

Class prediction by nearest shrunken centroids, with applications to DNA microarrays

Robert Tibshirani^{*}, Trevor Hastie[†],
Balasubramanian Narasimhan[‡],
and
Gilbert Chu[§]

June 12, 2002

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 and identifies the subsets of the genes that best characterize each class. The method is general, and can be used in other high-dimensional classification problems. The method is illustrated on data from two studies: lymphoma and cancer cell lines.

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

^{*}Depts. of Health, Research & Policy, and Statistics, Stanford Univ, tibs@stat.stanford.edu

[†]Depts. of Statistics, and Health, Research & Policy, Sequoia Hall, Stanford Univ., CA 94305. hastie@stat.stanford.edu

[‡]Depts. of Statistics, and Health, Research & Policy, Sequoia Hall, Stanford Univ., CA 94305. naras@stat.stanford.edu

[§]Depts. of Biochemistry and Medical Oncology, Stanford Univ., CA 94305. chu@cmgm.stanford.edu

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 et al. (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, and it is especially important to identify which genes contribute towards the classification. In this paper we propose a very 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 4026 genes from samples of 59 lymphoma patients. The samples are classified into diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL) and chronic lymphocytic lymphoma (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 begin with a nearest centroid classification. Figure 1 (light grey bars) shows the training-set centroids (average expression of each gene) for each of the 3 classes. The overall gene expression has been subtracted, so that these values are differences from the overall centroid.

To apply nearest centroid classification, we take the gene expression profile of the test sample (array), and compute its squared distance from each of the 3 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 4026 genes.

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 the red bars in Figure 1. Classification is then made to the nearest (shrunken) centroid. The resulting procedure has zero test errors. In addition, only 48 genes have a non-zero red bar for one or more classes in Figure 1, and hence are the only ones that contribute towards the classification. The amount of shrinkage is determined either by examination of the error on a test set, or 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

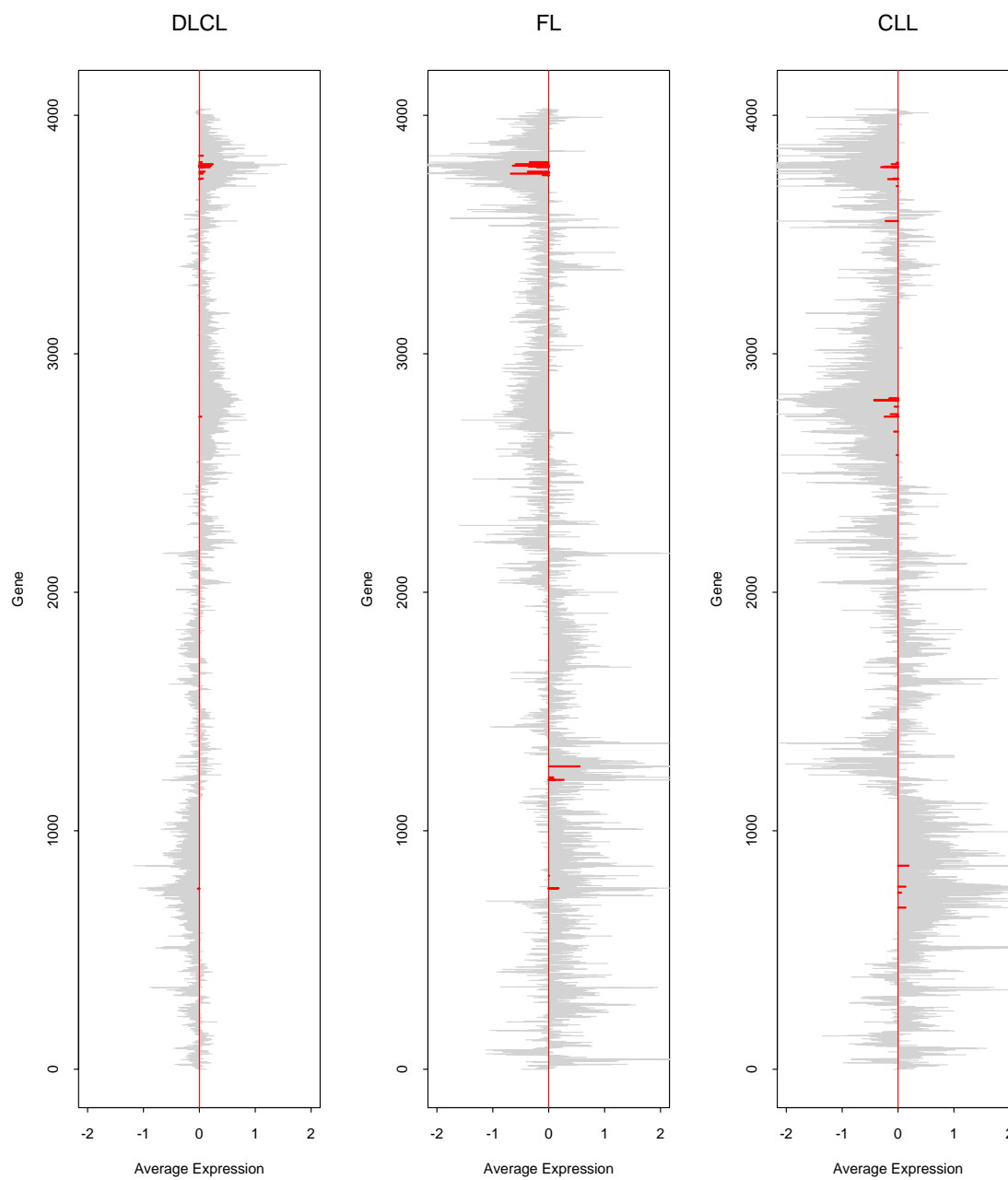


Figure 1: Centroids (grey) and shrunken centroids (red) for the lymphoma dataset. Each centroid has the overall centroid subtracted, and hence what we see are contrasts. The horizontal units are log ratios of expression. Going from left to right, the number of training samples is 27,5,7. The order of the genes is arbitrary.

