Statistical Science
2003, Vol. 18, No. 1, 104–117

Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan and Gilbert Chu

Abstract. We propose a new method for class prediction in DNA microarray studies based on an enhancement of the nearest prototype classifier. Our technique uses "shrunken" centroids as prototypes for each class to identify the subsets of the genes that best characterize each class. The method is general and can be applied to other high-dimensional classification problems. The method is illustrated on data from two gene expression studies: lymphoma and cancer cell lines.

Key words and phrases: Sample classification, gene expression arrays.

1. INTRODUCTION

Class prediction with high-dimensional features is an important problem and has recently received a great deal of attention in the context of DNA microarrays. The task is to classify and predict the diagnostic category of a sample, based on its gene expression profile. Recent proposals for this problem include Golub et al. (1999), Hedenfalk et al. (2001), Hastie, Tibshirani, Botstein and Brown (2001) and the artificial neural network approach in Khan et al. (2001).

The microarray problem is a unique and challenging classification task because there are a large number of inputs (genes) from which to predict classes and a relatively small number of samples. It is especially important to identify which genes contribute toward the classification. This can aid in biological understanding of the disease process and is also important in development of clinical tests for early diagnosis. In this article we propose a simple approach to the problem that performs well and is easy to understand and interpret.

As an example, we consider data from Alizadeh et al. (2000), which is available from the authors' web site. These data consist of expression measurements on 4,026 genes from samples of 59 lymphoma patients. The samples are classified into diffuse large B-cell lymphoma and leukemia (DLCL), follicular lymphoma (FL) and chronic lymphocytic leukemia (CLL). We selected a random subset of 20 samples and set them aside as a test set; the remaining 39 samples formed the training set.

We began with a nearest centroid classification. Figure 1 (light grey bars) shows the training-set centroids (average expression of each gene) within each of the three classes. The overall average expression of the corresponding gene has been subtracted, so that these values are differences from the overall centroid.

To apply the nearest centroid classification, we take the gene expression profile of the test sample and compute its squared distance from each of the three class centroids. The predicted class is the one whose centroid is closest to the expression profile of the test sample. This procedure makes zero errors on the 20 test samples, but has the major drawback that it uses all 4,026 genes.

Robert Tibshirani is Professor, Departments of Health Research and Policy, and Statistics, Stanford University, Stanford, California 94305-5405 (e-mail: tibs@stat.stanford.edu). Trevor Hastie is Professor, Departments of Statistics, and Health Research and Policy, Stanford University, Stanford, California 94305-5405 (e-mail: hastie@stat.stanford.edu). Balasubramanian Narasimhan is Senior Research Associate, Departments of Statistics, and Health Research and Policy, Stanford University, Stanford, California 94305-5405 (e-mail: naras@stat.stanford.edu). Gilbert Chu is Professor, Departments of Biochemistry, and Medical Oncology, Stanford University, Stanford, California 94305-5151 (e-mail: chu@cmgm.stanford.edu).

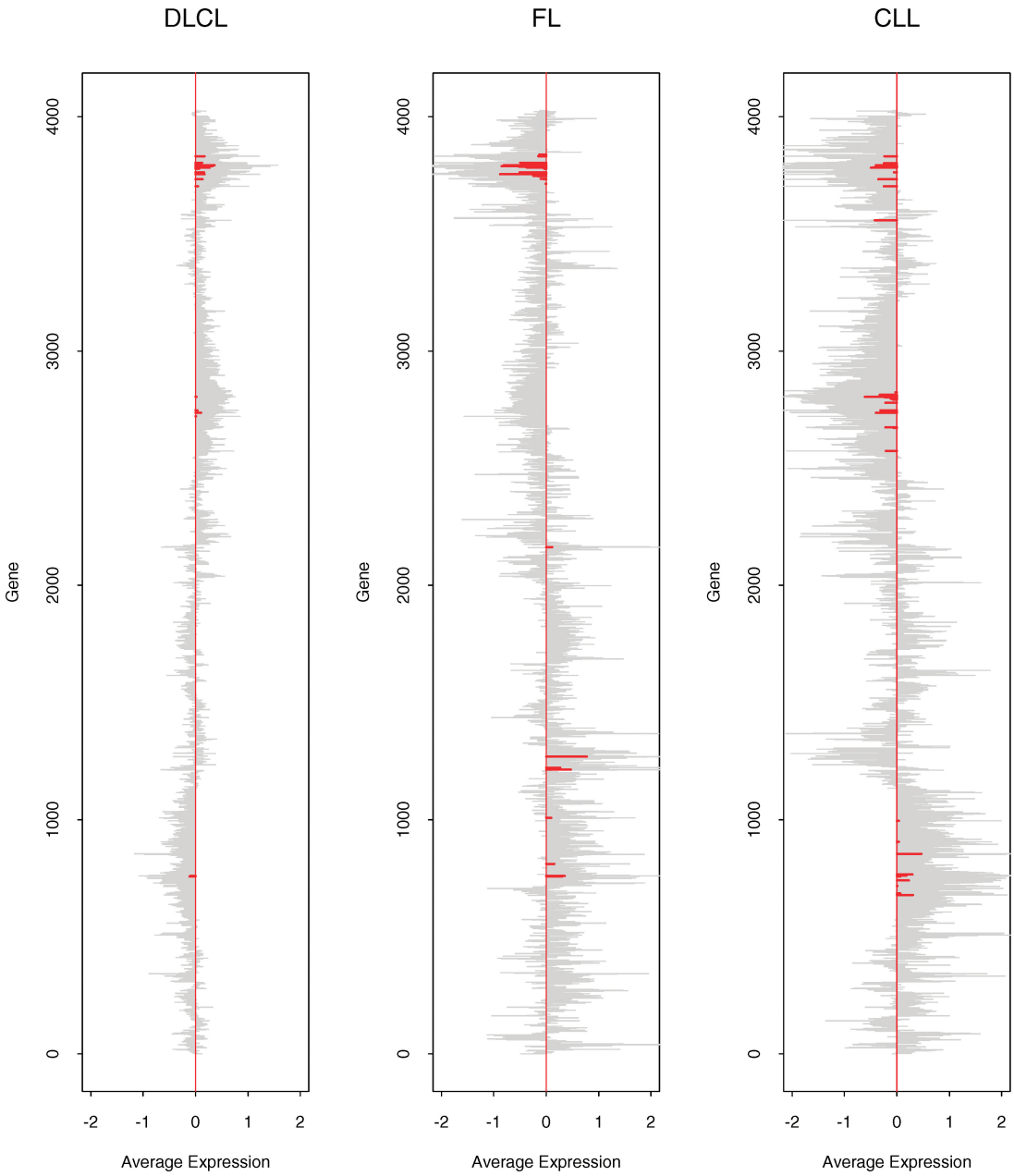

FIG. 1. *Centroids (grey) and shrunken centroids (red) for the lymphoma/leukemia data set. Each centroid has the overall centroid subtracted; hence, what we see are contrasts. The horizontal units are log ratios of expression. Going from left to right, the number of training samples is* 27, 5 *and* 7. *The order of the genes is determined by hierarchical clustering.*

We propose the "nearest shrunken centroid" method, which uses denoised versions of the centroids as prototypes for each class. The optimally shrunken centroids, derived using a method described below, are shown as red bars in Figure 1. Classification is then made to the nearest (shrunken) centroid. The resulting procedure has zero test errors. In addition, only 81 genes have a nonzero red bar for one or more classes in Figure 1 and, hence, are the only ones that contribute toward the classification. The amount of shrinkage is determined by cross-validation.

In the preceding example, the (unshrunken) nearest centroid method had the same error rate as the nearest shrunken centroid procedure. This is not always the case. Table 1 shows results taken from Tibshirani, Hastie, Narasimhan and Chu (2002) on classification of small round blue cell tumors. The data are taken from Khan et al. (2001). There are 25 test samples

TABLE 1
Results on classification of small round blue cell tumors

Method	Test error rate	Number of genes used
Nearest centroid	4/25	2,308
Nearest shrunken centroids	0/25	43
Neural network	0/25	96
Regularized discriminant analysis	0/25	2,308

and 2,308 genes. The neural network and regularized discriminant analysis methods used in the table are described in Section 7.

