comes from class two. That is,

$$\Pr(\text{error}) = \Pr(\hat{Y} = 2 \mid Y = 1) \times \pi_1 + \Pr(\hat{Y} = 1 \mid Y = 2) \times \pi_2. \qquad \text{(h)}$$

Consider the first component of this sum. We can rewrite it as

$$\begin{aligned}\Pr(\hat{Y} = 2 \mid Y = 1) &= 1 - \Pr(\hat{Y} = 1 \mid Y = 1) \\ &= 1 - \Pr\Big(L(\boldsymbol{X}; \boldsymbol{\mu_1}, \boldsymbol{\Sigma}) \times \pi_1 > L(\boldsymbol{X}; \boldsymbol{\mu_2}, \boldsymbol{\Sigma}) \times \pi_2\Big),\end{aligned}$$

where $L(\cdot\,;\,\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Continuing,

$$\begin{aligned}\Pr(\hat{Y} = 2 \mid Y = 1) &= 1 - \Pr\Bigg(\exp\left\{-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu_1})\right\} \times \pi_1 \\ &\qquad > \exp\left\{-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu_2})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu_2})\right\} \times \pi_2 \mid Y = 1\Bigg) \\ &= 1 - \Pr\Bigg((\boldsymbol{X} - \boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu_1}) - 2\log(\pi_1) \\ &\qquad < (\boldsymbol{X} - \boldsymbol{\mu_2})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu_2}) - 2\log(\pi_2) \mid Y = 1\Bigg) \\ &= 1 - \Pr\Bigg(2(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}X \\ &\qquad < \boldsymbol{\mu_2}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} - \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + 2\log(\frac{\pi_1}{\pi_2}) \mid Y = 1\Bigg).\end{aligned}$$

Using the fact that $\boldsymbol{X} \sim N(\boldsymbol{\mu_1},\, \boldsymbol{\Sigma})$, we can derive the distribution of the random variable $2(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ as $N(2(\boldsymbol{\mu_2}-\boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1}, 4\|\boldsymbol{\mu_1}-\boldsymbol{\mu_2}\|^2_{\boldsymbol{\Sigma}})$, where $\|\boldsymbol{\mu_1}-\boldsymbol{\mu_2}\|_{\boldsymbol{\Sigma}} = \big[(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\big]^{\frac{1}{2}}$ is the Mahalanobis distance between $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$. So then,

$$\Pr(\hat{Y} = 2 \mid Y = 1) = 1 - \Pr\Big(Z < \frac{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|^2_{\boldsymbol{\Sigma}} + 2\log(\frac{\pi_1}{\pi_2})}{2\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|_{\boldsymbol{\Sigma}}}\Big),$$

where $Z \sim N(0,1)$. Letting $\phi(\cdot)$ be the standard normal probability distribution function ($\phi(x) = \Pr(Z \leq x)$),

$$\Pr(\hat{Y} = 2 \mid Y = 1) = 1 - \phi\Big(\frac{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|^2_{\boldsymbol{\Sigma}} + 2\log(\frac{\pi_1}{\pi_2})}{2\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|_{\boldsymbol{\Sigma}}}\Big).$$

Equation (h) is then

$$\Pr(\text{error}) = \left[1 - \phi\Big(\frac{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|^2_{\boldsymbol{\Sigma}} + 2\log(\frac{\pi_1}{\pi_2})}{2\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|_{\boldsymbol{\Sigma}}}\Big)\right] \times \pi_1 + \left[1 - \phi\Big(\frac{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|^2_{\boldsymbol{\Sigma}} + 2\log(\frac{\pi_2}{\pi_1})}{2\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|_{\boldsymbol{\Sigma}}}\Big)\right] \times \pi_2. \quad \text{(i)}$$

Now suppose there are $K \geq 2$ classes. Then,

$$\Pr(\text{error}) = \sum_{j=1}^{K} \Big[\Pr(\hat{Y} \neq j \mid Y = j) \times \pi_j\Big]. \quad \text{(j)}$$