Table 1: *Results on classification of small, round blue cell tumors*

Method	Test error rate	Number of genes used
Nearest centroid	4/25	2308
Nearest shrunken centroids	0/25	43
Neural network	0/25	96
Regularized discriminant analysis	0/25	2308

always the case. Table 1 shows results taken from Tibshirani et al. (2001) on classification of small, round blue cell tumors. The data are taken from Khan et al. (2001). There are 25 test samples and 2308 genes. The neural network and regularized discriminant analysis methods used in the table are described in Section 7.

In Tibshirani et al. (2001) we give a brief description of the method, 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. Section 3 discusses an application of the method to capturing heterogeneity in a treatment versus control comparison. We detail our procedure for adaptive choice of thresholds in Section 4. Additional issues and comparisons are discussed in Sections 5–8. 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 $i = 1, 2, \dots, p$ and samples $j = 1, 2, \dots, n$. We have classes $1, 2, \dots, K$, and let C_k be indices of the n_k samples in class k . The i th component of the centroid for class k is $\bar{x}_{ik} = \sum_{i \in C_k} x_{ij} / n_k$, the mean expression value in class k for gene i ; the i th component of the overall centroid is $\bar{x}_i = \sum_{j=1}^n x_{ij} / n$.

In words, we shrink the class centroids towards the overall centroids. However, we first normalize by the within class-standard deviation for each

gene. Let

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

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

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

and $m_k = \sqrt{1/n_k - 1/n}$ makes the denominator equal to the estimated standard error of the numerator in d_{ik} . Thus d_{ik} is a t-statistic for gene i , comparing class k to the average class. We can write

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

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

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

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

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

where $+$ 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 produces “better” (more reliable) estimates of the true means (Donoho & Johnstone 1994). The proposed method has the nice 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 for a gene i , d_{ik} is shrunken 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 to the nearest centroid computation. We choose Δ by test set validation or cross-validation, as illustrated below.

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

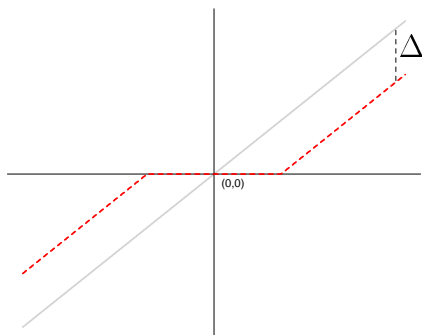


Figure 2: *Soft threshold function*

The top panel of Figure 3 shows the training and test errors as the shrinkage parameter Δ is varied. The “Size” axis at 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 $\Delta = 0.918$. The upper axis shows the number of *active* genes with at least one non-zero component d'_{ik} , as Δ is varied. At $\Delta = .918$ there are 2938 active genes. The number of genes with non-zero d'_{ik} in each class were (2471, 1312, 2340).

In this example, we have set aside a single test set for illustration of the methods. In practice, with a small number of samples one would instead carry out five or ten-fold cross-validation.

The 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 be easier to distinguish. In this case, many of the non-zero genes for that class may not be needed for accurate classification. Thus we might try vary the class thresholds to minimize total number of non-zero genes needed to achieve a given error rate. The details of how we do this are discussed in Section 4. 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 48 without increasing the test error.

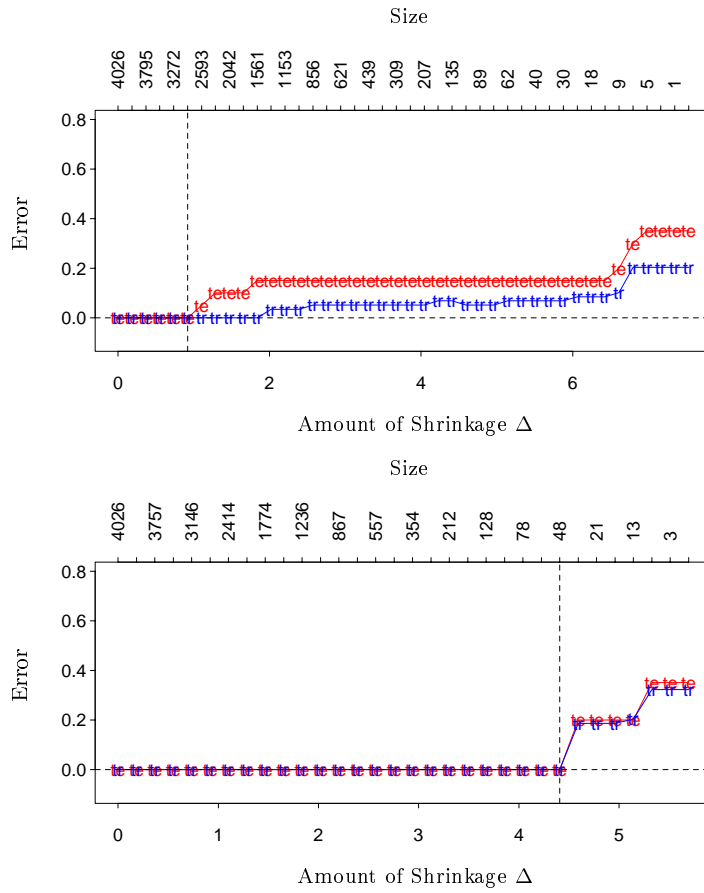


Figure 3: Training (*tr*, blue) and cross-validation error (*te*, red) as the threshold parameter Δ is varied. In the top panel, the default thresholding scaling is used: a solution with $\Delta = 0.918$ and 2938 genes is chosen. In the bottom, adaptive threshold scaling was used: the value $\Delta = 4.41$ is chosen, resulting in a subset of just 48 genes, with the same test error rate as in the top panel.