We gave a brief description of the nearest shrunken centroid method in Tibshirani, Hastie, Narasimhan and Chu (2002), focussing on the biological findings from two different applications. Here we give a broader and more thorough statistical treatment.

In Section 2 we describe the basic method. We detail our procedure for adaptive choice of thresholds in Section 3. Additional issues and comparisons are discussed in Sections 4–8, including application of the method to capturing heterogeneity with an "abnormal" class compared to a control class, in Section 6. Finally we conclude with a brief discussion in Section 9.

2. NEAREST SHRUNKEN CENTROIDS

2.1 Details of the Proposal

Let x_{ij} be the expression for genes $i = 1, 2, \ldots, p$ and samples $j = 1, 2, \ldots, n$. Each sample belongs to one of K classes $1, 2, \ldots, K$. Let C_k be indices of the n_k samples in class k. The ith component of the centroid for class k is $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$, the mean expression in class k for gene i; the ith component of the overall centroid is $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$.

Our idea is to shrink the class centroids toward the overall centroid. However, we first normalize by the within-class standard deviation for each gene. Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot s_i}, \tag{1}$$

where s_i is the pooled within-class standard deviation for gene i,

$$s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2, \tag{2}$$

and $m_k = \sqrt{1/n_k - 1/n}$ makes the denominator in Equation (1) equal to the estimated standard error of the numerator. Thus d_{ik} is a t-statistic for gene i, comparing class k to the average class. (In fact, we also add a regularization parameter s_0 to the values s_i; see Section 9.) We can write

$$\bar{x}_{ik} = \bar{x}_i + m_k s_i d_{ik}. \tag{3}$$

Our proposal shrinks each d_{ik} toward zero, giving d'_{ik} and new shrunken centroids or prototypes

$$\bar{x}'_{ik} = \bar{x}_i + m_k s_i d'_{ik}. \tag{4}$$

The shrinkage we use is called *soft thresholding*: The absolute value of each d_{ik} is reduced by an amount Δ and is set to zero if the result is less than zero. Algebraically, this is expressed as

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+, \tag{5}$$

where the subscript plus means *positive part* ($t_+ = t$ if $t > 0$ and zero otherwise). This is shown in Figure 2. Since many of the $\bar{x}_{ik}$ will be noisy and close to the overall mean $\bar{x}_i$, soft thresholding usually produces "better" (more reliable) estimates of the true means (Donoho and Johnstone, 1994). The proposed method has the attractive property that many of the components (genes) are eliminated as far as class prediction is concerned if the shrinkage parameter Δ is large enough. Specifically, if Δ causes d_{ik} to shrink to zero for all classes k, then the centroid for gene i is $\bar{x}_i$, the same for all classes. Thus gene i does not contribute

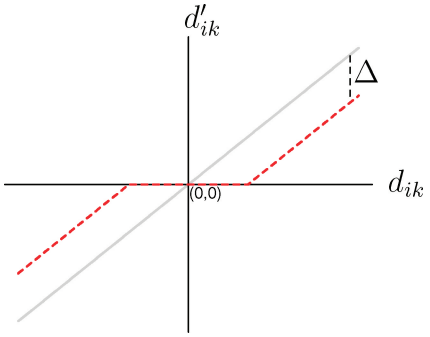

FIG. 2. *Soft threshold function.*

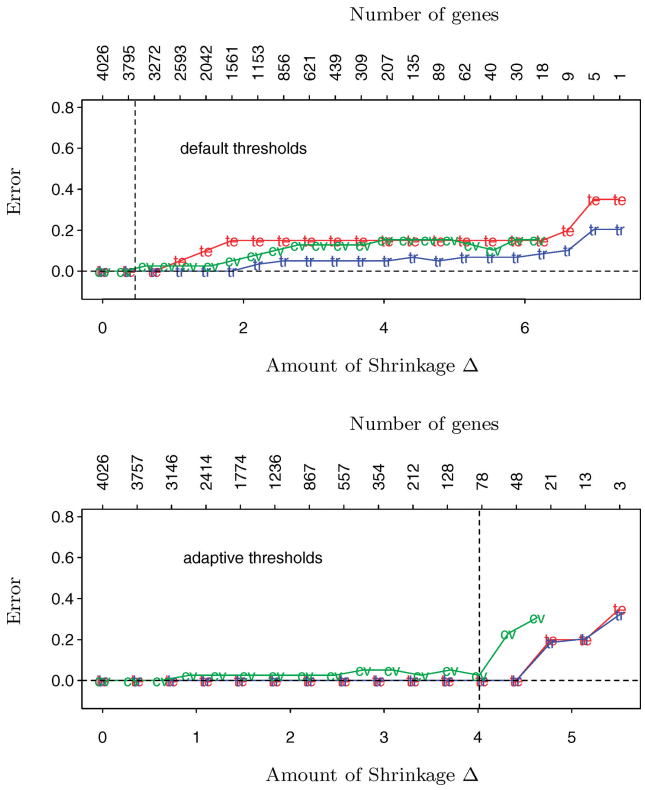

FIG. 3. *Lymphoma/leukemia data: training error (tr, blue), cross-validation error (cv, green) and test error (te, red) as the threshold parameter* Δ *is varied. In the top panel, the default soft threshold scaling is used: a solution with* $\Delta = 0.463$ *and* 3,666 *genes is chosen. In the bottom panel, adaptive threshold scaling was used; the value* $\Delta = 4.01$ *is chosen, resulting in a subset of just* 81 *genes, with the same test error rate as in the top panel.*

to the nearest centroid computation. We chose Δ by cross-validation, as illustrated below.

Note that the standardization by s_i in (1) has the effect of giving higher weight to genes that have stable expression within samples of the same class. This same standardization is inherent in other common statistical methods, such as linear discriminant analysis (see Section 7).

The top panel of Figure 3 shows the training, 10-fold cross-validation and test errors as the shrinkage parameter Δ is varied. The top of the plot indicates the number of genes retained (for the training data) at that particular threshold. The left end of the figure represents no shrinkage, while the right end represents complete shrinkage. The test error is minimized near $\Delta = 0.463$; when the curve is flat near the minimum, we typically chose the largest value of Δ (smallest number of genes) that achieves the minimal error. The upper axis shows the number of *active* genes with at least one nonzero component d'_{ik}, as Δ is varied. At $\Delta = 0.463$ there are about 3,666 active genes. The numbers of genes with nonzero d'_{ik} in each class are (3200, 2497, 3133).

Note that the selection of genes for a given value of Δ is carried out separately for each of the 10 cross-validation trials. This is important to avoid selection bias and an unrealistically optimistic cross-validation error rate. As pointed out by Ambroise and McLachlan (2002), a number of authors have made the mistake of selecting genes based on all of the training data (expression values and classes) and then subjecting only the selected genes to cross-validation. This can produce a wildly optimistic estimate for misclassification error: it is easy to simulate two-class examples in which the class labels are independent of the expression values (test error = 50%), but cross-validation after selection reports an error of zero.

Formula (1) takes into account the size of each class and effectively applies a larger threshold to a smaller (higher variance) class. Even after this adjustment, some classes may be farther away than others from the overall centroid and, hence, may be easier to distinguish. In this case, many of the nonzero genes for that class may not be needed for accurate classification. Thus we might try to vary the class thresholds to minimize the total number of nonzero genes needed to achieve a given error rate. The details of how we do this are discussed in Section 3. In this case the procedure increased the thresholds for the first and third classes, and was very successful: as shown in the bottom panel of Figure 3, it reduced the number of genes to just 81 without increasing the test error.

2.2 Finding the Predictors that Matter

Figure 4 shows the shrunken differences d_{ik} for the 81 genes that have at least one nonzero difference. Figure 5 shows the heat map of the chosen 81 genes. Within each of the horizontal partitions, we have ordered the genes by hierarchical clustering, and similarly for the samples within each vertical partition. Clear separation of the classes is evident. The top set of genes characterizes CLL with some genes overexpressed and others underexpressed. Similarly the middle set of genes characterizes FL. The genes in the bottom set of the figure are overexpressed in DLCL, and underexpressed in FL and CLL.