Again beginning with the first component of the sum, and proceding as above,

$$\begin{aligned}
\Pr(\hat{Y} \neq 1 \mid Y = 1) &= 1 - \Pr(\hat{Y} = 1 \mid Y = 1) \\
&= 1 - \Pr(L(\boldsymbol{X}; \boldsymbol{\mu_1}, \boldsymbol{\Sigma}) \times \pi_1 > \max_{i \neq 1} \{L(\boldsymbol{X}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma}) \times \pi_i\}) \\
&= 1 - \phi\Big(\min_{i \neq 1} \left\{\frac{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_i}\|^2_{\boldsymbol{\Sigma}} + 2\log(\frac{\pi_1}{\pi_i})}{2\|\boldsymbol{\mu_1} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}}}\right\}\Big).
\end{aligned}$$

The remaining components of (j) are analagous, and

$$\Pr(\text{error}) = \sum_{j=1}^{K} \left\{\left[1 - \phi\Big(\min_{i \neq j} \left\{\frac{\|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|^2_{\boldsymbol{\Sigma}} + 2\log(\frac{\pi_j}{\pi_i})}{2\|\boldsymbol{\mu_j} - \boldsymbol{\mu_i}\|_{\boldsymbol{\Sigma}}}\right\}\Big)\right] \times \pi_j\right\}. \quad \text{(k)}$$

Proof of Theorem 1. Let $\boldsymbol{s_0}$ be the subset of $m_0 \leq m$ features with the lowest value of equation (d), and let $\boldsymbol{s_0}'$ be any other subset of m_0 features. By Lemma 1, the misclassification rate of the nearest-centroid classifier based on $\boldsymbol{s_0}'$ is greater than or equal to that of the classifier based on $\boldsymbol{s_0}$.

Proof of Theorem 2. We begin by writing

$$\begin{aligned}\mathrm{E}(\|\boldsymbol{\mu_k}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}\|^2_{\boldsymbol{\Sigma}}) &= \mathrm{E}\left\{\boldsymbol{\mu}_k-\left[\omega_k\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}+(1-\omega_k)\hat{\boldsymbol{\mu}}_k\right]\right\}^T\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{\mu}_k-\left[\omega_k\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}+(1-\omega_k)\hat{\boldsymbol{\mu}}_k\right]\right\}\\ &= \mathrm{E}(\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)-2\omega_k\mathrm{E}(\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)\\ &\quad+\omega^2\mathrm{E}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k),\end{aligned}$$

$k=1,2,\ldots,K$. Differentiating with respect to ω_k gives

$$\hat{\omega}_k=\frac{\mathrm{E}(\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)}{\mathrm{E}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)},$$

$k=1,2,\ldots,K$. Note that

$$\begin{aligned}\mathrm{E}(\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k) &= -\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)\\ &= \mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_k-\bar{\boldsymbol{\mu}}_k)-\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k)\\ &= \mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)-\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k),\end{aligned}$$

$k=1,2,\ldots,K$. Also,

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\hat{\boldsymbol{\mu}}_k) &= \mathrm{E}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k+\bar{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k+\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k+\bar{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k+\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k)\\ &= \mathrm{E}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k)+(\bar{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)\\ &\quad+\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)-2\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-\bar{\boldsymbol{\mu}}_k),\end{aligned}$$

$k=1,2,\ldots,K$. Letting $V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k)=\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)$, $C_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k,\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}})=\mathrm{E}(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^{\boldsymbol{o}}_{\boldsymbol{k}}-$

$\bar{\boldsymbol{\mu}}_k$) (and so on), we can now write

$$\hat{\omega}_k = \frac{V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k) - C_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k,\, \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}})}{(\bar{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) + V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k) + V_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}) - 2C_{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}_k,\, \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}})},$$

$k = 1, 2, \ldots, K$.