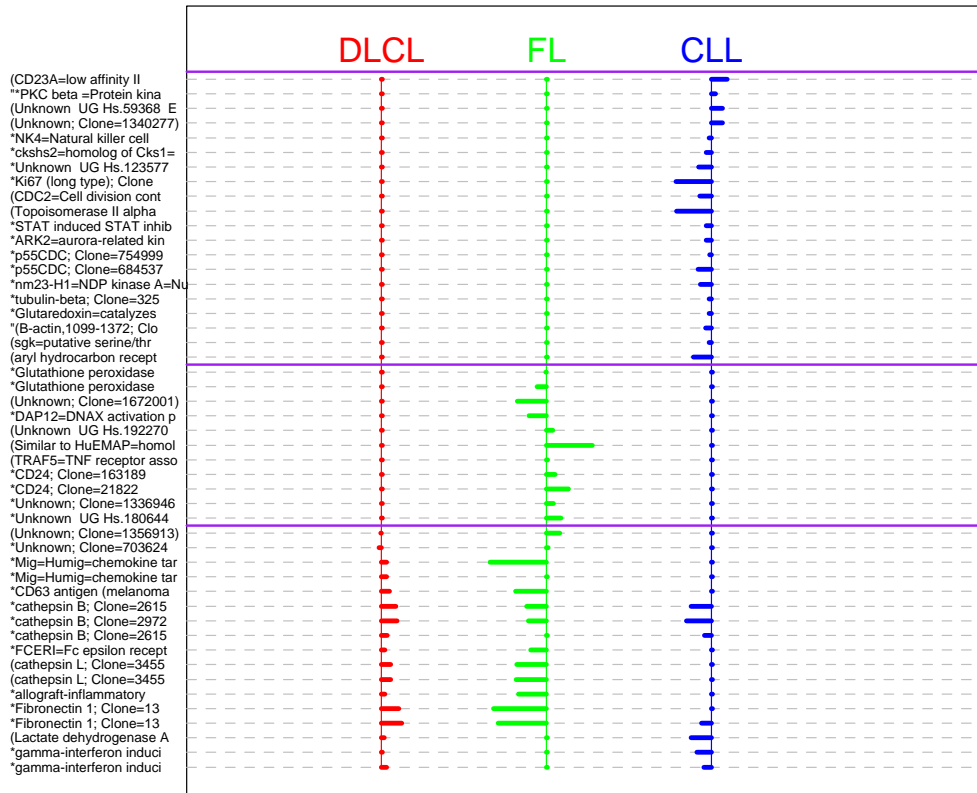


Figure 4: *Shrunken differences d_{ik} for the 48 genes having at least one non-zero difference.*

2.2 Finding the predictors that matter

Figure 4 shows the shrunken differences d_{ik} for only the 48 genes having at least one non-zero difference. Figure 5 shows the heat-map of the chosen 48 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 over-expressed and others under-expressed. Similarly the middle set of genes characterizes FL. The bottom set of genes in the figure tend to be over-expressed in DLCL, and under-expressed in FL and CLL.

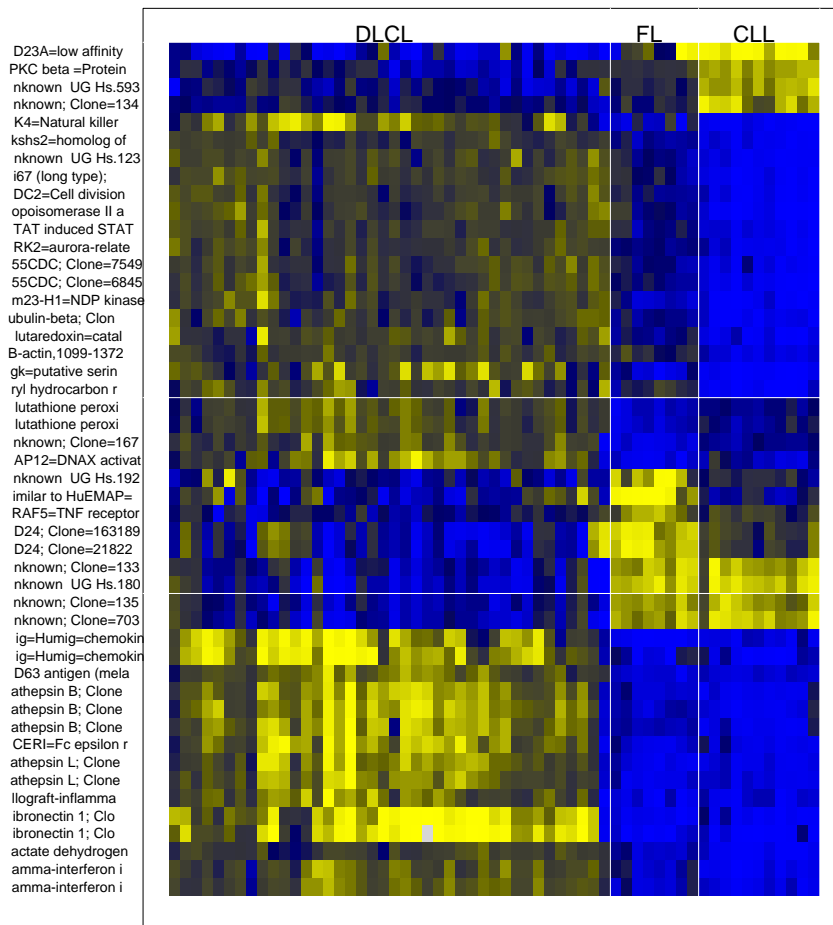


Figure 5: Heat-map of the chosen 48 genes. Within each of the horizontal partitions, we have ordered the genes by hierarchical clustering, and similarly for the samples within each vertical partition. The data for all 59 samples is shown.