2.3 The Log-Likelihood

It is quite common to have a small number of samples in each class, especially when the number of

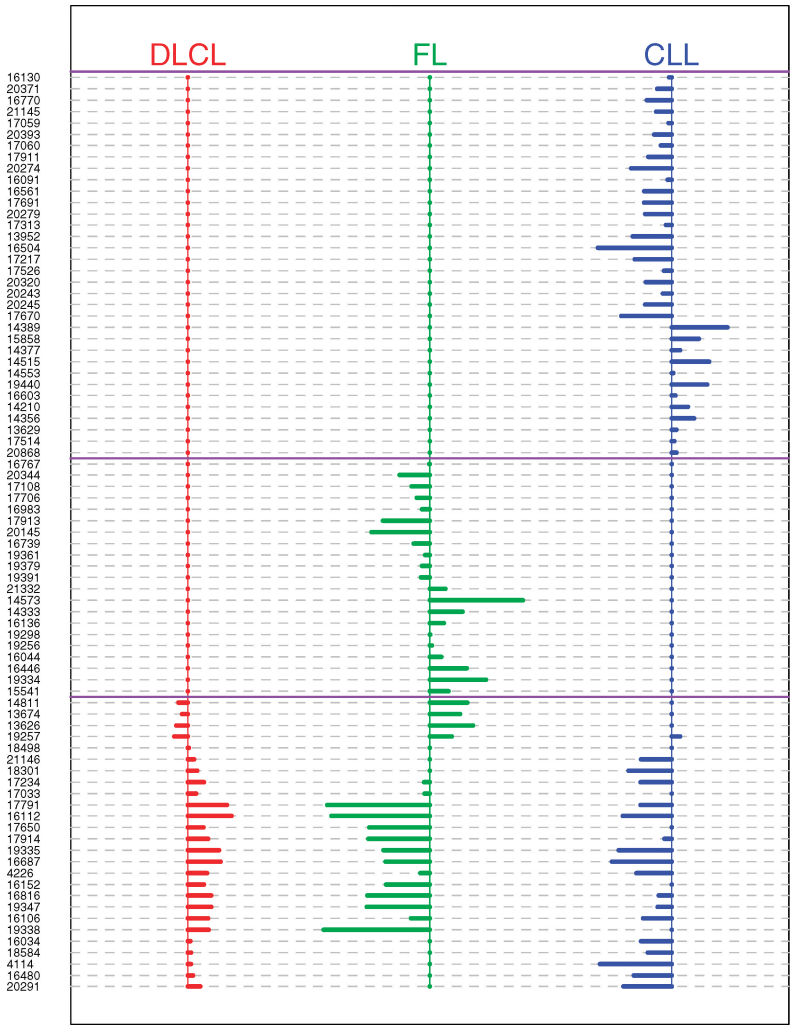

FIG. 4. *Shrunken differences d_{ik} for the* 81 *genes that have at least one nonzero difference.*

classes is large. This can result in a cross-validation curve that has discrete jumps and high variability.

To help with this problem, we can use the mean cross-validated log-likelihood rather than misclassification error. Since our model produces class probability estimates [see Equation (8) in Section 2.2], the log-likelihood of a test sample x^* with class label y^* is $\log \hat{p}_{y^*}(x^*)$. The mean log-likelihood curve is typically smoother than the misclassification error curve.

Figure 6 shows the test set log-likelihood and misclassification error curves for the lymphoma data. (This is for illustration only; we are not suggesting use of the test error to select Δ.) They give a similar picture, although the choice of the smallest model where the log-likelihood starts to dip yields more genes than that from the misclassification error curve. In the next section we make use of the log-likelihood in estimation of class probabilities.

2.4 Class Probabilities and Discriminant Functions

We classify test samples to the closest shrunken centroid, again standardizing by s_i. We also make a correction for the relative abundance of members of each class. Details are given next.

Suppose we have a test sample (vector) with expression levels $x^* = (x_1^*, x_2^*, \ldots, x_p^*)$. We define the *discriminant score* for class k as

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log \pi_k. \tag{6}$$

The first part of (6) is simply the standardized squared distance of x^* to the kth shrunken centroid. The second part is a correction based on the class *prior probability* π_k, where $\sum_{k=1}^{K} \pi_k = 1$. This prior gives the overall proportion of class k in the population. The

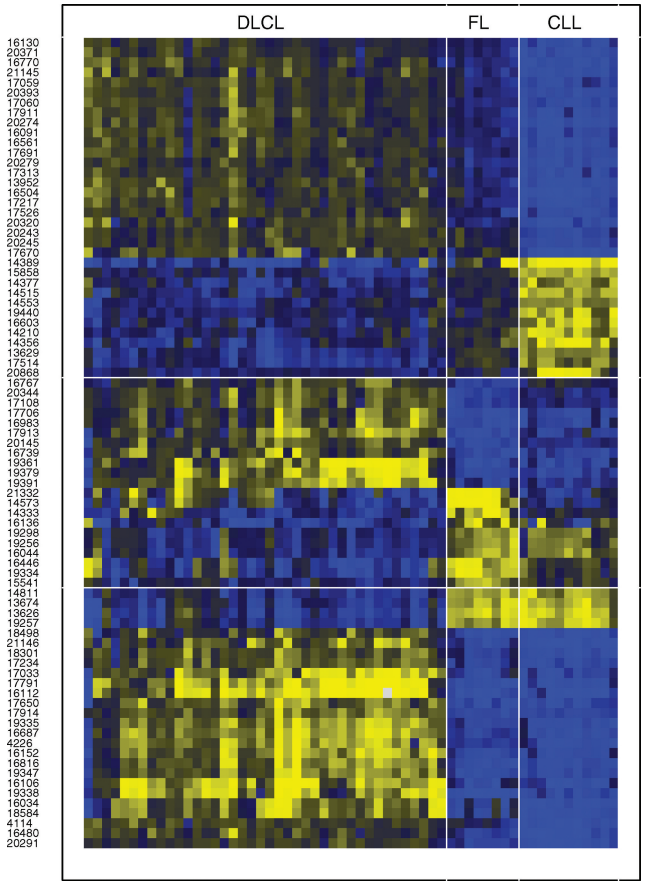

FIG. 5. *Heat map of the chosen* 81 *genes. Within each of the horizontal partitions, we have ordered the genes by hierarchical clustering, and similarly for the samples within each vertical partition. The data for all* 59 *samples are shown.*

classification rule is then

$$C(x^*) = \ell \quad \text{if } \delta_\ell(x^*) = \min_k \delta_k(x^*). \tag{7}$$

If the smallest distances are close and hence ambiguous, the prior correction gives a preference for larger classes, since they potentially account for more errors. We usually estimate the π_k by the *sample priors* $\hat{\pi}_k = n_k/n$. If the sample prior is not representative of the population, then more realistic priors or even uniform priors $\pi_k = 1/K$ can instead be used. We can use the discriminant scores to construct estimates of the class probabilities by analogy to Gaussian linear discriminant analysis:

$$\hat{p}_k(x^*) = \frac{e^{-(1/2)\delta_k(x^*)}}{\sum_{\ell=1}^{K} e^{-(1/2)\delta_\ell(x^*)}}. \tag{8}$$

The left panel of Figure 7 displays these probabilities for the lymphoma data. For illustration, we used the largest value of Δ $(= 4.41)$ that minimizes the test error in the bottom panel of Figure 3, rather than the cross-validation-minimizing value of 4.01 used earlier. The value $\Delta = 4.41$ yields 48 genes. We derived the probabilities using the centroids that were defined by applying this value of Δ to the test set.

In Figure 6, the value $\Delta = 4.04$ gives exactly the same test error (in fact, the same class predictions) as $\Delta = 4.41$, but gives a higher log-likelihood value. The estimated probabilities resulting from $\Delta = 4.04$ are shown in the right panel of Figure 7. These probabilities are more extreme than those in the left panel. The rightmost probabilities are preferred, since they produce a higher log-likelihood score.