Proof of Corollary 1. Assuming all genes are independent ($\boldsymbol{\Sigma} = \mathtt{diag}(\boldsymbol{\sigma}^2)$),

$$\begin{aligned}
V_{\boldsymbol{\sigma}}(\hat{\boldsymbol{\mu}}_k) &= \mathrm{E} \sum_{i=1}^{m} \frac{1}{\sigma_i^2} (\hat{\mu}_{ik} - \mu_{ik})^2 = \frac{m}{n_k} \\
V_{\boldsymbol{\sigma}}(\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}) &= \mathrm{E} \sum_{i=1}^{m} \frac{1}{\sigma_i^2} (\hat{\mu}_k^o - \bar{\mu}_k)^2 = \frac{1}{n_k} (\frac{1}{m} \sum_{i=1}^{m} \frac{1}{\sigma_i^2}) (\frac{1}{m} \sum_{i=1}^{m} \sigma_i^2) \\
C_{\boldsymbol{\sigma}}(\hat{\boldsymbol{\mu}}_k,\, \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{\boldsymbol{o}}) &= \mathrm{E} \sum_{i=1}^{m} \frac{1}{\sigma_i^2} (\hat{\mu}_{ik} - \mu_{ik}) (\hat{\mu}_k^o - \bar{\mu}_k) = \frac{1}{n_k},
\end{aligned}$$

$k = 1, 2, \ldots, K$. Hence,

$$\hat{\omega}_k = \frac{m - 1}{n_k \sum_{i=1}^{m} \frac{(\bar{\mu}_k - \mu_{ik})^2}{\sigma_i^2} + m - 2 + (\frac{1}{m} \sum_{i=1}^{m} \frac{1}{\sigma_i^2}) (\frac{1}{m} \sum_{i=1}^{m} \sigma_i^2)},$$

$k = 1, 2, \ldots, K$.

6 Discussion

We have derived the theoretically optimal subset of a given size for a nearest centroid classifier. We have also considered the estimation of this optimal subset. Although an exhaustive search would be ideal, it is not practical in many settings. We have thus proposed a greedy algorithm for estimating optimal subsets and demonstrated that our algorithm can produce more accurate classifiers in both simulated and real applications. However, it is likely that improvements can be

made to the algorithm. Furthermore, we did not have success in estimating covariance matrices in the examples considered. Improvements in classification accuracy may be possible in other settings or by other procedures for estimating covariances.

We have also considered the applicability of shrinkage ideas to nearest centroid classifiers. In particular, we have considered classifiers built with centroids that have been shrunken across their features. Using a novel shrinkage procedure, it was demonstrated on several previously published datasets that shrunken centroids can improve prediction accuracy, making this approach worth exploring further.

The algorithms described here have been implemented in the point-and-click Classification to Nearest Centroids (ClaNC) software, available from the authors' websites.

Acknowledgements

This research was supported in part by the Cancer-Epidemiology and Biostatistics Training Grant 5T32CA009168-29 and NIH grant 1 U54 GM2119-03.

References

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Jr., J. H., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**: 503–511.

Ambroise, C. & McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data, *Proceedings of the National Academy of Sciences* **99**: 6562–6566.

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

Dabney, A. (2005). Classification of microarrays to nearest centroids, *Bioinformatics*, In press.

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

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

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O., Wilfond, B., Borg, A. & Trent, J. (2001). Gene expression profiles in hereditary breast cancer, *New England Journal of Medicine* **344**: 539–548.

Jaeger, J., Sengupta, R. & Ruzzo, W. (2003). Improved gene selection for classification of microarrays, *Pac. Symp. Biocomput.* pp. 53–64.

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C. & Meltzer, P. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine* **7**: 673–679.

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

Mardia, K., Kent, J. & Bibby, J. (1979). *Multivariate Analysis*, Academic Press, London.

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

Scherf, U., Ross, D., Waltham, M., Smith, L., Lee, J., Tanabe, L., Kohn, K., Reinhold, W., Myers, T., Andrews, D., Scudiero, D., Eisen, M., Sausville, E., Pommier, Y., Botstein, D., Brown, P. & Weinstein, J. (2000). A gene expression database for the molecular pharmacology of cancer, *Nature Genetics* **24**: 236–244.

Stein, C. (1956). Inadmissability of the usual estimator for the mean of a multivariate distribution., *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**: 197–206.