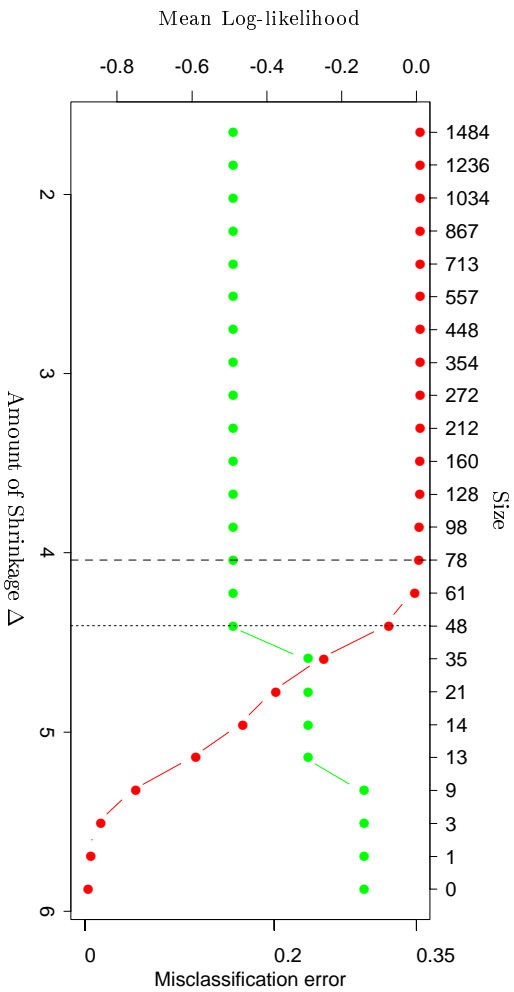


Figure 6: *Test set mean log-likelihood curve (red), and test set misclassification error curve (green). The latter has been translated so that it fits in the same plotting region. The broken line shows where the log-likelihood starts to dip, while the dotted line shows where the misclassification error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

2.3 The log-likelihood

It is quite common to have a small number of samples in each class, especially when the number of 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 (8), 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. They give a similar picture, although 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 $x^* = (x_1^*, x_2^*, \dots, x_p^*)$. We define the *discriminant score* for class k

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

The first part of (6) is simply the standardized squared distance of x^* to the k th 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 classification rule is then

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

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 can be used instead, or even uniform priors $\pi_k = 1/K$.

We can use the discriminant scores to construct estimates of the class probabilities, by analogy to Gaussian linear discriminant analysis:

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

The left panel of Figure 7 displays these probabilities for the lymphoma data. We used the centroids from the maximally shrunken 48 gene model ($\Delta = 4.41$) applied to the test set.

Now looking at 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. They are more extreme than those in the left panel. The rightmost probabilities are preferred, since they produce a higher log-likelihood score.

3 Capturing heterogeneity

In discriminating an “abnormal” from a “normal” group, the average gene expression may not differ between the groups. But the variability in expression may be greater in the abnormal group, due to heterogeneity in the abnormal population. This is illustrated in Figure 8. Nearest centroid classification will not work here, since the class centroids are not separated. To attack this problem, we can 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 1000 genes in 40 samples, 20 from a normal group 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. Nearest centroid shrinkage on the transformed features x'_{ij} showed a test error rate of near zero, with 150 or more non-zero genes. Figure 9 shows 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, 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 abnormal class is not heterogeneous, but instead mean-shifted. The right panel of Figure 9 investigates this. The expression of the first 200 genes in the abnormal class have 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 successively used the heterogeneity model in predicting radiation sensitivity from lymphoid cell lines [Reiger *et. al.*, in preparation)]. Since that work is not yet published, we cannot give details of the results here.

4 Adaptive choice of thresholds

In this section describe the procedure for adaptive threshold choice in the nearest shrunken centroid method. We define a scaling vector $(\theta_1, \theta_2, \dots, \theta_K)$;

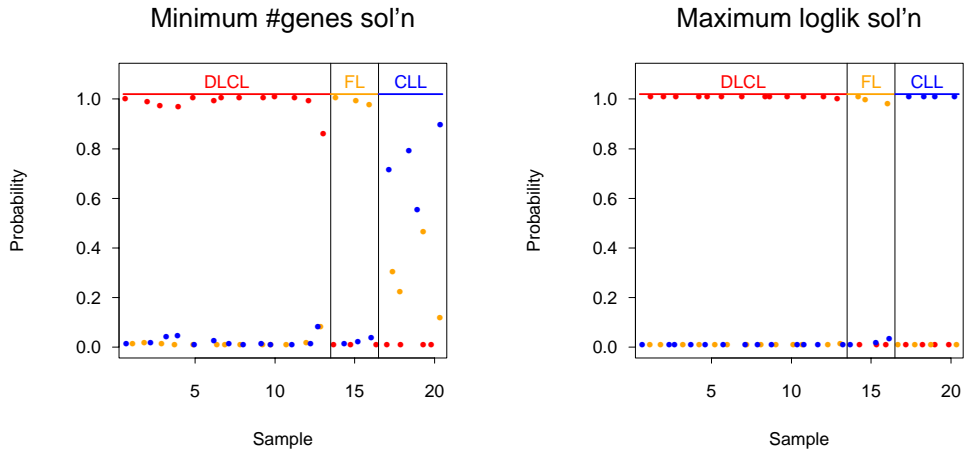


Figure 7: *Estimated test set probabilities using the 48 gene model from minimizing misclassification error (left) and the 78 gene model from maximizing the log-likelihood (right). Probabilities are partitioned by the true class. There are no classification errors in the test set.*