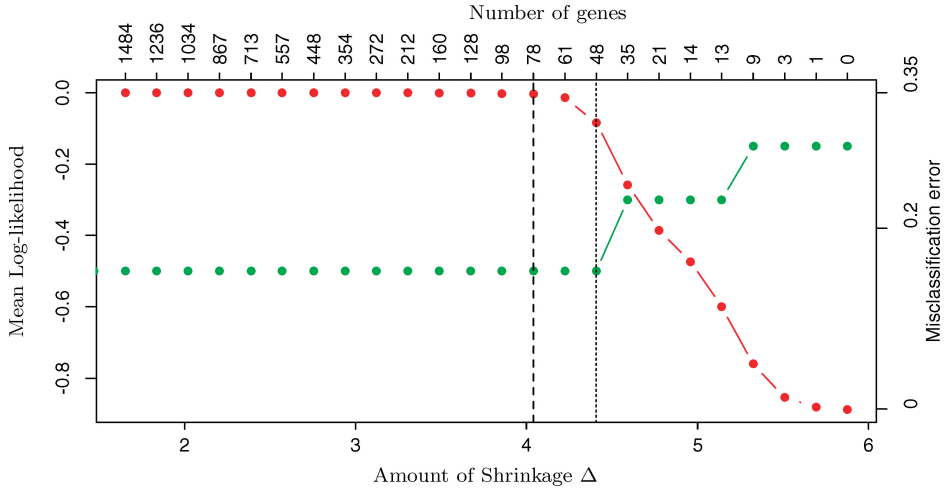

FIG. 6. *Test set mean log-likelihood curve (red) and test set misclassification error curve (green). The latter has been translated so that it fits in the same plotting region. The broken line shows where the log-likelihood curve starts to dip, while the dotted line shows where the misclassification error starts to rise.*

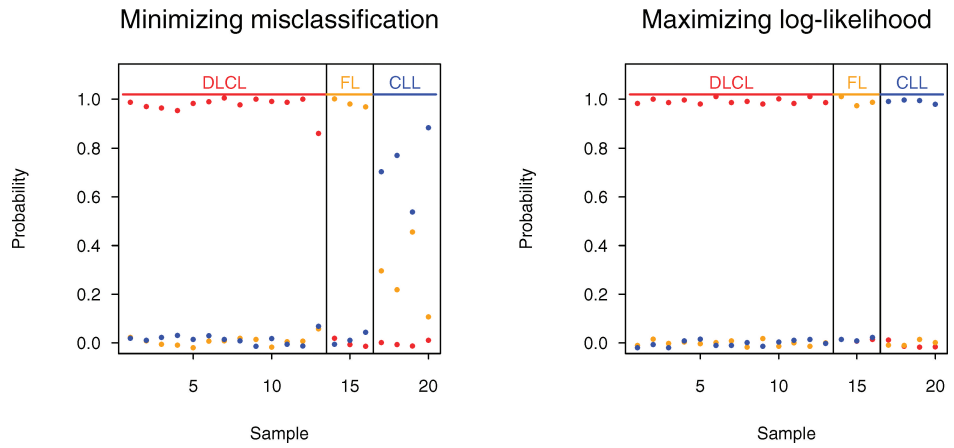

FIG. 7. *Estimated test set probabilities using the* 48 *gene model from minimizing misclassification error* (*left*) *and the* 78 *gene model from maximizing the log-likelihood* (*right*). *Probabilities are partitioned by the true class. There are no classification errors in the test set.*

3. ADAPTIVE CHOICE OF THRESHOLDS

In this section we describe the procedure for adaptive threshold choice in the nearest shrunken centroid method. We define a scaling vector $(\theta_1, \theta_2, \ldots, \theta_K)$ and initially set $\theta_k = 1$ for all k. These scalings are included in the denominator of expression (1), that is,

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \theta_k \cdot s_i}. \tag{9}$$

We scale the values so that $\min_j(\theta_j) = 1$: values greater than 1 mean that a larger threshold is effectively used for class k.

We applied the following procedure:

1. Find the class k with the largest number of training errors averaged over the grid of Δ values used.
2. Decrease θ_k by 10% and then rescale all θ_j so that $\min_j(\theta_j) = 1$.
3. Repeat the above steps for a number of iterations (here 10) and find the solution that gives the lowest average error, among the values of $(\theta_1, \theta_2, \ldots, \theta_K)$ visited.

Note that this procedure is based entirely on the training set and does not use information from cross-validation or a test set. It is admittedly heuristic, but does produce useful results in practice.

For the lymphoma data, we obtained the solution $(\theta_1, \theta_2, \theta_3) = (1.88, 1.00, 1.52)$, which is the value we used to produce Figures 1 and 4. Most of the errors in the original solution occurred in class FL; the new thresholds are larger for classes DLCL and CLL, and hence many fewer genes are used to discriminate these classes. Remarkably, the total number of genes used has decreased from 3,666 to 81 without raising the test error.

To test this procedure further, we simulated some data consisting of 10 samples in each of four classes and 1,000 genes. We ran two different simulations, with the results shown in the top and bottom panels of Figure 8. For a concise description, let $r(a, n)$ represent the number a repeated n times. All ex-

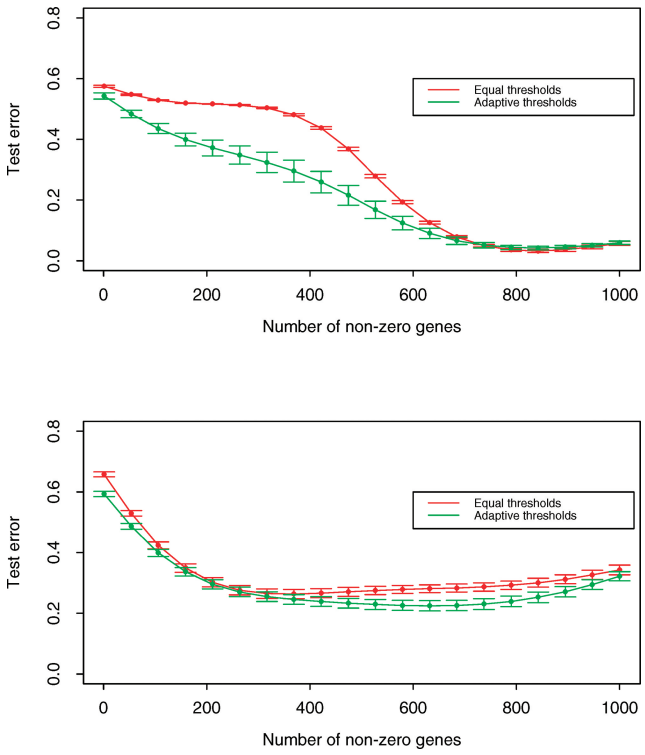

FIG. 8. *Simulated data: mean* $\pm$ 1 *standard deviation of the test error over five simulations, for default* (*equal*) *thresholds* (*red*) *and adaptive thresholds* (*green*). *In the setup for the top panel, the class centroids are unevenly spaced; in the bottom panel, the within-class variances are unequal.*