Storey, J. (2005). The optimal discovery procedure I: A new approach to simultaneous significance testing, *Technical Report 259*, Department of Biostatistics, University of Washington, Seattle.

Storey, J., Dai, J. & Leek, J. (2005). The optimal discovery procedure II: Applications to comparative microarray experiments, *Technical Report 260*, Department of Biostatistics, University of Washington, Seattle.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences* **99**: 6567–6572.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics* **17**: 520–525.

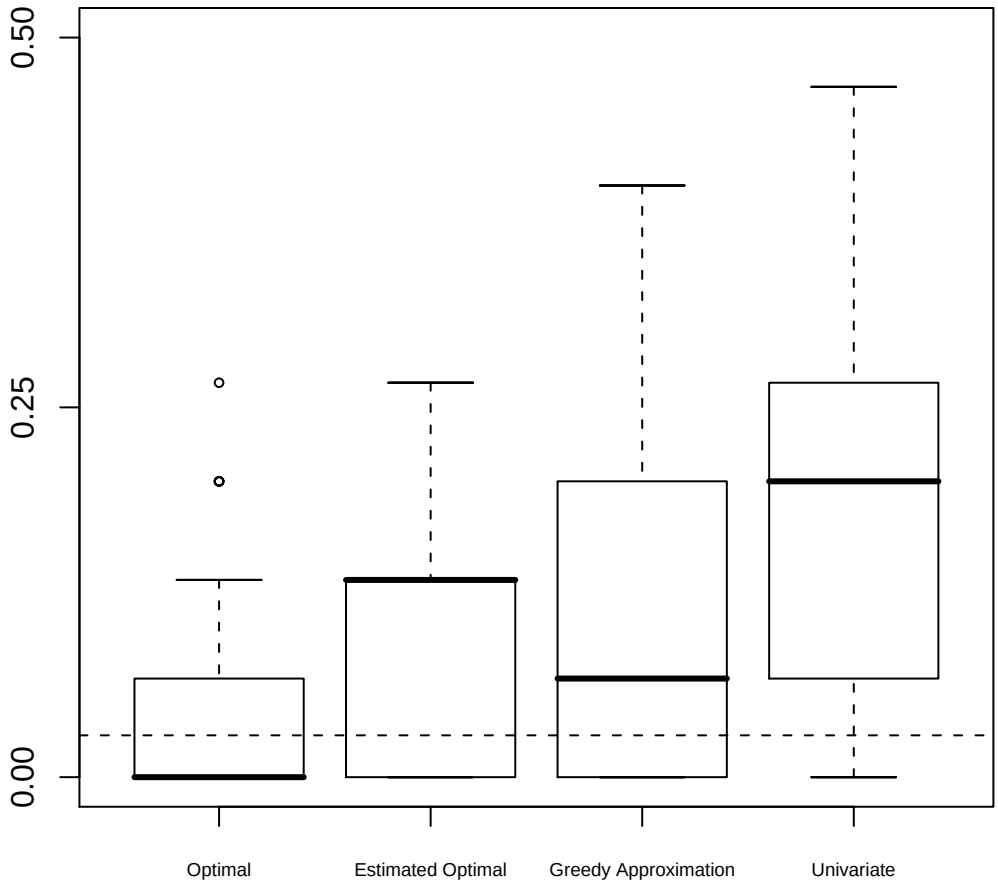

Figure 1: Comparisons of classification error rates on simulated data using different methods for choosing subsets of five from 10 features. "Optimal" uses the (unknown in practice) true optimal subset of five features. "Estimated Optimal" uses the subset derived by an exhaustive search of the data. "Greedy Approximation" uses our greedy algorithm instead of an exhaustive search. "Univariate" uses F-statistics to score each feature individually, without respect for how subsets work *jointly*. The horizontal dashed line indicates the true misclassification error rate using the optimal subset.