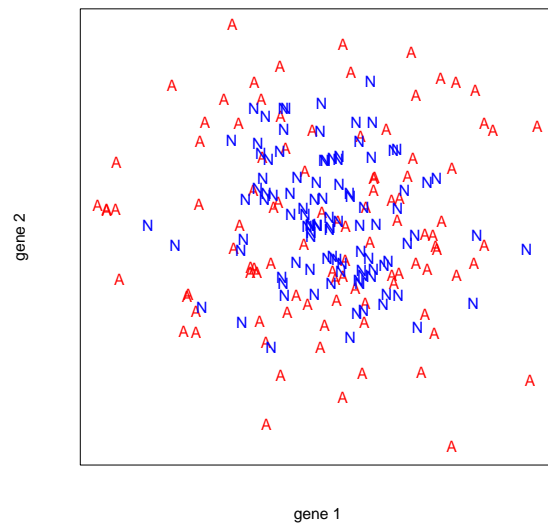


Figure 8: *Illustration of heterogeneity in gene expression. Abnormal group “A” has the same average gene expression as the normal group “N”, but shows larger variability.*

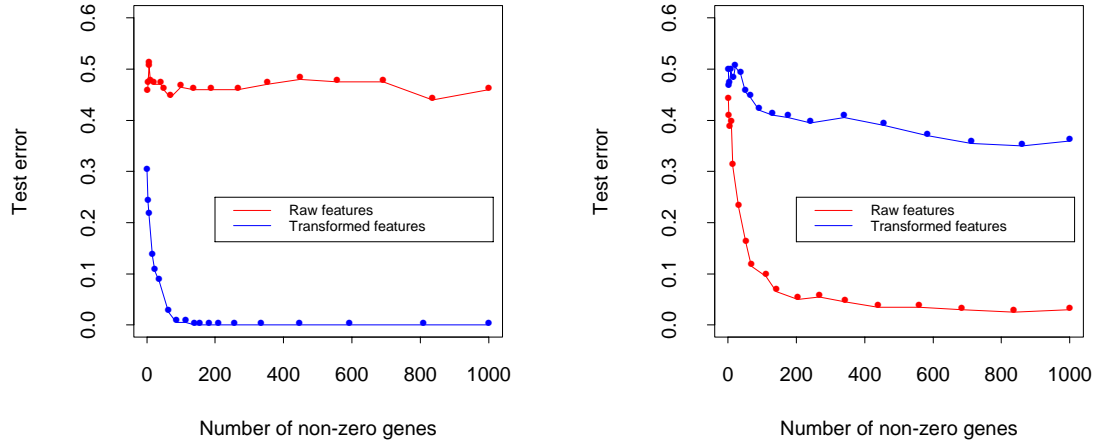


Figure 9: *Left panel: test error for data simulated from heterogeneous two-class problem, using nearest shrunken centroids on raw expression values (red) and transformed expression values $|x_{ij} - \bar{m}_i|$ (blue); Right panel: as in left panel, but data are simulated from mean-shifted two-class problem.*

initially we set $\theta_k = 1 \forall k$. These scalings are included in the denominator of expression (1), that is

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

We scale the values so that $\min_j(\theta_j) = 1$: values greater than one 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 for a number of iterations (here 10), and find the solution giving lowest average error, among the values of $(\theta_1, \theta_2, \dots, \theta_K)$ visited.

For the lymphoma data, we obtained the solution (1.88, 1.00, 1.52) which is the value we used to produce Figure 1. Most of the errors in the original solution occurred in class FL: the new thresholds are larger for classes DLCL and CLL, and hence much fewer genes are used to discriminate these classes. Remarkably, the total number of genes used has decreased from 2938 to 48, without raising the test error.

To test this procedure further, we simulated some data with ten samples in each of four classes, and 1000 genes. We ran two different simulations, with the results shown in the top and bottom panels of Figure 10. For a concise description, let $r(a, n)$ represent the number “a” repeated n times. All expression values were independent Gaussian with variance one. In the first simulation, the class centroids were $[r(3, 500), r(.4, 500)]$, $[r(.5, 100), r(0, 900)]$, $[r(0, 100), r(.5, 100), r(0, 800)]$ and $[r(0, 100), r(0, 100), r(.5, 100), r(0, 700)]$. Thus the first class is far from the others, in the space spanned by the first 500 genes. The top panel of Figure 10 shows the mean \pm one standard error of the test error over five simulations. The methods used were the default (equal) thresholds (red), and adaptive thresholds (green). The average value of the adaptive threshold was 2.0, 1.0, 1.0, 1.0. The adaptive threshold method has generally lower test error.

In the second simulation, the means in the four classes were $[r(.5, 300), r(0, 700)]$, $[r(.5, 150), r(-.5, 150), r(0, 700)]$, $[r(-.5, 150), r(.5, 150), r(0, 700)]$, and

$[r(-.5, 150), r(-.5, 150), r(0, 700)]$. 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 10 shows the results: again the adaptive threshold does better in terms of test error; the average value of the adaptive threshold was 1.4, 1.1, 1.2, 1.0. With equal thresholds, the majority of non-zero genes were in class 1: under the adaptive thresholds, the distribution was more balanced. Note however the test error for the adaptive method has its minima well beyond the true number of non-zero genes (300).

5 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:

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

This is sometimes known as *hard thresholding*. It differs from soft thresholding in that differences greater than Δ are unchanged, rather than shrunk towards 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 1000 genes and 40 samples, with 20 in each of two classes. For the first 100 genes, we added a random effect $\mu_i \sim N(0, .5^2)$ to each expression level in class two, for each gene i . Hence 100 of the 1000 genes are differentially expressed in the two classes, by varying amounts. The left panel of Figure 11 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; the right panel shows that soft thresholding does a much better job of estimating the gene expression differences.