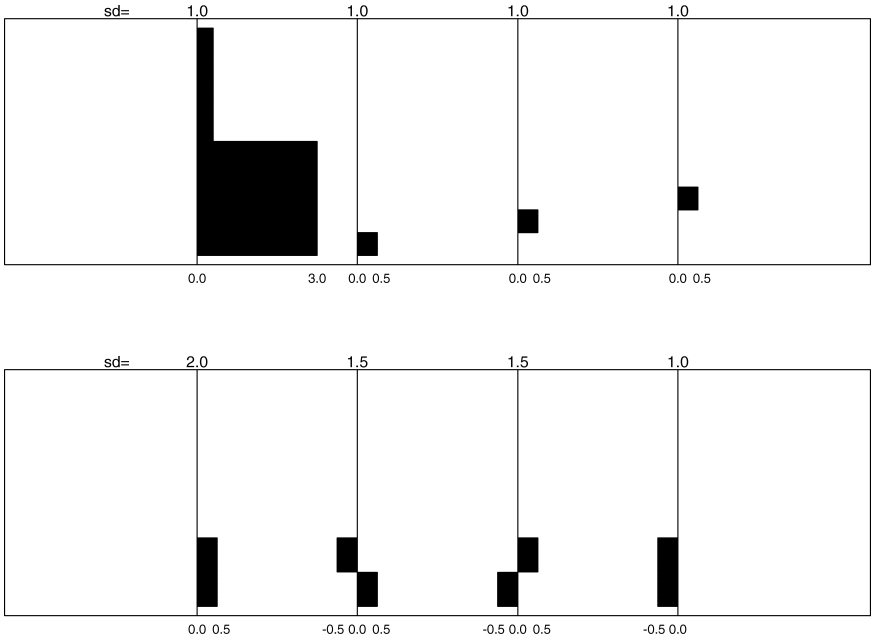

FIG. 9. *Centroids for each of four classes for the two simulation scenarios. The standard deviations for each class are indicated at the top of the plot.*

pression values were independent Gaussian with variance 1. In the first simulation, the class centroids were $[r(3, 500), r(0.4, 500)]$, $[r(0.5, 100), r(0, 900)]$, $[r(0, 100), r(0.5, 100), r(0, 800)]$ and $[r(0, 100), r(0, 100), r(0.5, 100), r(0, 700)]$. The centroids are shown in the top panel of Figure 9.

Thus the first class is far from the others, in the space spanned by the first 500 genes. The top panel of Figure 8 shows the mean $\pm$ 1 standard deviation of the test error over five simulations. The methods used were default (equal) thresholds (red) and adaptive thresholds (green). The average values of the adaptive threshold were 2.0, 1.0, 1.0 and 1.0. The adaptive threshold method generally has lower test error.

In the second simulation, the means in the four classes were $[r(0.5, 300), r(0, 700)]$, $[r(0.5, 150), r(-0.5, 150), r(0, 700)]$, $[r(-0.5, 150), r(0.5, 150), r(0, 700)]$ and $[r(-0.5, 150), r(-0.5, 150), r(0, 700)]$. The centroids are shown in the bottom panel of Figure 9. The standard deviations in each class were 2, 1.5, 1.5 and 1.0. Thus each class centroid is equidistant from the overall centroid (the origin), but the within-class standard deviations are different. The bottom of Figure 8 shows the results: again the adaptive threshold does better in terms of test error; the average values of the adaptive threshold were 1.4, 1.1, 1.2 and 1.0. With equal thresholds, the majority of nonzero genes were in class 1: under the adaptive thresholds, the distribution was more balanced.

4. SOFT VERSUS HARD THRESHOLDING

An alternative to the soft thresholding (5) would be to keep all differences greater in absolute value than Δ and discard the others; that is,

$$d'_{ik} = d_{ik} \cdot I(|d_{ik}| > \Delta). \tag{10}$$

This is sometimes known as *hard thresholding*. It differs from soft thresholding in that differences greater than Δ are unchanged, rather than shrunken toward zero by the amount Δ. One drawback of hard thresholding is its "jumpy" nature: as the threshold Δ is increased, a gene with a full contribution d_{ik} suddenly is set to zero.

To investigate the relative behavior of hard versus soft thresholding, we generated standard normal expression data for 1,000 genes and 40 samples, with 20 samples in each of two classes. For the first 100 genes, we added a random effect $\mu_i \sim N(0.0, 0.5^2)$ to each expression level in class 2 for each gene i. Hence 100 of the 1,000 genes are differentially expressed in the two classes by varying amounts. This experiment was repeated 10 times and the results were averaged. The left panel of Figure 10 shows the test error for hard and soft thresholding, as the threshold Δ is varied, while the right panel displays the mean squared error $\sum_i (\hat{\mu}_i - \mu_i)^2/p$, where $\hat{\mu}_i = \sum_{j=1}^{20} x'_{ij}/20 - \sum_{j=21}^{40} x'_{ij}/20$. In the left panel, we see that soft thresholding yields lower test error at its minimum; the right

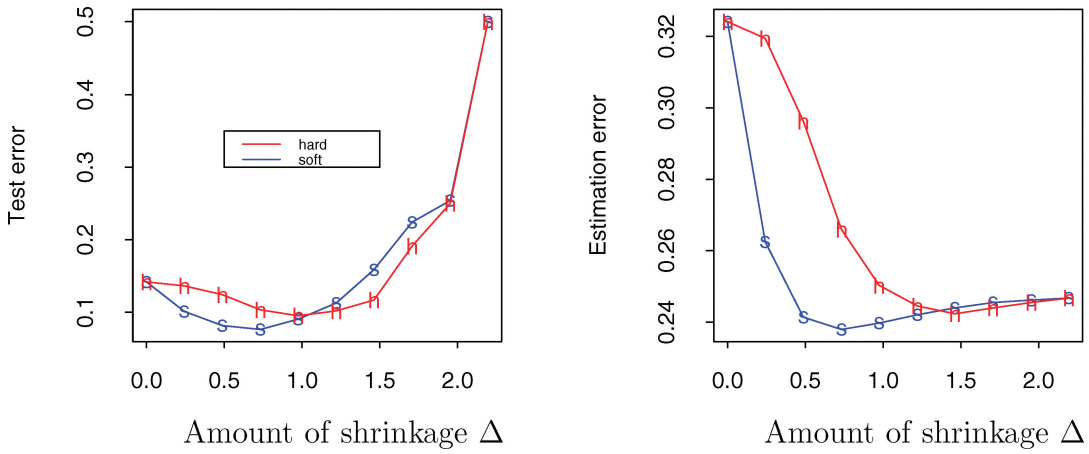

FIG. 10. *Simulated data in two classes. Left: Test misclassification error as the threshold* Δ *is varied, using hard thresholding* (h) *and soft thresholding* (s). *Right: The estimation error* $\sum(\hat{\mu}_i - \mu_i)^2/p$, *where* μ_i *and* $\hat{\mu}_i$ *are the true and estimated difference in expression between class* 1 *and class* 2 *for gene* i. *Results are averages over* 10 *simulations: standard error of the average is about* 0.015 *in the left panel and* 0.01 *in the right panel.*

panel shows that soft thresholding does a much better job of estimating the gene expression differences.

5. NATIONAL CANCER INSTITUTE CANCER LINES AND SUBCLASS DISCOVERY

Here we describe how to use nearest centroid shrinkage to discover subclasses. We consider data from Ross et al. (2000) that consist of measurements on 6,830 genes on 61 cell lines. The samples have been categorized into eight different cancer classes: breast (BRE), CNS, colon (COL), leukemia (LEU), melanoma (MEL), non-small cell lung cancer (NSC), ovarian (OVA) and renal (REN). We randomly chose a training set of size 40 and a test set of size 21, so that the classes were well represented in both sets. Default (equal) soft thresholding was used, with the prior probabilities set to the sample class proportions. The results are shown in Figure 11. The best cross-validated error rate occurs at about 5,000 genes, giving a test error of 5/21. Adaptive thresholding failed to improve this result.

We also tried both support vector machines (Ramaswamy et al., 2001) and regularized discriminant analysis (Section 7). Both gave five errors on the test set. However, neither method gave a simple picture of the data.