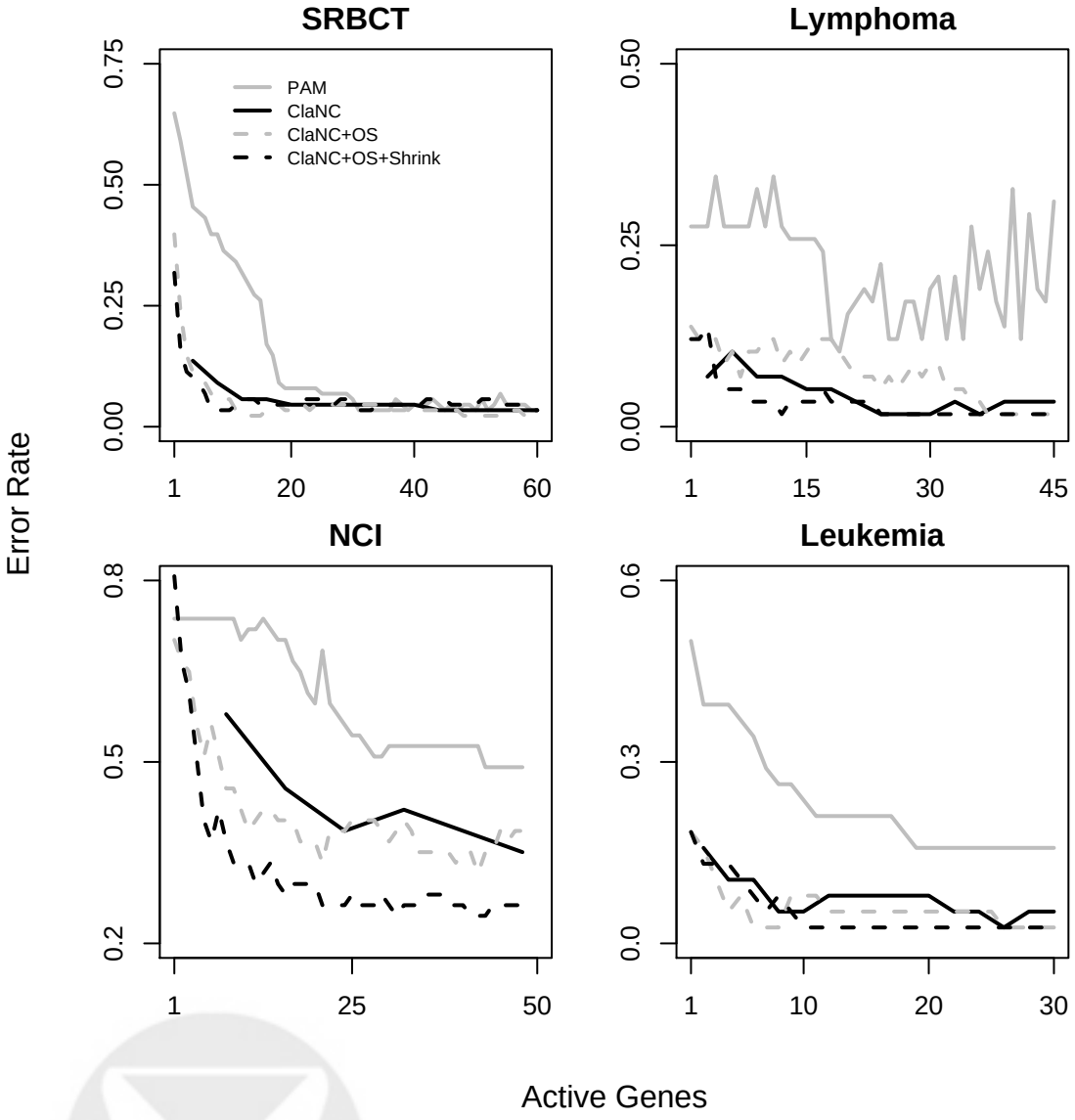

Figure 2: Comparisons of classification error rates on four previously published microarray datasets using different methods for choosing subsets of given size. The number of features (here, genes) are shown on the x-axis, and misclassification error rates are shown on the y-axis. PAM is the Prediction Analysis of Microarrays method. ClaNC is the Classification to Nearest Centroids method. The "OS" after "ClaNC" indicates that optimal subsets have been estimated using our algorithm. "Shrink" indicates that the class centroids have been shrunken.