6 Example: NCI cancer lines

This data is taken from Ross et al. (2000), and consists of measurements on 6830 genes on 61 cell lines. The samples have been categorized into 8 differ-

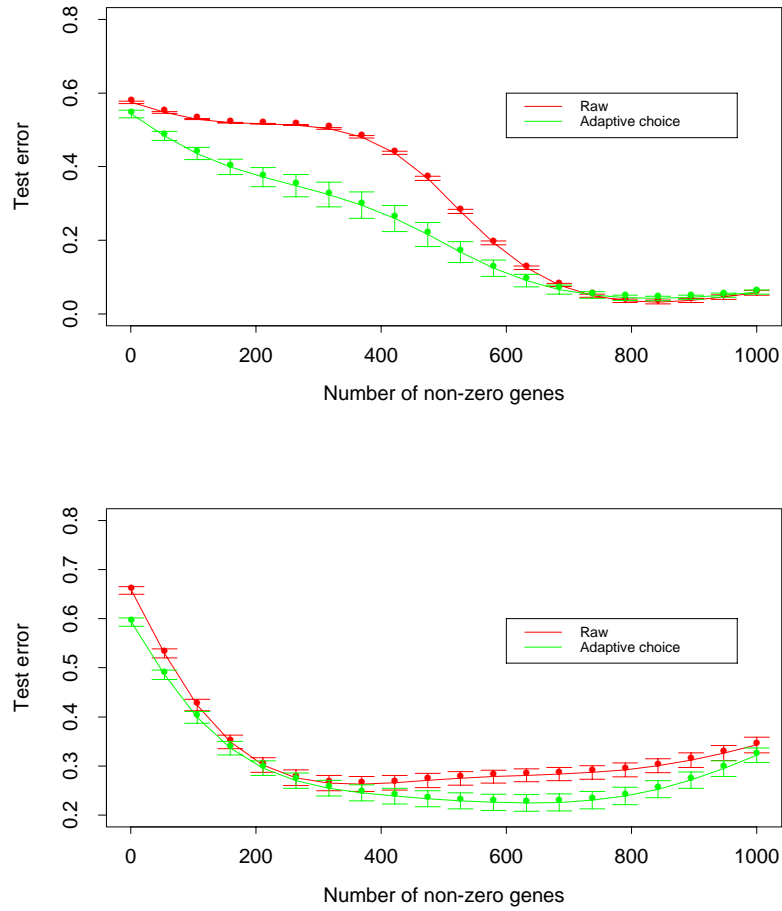


Figure 10: *Simulated data: mean \pm one standard error of the test error over five simulations, for default (equal) thresholds (red) and adaptive thresholds (green). In the setup for the top panel, the class centroids are unevenly spaced. In the bottom panel, the within-class variances are unequal.*

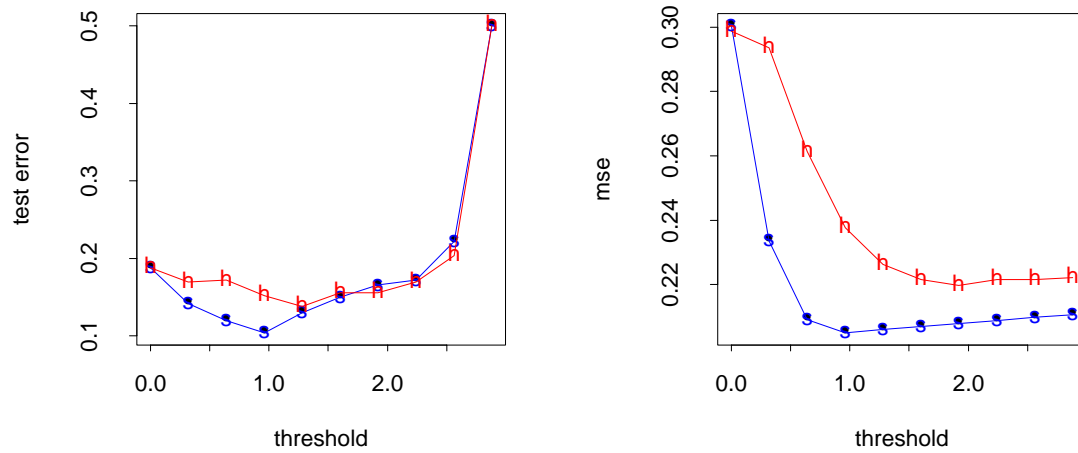


Figure 11: *Simulated data in two classes: left panel shows test misclassification error as the threshold Δ is varied, using hard thresholding (h) and soft thresholding (s). Right panel shows the estimation error $\sum(\hat{\mu}_i - \mu_i)^2/p$, where μ_i and $\hat{\mu}_i$ are the true and estimated difference in expression between class 1 and 2, for gene i .*

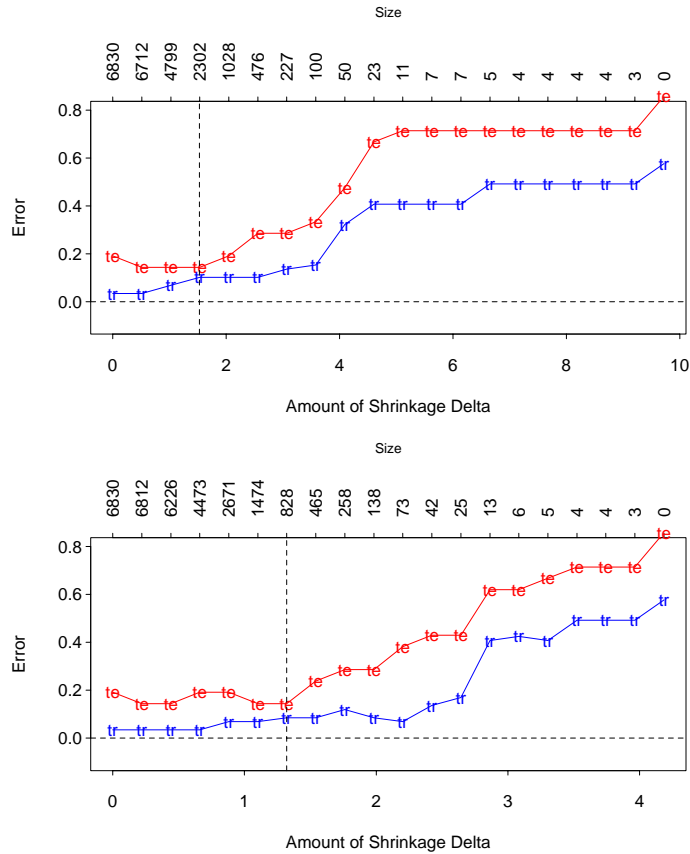


Figure 12: *NCI cancer cell lines: training and test error curves using default threshold scaling (top) and adaptive threshold scaling (bottom)*