Next we show a generalization of the nearest shrunken centroid approach that facilitates the discovery of potentially important subclasses. It may be valuable biologically to look for distinct subclasses of diseases in microarray analyses. We can generalize the nearest shrunken centroid procedure to facilitate the discovery of subclasses. Consider the problem illustrated in Figure 12. The values indicate average gene expression. There are two subclasses in class 2, and each of these can be distinguished from class 1 based on a small set of genes. However, nearest shrunken centroids will fail here, because the overall centroids for each class are the same. Linear separating classifiers, such as support vector machines (SVM), and linear discriminant analysis will also do poorly here. Either could be made to work with a suitable nonlinear transformation of the features (or choice of kernel for the

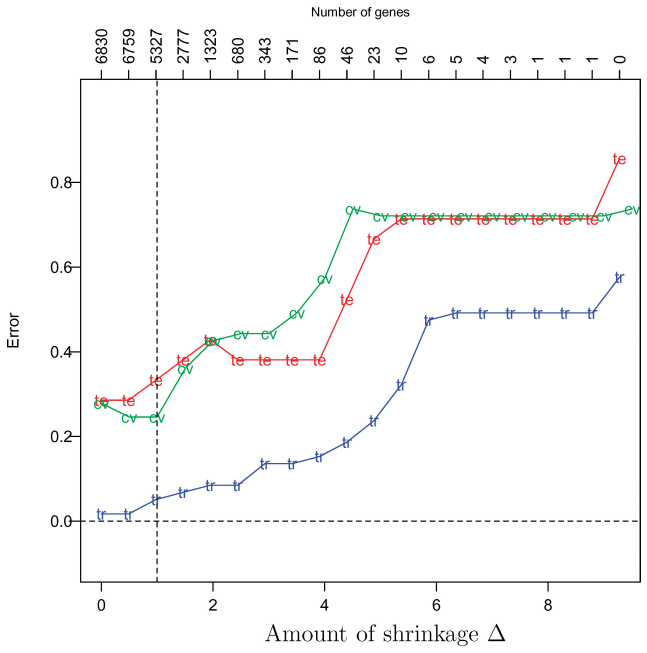

FIG. 11. *NCI cancer cell lines: training, cross-validation and test error curves.*

TABLE 2
NCI subclass results: test errors (out of 21 samples) for nearest shrunken centroid model with no subclasses (second column from left) and two subclasses per class (third column from left); the columns on the right show the resulting number of errors when a pair of subclasses for a given class is fused into one subclass

Number of genes	Zero subclasses	Two subclasses	Fusing subclasses for each class							
			BRE	CNS	COL	LEU	MEL	NSC	OVA	REN
6830	5	6	6	6	6	6	6	7	5	8
6827	5	6	5	5	7	6	6	7	6	6
6122	5	5	6	5	5	5	5	5	5	5
3571	7	6	8	7	6	6	6	6	7	6
1695	9	6	8	7	6	6	6	6	7	6
696	9	7	9	6	7	7	7	7	8	8
293	9	6	8	7	7	6	6	7	7	6
119	10	6	8	8	8	6	8	7	7	8
42	10	12	13	14	14	12	12	12	12	12
17	14	14	14	14	16	14	14	14	14	13

SVM); while these may give low prediction error, they may not reveal the biologically important subclasses that are present.

For any class, our idea is to apply r-means clusters to the samples in that class, resulting in r subclasses for that class. Doing this for each of the K classes results in a total of $K \cdot r$ subclasses. We apply nearest shrunken centroids to this $r \cdot K$ class problem. If the predicted class from this large problem is h, then our final predicted class is the class k that contains subclass h.

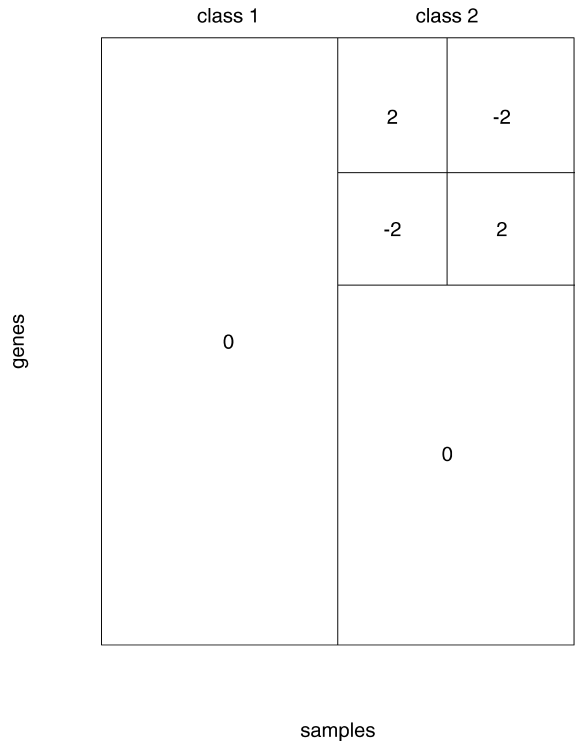

FIG. 12. *Two class problem with distinct subclasses. Numbers indicate the average gene expression.*

With typical sample sizes, the choice $r = 2$ will be most reasonable. Table 2 shows the results on the National Cancer Institute (NCI) data. Without subclasses, the test error rates start to rise when fewer than 2,000 or 3,000 genes are used. Using subclasses, we achieve about the same error rate with as few as 119 genes. The right part of the table shows that for 119 the subclasses are most important for BRE, CNS, COL, MEL and REN. The 119 gene solution is displayed in Figure 13 and shows some distinct subclasses among some of the main classes.

6. CAPTURING HETEROGENEITY

In discriminating an "abnormal" from a "normal" group, the average gene expression may not differ between the groups. However, the variability in expression may be greater in the abnormal group, due to heterogeneity in the abnormal population. This is illustrated in Figure 14. Nearest centroid classification will not work in this case, since the class centroids are not separated. The subclass method of the previous section might help: we propose an alternative approach here.

We define new features $x'_{ij} = |x_{ij} - \bar{m}_i|$, where $\bar{m}_i$ is the mean expression for gene i in the normal group. Then we apply nearest shrunken centroids to the new features x'_{ij}.

To illustrate this, we generated the expression of 1,000 genes in 40 samples—20 from a normal group

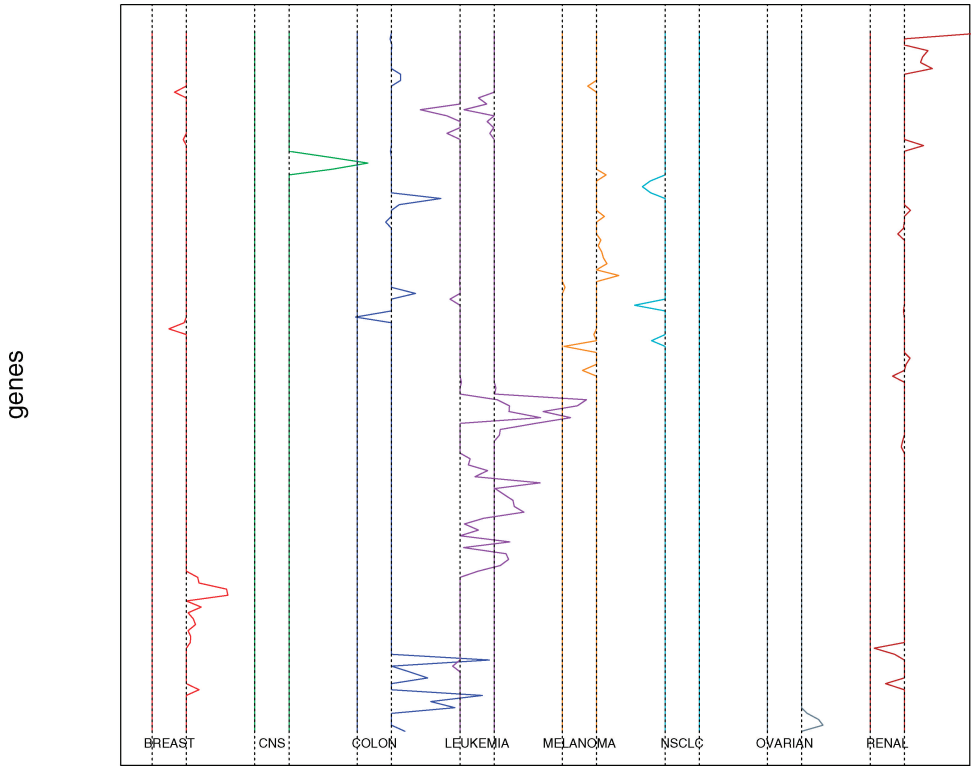

FIG. 13. *NCI subclass results. Shown are pairs of centroids for each class for the genes that survived the thresholding.*

and 20 from an abnormal group. All expression values were generated independently as standard Gaussian except for the first 200 genes in the abnormal group, which had mean zero, but standard deviation 2. An independent test set of size 200 was also generated.