ent cancer classes: **Breast**, **CMS**, **Colon**, **Leukemia**, **Melanoma**, **NSCLC**, **Ovarian** and **Renal**. 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. The results are shown in Figures 12 and 13. With adaptive thresholding, a minimum test error of $3/21 = 14.2\%$ is achieved with the 830 genes shown in Figure 13. The thresholds for the 8 classes were $(2.32, 1.52, 2.32, 2.32, 1.52, 1.00, 1.52, 1.23)$. By comparison, regularized linear discriminant analysis (Section 7) achieved a minimum test error of $4/21$.

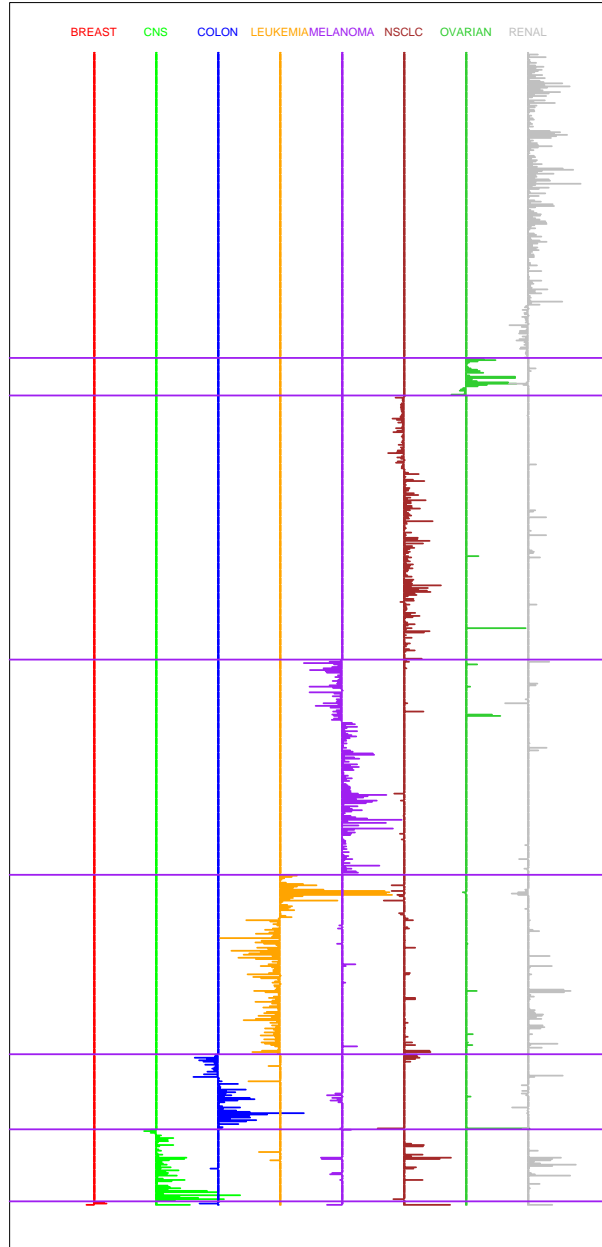


Figure 13: *NCI cancer cell lines: 830 genes with non-zero contributions for characterizing the 8 classes.*

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 in computing distance to centroids:

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

Here we are using a 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 that

- 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 $d'_{ik} = 0$, $k = 1, \dots, K$ due to the soft-thresholding in (5). Such genes contribute no discriminatory information in (6), and in fact cancel in (8)

Both our scores (6) and (11) are effectively *linear* in the 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 towards class discrimination), and multiply by -2 , we get

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

which is linear in the 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

$$\tilde{\delta}_k^{LDA}(x^*) = x^{*T} W^{-1} \bar{x}_k - \frac{1}{2} \bar{x}_k'^T W^{-1} \bar{x}_k' + \log \pi_k \quad (13)$$

Our nearest prototype scores (6) are a restricted form of the LDA scores (11). Since $p \gg n$, the within-class covariance W is singular, and so this solution is undefined. Nearest shrunken centroids makes W diagonal which solves the singularity problem; in addition it shrinks the \bar{x}_k .

Regularized Discriminant Analysis Friedman (1989) leaves the centroids alone, and modifies the covariance matrix in a different way:

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

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, nor centroid shrinking). A slightly modified version uses $W + \lambda D$, where $D = \text{diag}(s_1^2, s_2^2, \dots, s_p^2)$. 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.

In the NCI example of Figures 12 and 13, RDA yielded a minimum test error of 4/21, compared to 3/21 for nearest shrunken centroids. 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.

The neural network approach of Khan et al. (2001) can also be interpreted as a form of dampened discriminant analysis. In that paper they actually use a *linear* network, using the first 10 principal components (eigengenes). In statistical parlance, this is known as principal components regression, and requires no iterative learning procedure and learning curves. Principal component regression is a hard-thresholded version of ridge regression. The authors also use a model-averaging procedure, similar to bagging Breiman (1996) but based on three-fold cross-validation, to regularize the procedure further. Their procedure is far more complex than nearest shrunken centroids. With so many genes and so few samples, it is very likely that restricted versions of simpler statistical methods will do as well or better than neural networks.