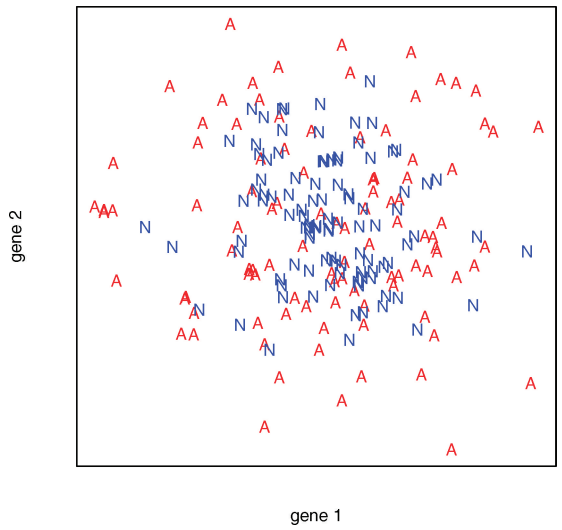

FIG. 14. *Illustration of heterogeneity in gene expression. Abnormal group A has the same average gene expression as the normal group N, but shows larger variability.*

Nearest centroid shrinkage on the transformed features x'_{ij} showed a test error rate of near zero, with 150 or more nonzero genes. Figure 15 compares the results of nearest shrunken centroids on the raw expression values x_{ij} and the transformed expression values x'_{ij}. Nearest centroid shrinkage on the raw values does poorly with an error rate greater than 40%, while use of the transformed values reduces the error rate to near zero.

By transforming to the distance from the normal centroid, the use of the features x'_{ij} might also provide discrimination in situations where the abnormal class is not heterogeneous, but is instead mean-shifted. The right panel of Figure 15 investigates this. The expression of the first 200 genes in the abnormal class has mean 0.5 and standard deviation 1 (versus 0 and 1 for the normal class). Now nearest shrunken centroids on the raw features is much more powerful, while use of the transformed features works poorly. We conclude that use of neither the raw nor transformed features dominates the other, and both should be tried on a given problem.

We have successfully used the heterogeneity model to predict toxicity from radiation sensitivity using transcriptional responses to DNA damage in lymphoid cells (Rieger et al., 2003).

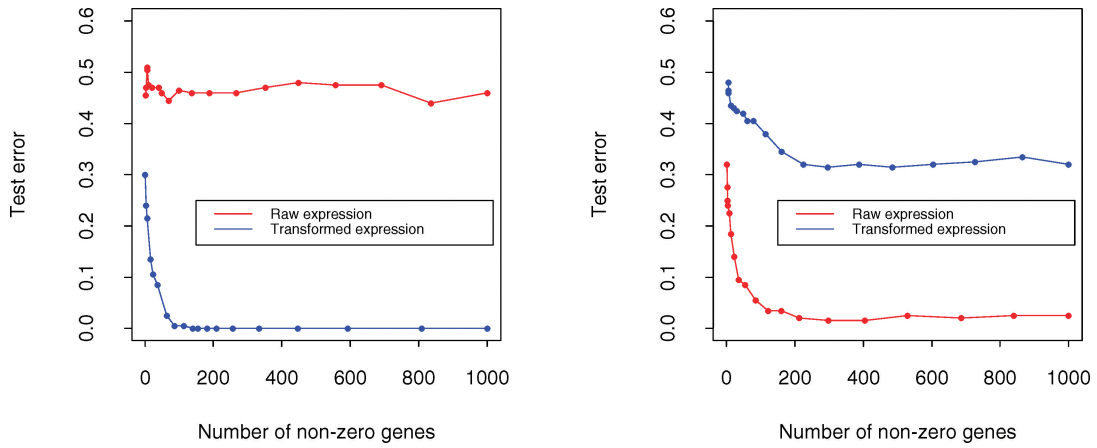

FIG. 15. *Left: Test error for data simulated from the heterogeneous two-class problem, using nearest shrunken centroids on raw expression values (red) and transformed expression values* $|x_{ij} - \bar{m}_i|$ *(blue). Right: Same as in left panel, but data are simulated from the mean-shifted homogeneous two-class problem.*

7. RELATIONSHIP TO OTHER APPROACHES

The discriminant scores (6) are similar to those used in linear discriminant analysis (LDA), which arise from using the *Mahalanobis* metric to compute the distance to centroids:

$$\delta_k^{\mathrm{LDA}}(x^*) = (x^* - \bar{x}_k)^T W^{-1}(x^* - \bar{x}_k) - 2\log\pi_k. \tag{11}$$

Here we use vector notation and W is the pooled within-class covariance matrix. With thousands of genes and tens of samples ($p \gg n$), W is huge and any sample estimate will be singular (and hence its inverse is undefined). Our scores can be seen to be a heavily restricted form of LDA, necessary to cope with the large number of variables (genes). The differences are the following:

- We assume a diagonal within-class covariance matrix for W; without this, LDA would be ill-conditioned and fail.
- We use shrunken centroids rather than centroids as a prototype for each class.
- As the shrinkage parameter Δ increases, an increasing number of genes will have *all* their $d'_{ik} = 0$, $k = 1, \ldots, K$, due to the soft thresholding in (5). Such genes contribute no discriminatory information in (6), and in fact cancel in Equation (8).

Both our scores (6) and the LDA scores (11) are *linear* in x_i^*. If we expand the square in (6), discard the terms involving x_i^{*2} (since they are independent of the class index k and hence do not contribute toward class discrimination) and multiply by $-1/2$, we get

$$\tilde{\delta}_k(x^*) = \sum_{i=1}^{p} \frac{x_i^* \bar{x}'_{ik}}{s_i^2} - \frac{1}{2}\sum_{i=1}^{p} \frac{\bar{x}'^2_{ik}}{s_i^2} + \log\pi_k, \tag{12}$$

which is linear in x_i^*. Because of the sign change, our rule classifies to the largest $\tilde{\delta}_k(x^*)$. Likewise the LDA discriminant scores have the equivalent linear form

$$\tilde{\delta}_k^{\mathrm{LDA}}(x^*) = x^{*T} W^{-1}\bar{x}_k - \tfrac{1}{2}\bar{x}'^T_k W^{-1}\bar{x}'_k + \log\pi_k. \tag{13}$$

Regularized discriminant analysis (RDA; Friedman, 1989) leaves the centroids alone and modifies the covariance matrix in a different way,

$$\delta_k^{\mathrm{RDA}}(x^*) = (x^* - \bar{x}_k)^T (W + \lambda I)^{-1}(x^* - \bar{x}_k), \tag{14}$$

where λ is a parameter (like our Δ). The fattened $W + \lambda I$ is nonsingular, and as λ gets large, this procedure approaches the nearest centroid procedure (with no variance scaling or centroid shrinking). A slightly modified version uses $W + \lambda D$, where $D = \mathrm{diag}(s_1^2, s_2^2, \ldots, s_p^2)$. As λ gets large, this approaches the variance weighted nearest centroid procedure. In practice, we normalize this regularized covariance by dividing by $1 + \lambda$, leading to the convex combination $(1 - \alpha)W + \alpha D$, where $\alpha = \lambda/(1 + \lambda)$. Although the relative distances do not change, this is important when making the adjustment for the class priors.

Although RDA shows some promise, it is more complicated than our nearest shrunken centroid procedure. Furthermore, in the process of its regularization, it does not select a subset of genes as the shrunken centroid procedure does. We are considering other hybrid approaches of RDA and nearest centroids in ongoing research projects.