8 Nearest centroid classifier versus LDA

As discussed in the previous section, the nearest centroid classifier is equivalent to Fisher's linear discriminant analysis (LDA) 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 (unshrunk) nearest centroid classifier and standard (full within covariance) LDA. Now 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 . Let x_j be a p -vector of gene expression values in class j . Suppose $x_1 \sim N(0, \Sigma)$, $x_2 \sim N(\mu, \Sigma)$ with Σ being a full (non-diagonal) matrix. Then LDA gives the maximum likelihood unbiased estimate of μ , while nearest centroids yields a biased estimate. However the LDA method estimates more parameters ($p + p^2/2$) than nearest centroid procedure ($2p$) and hence will 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 $n = 40$: twenty samples in each of two classes. We set the ij th element of Σ to $\rho^{|i-j|}$ where ρ was varied from 0 to .8. Each of the components of the mean vector μ were set to ± 1 at random: such a mixed vector is needed to give full LDA a potential advantage over LDA with a diagonal covariance. The results of 20 simulations from this model are shown in Figure 14. Bias, variance and mean-squared error refer to estimation of μ . 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.

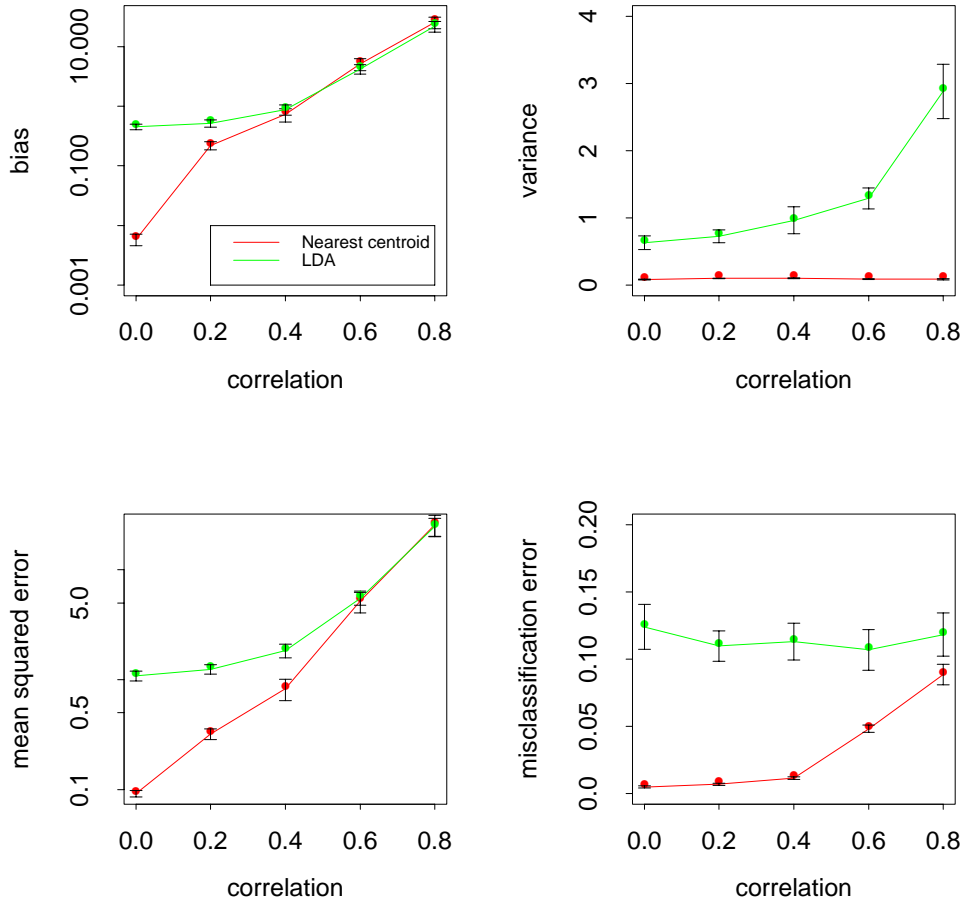


Figure 14: *Simulation results: bias, variance (top panels) and mean squared error, misclassification error (bottom panels) for linear discriminant analysis and nearest centroid classifier. Details of simulation are given in the text. Nearest centroid classifier outperforms LDA because of its smaller variance.*

9 Discussion

The nearest shrunken centroid classifier is potentially useful in any high-dimensional classification problem. Besides its application to gene expression arrays, it could also be applied to other kinds of emerging genomic data including protein arrays and SNP arrays.

Our proposal can also be applied in unsupervised problems. For example, it is by now standard to use hierarchical clustering methods on expression arrays to discover clusters in the samples (Eisen et al. 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 shrunken nearest centroid method are straightforward. One important detail: in the denominator of the statistics d_{ik} in (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 s_i over the set of genes. A similar strategy was used in the SAM methodology of (Tusher et al. 2001). We are currently developing a program similar to SAM, to implement nearest shrunken centroid classification.

References

- Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., J., P., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R. Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000), 'Identification of molecularly and clinically distinct subtypes of diffuse large b cell lymphoma by gene expression profiling', *Nature* **403**, 503–511.
- Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **26**, 123–140.
- Donoho, D. & Johnstone, I. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.
- Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci., USA*. **95**, 14863–14868.

- Friedman, J. (1989), ‘Regularized discriminant analysis’, *Journal of the American Statistical Association* **84**, 165–175.
- 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., Botstein, D. & Brown, P. (2001), ‘Supervised harvesting of expression trees’, *Genome Biology* **2(1)**, 1–12.
- 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., Fehrl, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Bor, A. & Trent, J. (2001), ‘Gene-expression profiles in hereditary breast cancer’, *N. Engl. J. Med.* **344**, 539–548.
- 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.
- Ross, D., Scherf, U., Eisen, M., Perou, C., Spellman, P., Iyer, V., Rees, C., Jeffery, S., Van de Rijn, M., 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.
- Tibshirani, R., Hastie, T., Narasimhan, B. & Tibshirani, R. (2001), Diagnosis of multiple cancer types by shrunken centroids of gene expression. To appear, Proc. Natl. Acad. Sci.
- Tusher, V., Tibshirani, R. & Chu, C. (2001), ‘Significance analysis of microarrays applied to transcriptional responses to ionizing radiation’, *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121.