8. NEAREST CENTROID CLASSIFIER VERSUS LDA

As discussed in the previous section, the nearest centroid classifier is equivalent to Fisher's linear discrimi-

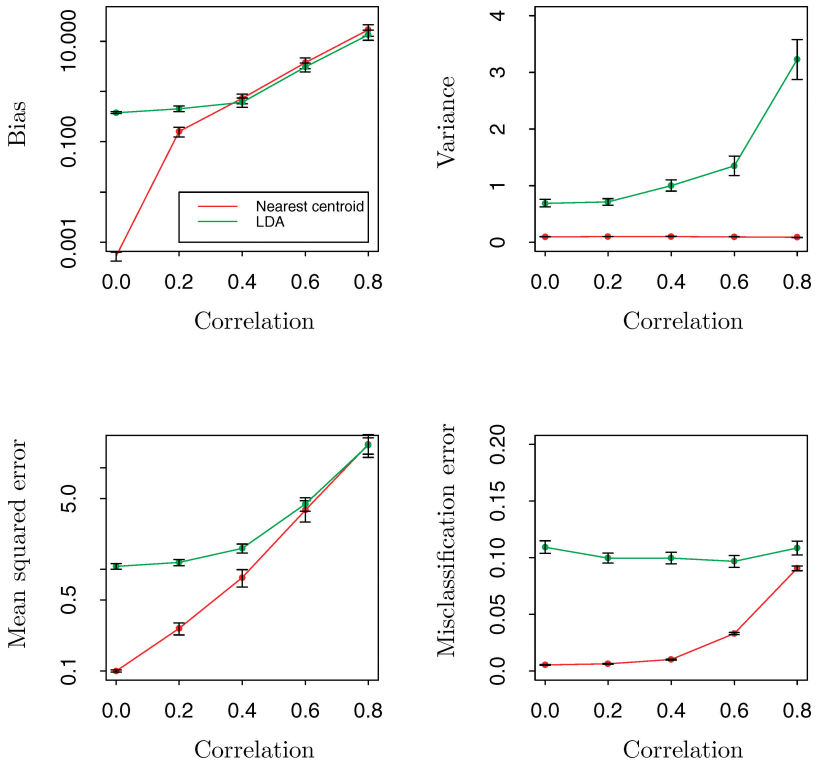

FIG. 16. *Simulation results: bias and variance (top panels) and mean-squared error and misclassification error (bottom panels) for linear discriminant analysis and the nearest centroid classifier. Details of the simulation are given in the text. The nearest centroid classifier outperforms LDA because of its smaller variance.*

nant analysis if we restrict the within-class covariance matrix to be diagonal. When is this restriction a good one?

Consider a two class microarray problem with p genes and n samples. For simplicity we consider the standard (unshrunken) nearest centroid classifier and standard (full within covariance) LDA. The recent thesis of Levina (2002) did some theoretical comparisons of these methods. She assumed $p \to \infty$, $n \to \infty$ and $p/n \to \gamma \in (0, 1)$, and analyzed the worst case error of each method. The relative performance of the two methods depends on the correlation structure of the features (samples). Her results show that if p is a large fraction of n, for a large class of correlation structures, nearest centroid classification outperforms full LDA.

Now in our problem, usually we have $p \gg n$: in that case, LDA is not even defined without some regularization. Hence to proceed we assume that p is a little less than n and hope that what we learn will extend to the case $p > n$. Let x_j be a p-vector of gene expression values in class j. Suppose $x_1 \sim N(0, \Sigma)$ and $x_2 \sim N(\mu, \Sigma)$, where Σ is a full (nondiagonal) matrix. Then LDA uses the maximum likelihood unbiased estimate of $\Sigma^{-1}\mu$, while nearest centroid uses a biased estimate. However, the LDA method estimates $\Sigma^{-1}\mu$ in a multivariate manner, and hence will tend to have higher variance. What is the resulting bias–variance tradeoff and how does it translate into misclassification error?

We did an experiment with $p = 30$ and $n = 40$, with 20 samples in each of two classes. We set the ijth element of Σ to $\rho^{|i-j|}$, where ρ was varied from 0 to 0.8. Each of the components of the mean vector μ was set to ± 1 at random: such a mixed vector is needed to give full LDA a potential advantage over LDA with a diagonal covariance. For each simulation, an independent test set of size 500 was also generated. The results of 100 simulations from this model are shown in Figure 16. Bias, variance and mean-squared error refer to estimation of $\Sigma^{-1}\mu$. For small correlations, the underlying (diagonal covariance) model for nearest centroids is approximately correct and the method wins; LDA shows a small improvement in bias for larger correlations, but this is more than offset by the increased variance. Overall the nearest centroid method has lower mean-squared error and test misclassification error in all cases.

Now for real microarray problems, $p \gg n$, and both LDA and nearest centroid methods can be improved by appropriate regularization or shrinkage. We have

not included regularization in the above comparison, but the above results suggest that the bias–variance tradeoff will cause the nearest centroid method to outperform full LDA.

9. DISCUSSION

The nearest shrunken centroid classifier is potentially useful in any high-dimensional classification problem. In addition to its application to gene expression arrays, it could also be applied to other kinds of emerging genomic data, including mass spectroscopy for protein measurements, tissue arrays and single nucleotide polymorphism arrays.

Our proposal can also be applied in conjunction with unsupervised methods. For example, it is now standard to use hierarchical clustering methods on expression arrays to discover clusters in the samples (Eisen, Spellman, Brown and Botstein, 1998). The methods described here can identify subsets of the genes that succinctly characterize each cluster.

Finally, we touch on computational issues. The computations involved in the nearest shrunken centroid method are straightforward. One important detail: in the denominator of the statistics d_{ik} in Equation (1) we add the same positive constant s_0 to each of the s_i values. This guards against the possibility of large d_{ik} values arising by chance from genes at very low expression levels. We set s_0 equal to the median value of the s_i over the set of genes. A similar strategy was used in the significance analysis of microarrays (SAM) methodology of Tusher, Tibshirani and Chu (2001).

We have developed a package in the Excel and R language called prediction analysis for microarrays. It implements all of the nearest shrunken centroids methodology discussed in this article and is available at the website http://www-stat.stanford.edu/~tibs/PAM.

REFERENCES

ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., MA, C., LOSSOS, I. S., ROSENWALD, A., BOLDRICK, J. C., SABET, H., TRAN, T., YU, X., POWELL, J. I., YANG, L., MARTI, G. E., MOORE, T., HUDSON, JR., J., LU, L., LEWIS, D. B., TIBSHIRANI, R., SHERLOCK, G., CHAN, W. C., GREINER, T. C., WEISENBURGER, D. D., ARMITAGE, J. O., WARNKE, R., LEVY, R., WILSON, W., GREVER, M. R., BYRD, J. C., BOTSTEIN, D., BROWN, P. O. and STAUDT, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** 503–511.

AMBROISE, C. and MCLACHLAN, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U.S.A.* **99** 6562–6566.

DONOHO, D. and JOHNSTONE, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95** 14 863–14 868.

FRIEDMAN, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84** 165–175.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M., 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.

HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. and BROWN, P. (2001). Supervised harvesting of expression trees. *Genome Biology* **2** (1) research/0003.

HEDENFALK, I., DUGGAN, D., CHEN, Y., RADMACHER, M., BITTNER, M., SIMON, R., MELTZER, P., GUSTERSON, B., ESTELLER, M., RAFFELD, M., YAKHINI, Z., BEN-DOR, A., DOUGHERTY, E., KONONEN, J., BUBENDORF, L., FEHRLE, W., PITTALUGA, S., GRUVBERGER, S., LOMAN, N., JOHANNSSON, O., OLSSON, H., WILFOND, B., SAUTER, G., KALLIONIEMI, O., BORG, A. and TRENT, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal Medicine* **344** 539–548.

KHAN, J., WEI, J., RINGNER, M., SAAL, L., LADANYI, M., WESTERMANN, F., BERTHOLD, F., SCHWAB, M., ANTONESCU, C., PETERSON, C. and MELTZER, P. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7** 673–679.

LEVINA, E. (2002). Statistical issues in texture analysis. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J., POGGIO, T., GERALD, W., LODA, M., LANDER, E. and GOLUB, T. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* **98** 15 149–15 154.

RIEGER, K., HONG, W., TUSHER, V., TANG, J., TIBSHIRANI, R. and CHU, G. (2003). Toxicity of radiation therapy associated with abnormal transcriptional responses to DNA damage. Submitted.

ROSS, D., SCHERF, U., EISEN, M., PEROU, C., REES, C., SPELLMAN, P., IYER, V., JEFFERY, 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.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B., and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99** 6567–6572.

TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98** 5116–5121.