

# Supervised Learning Methods for Gene-Expression Data

G.J. McLachlan <sup>\*,1</sup>,  
Ambroise, C., Ben-Tovim Jones, L., and Zhu, X.

*Department of Mathematics, University of Queensland, St. Lucia, Brisbane 4072,  
Australia*

---

## Abstract

Clinicians are starting to use microarrays as diagnostic tools for tumor tissue classification, yet in current studies expressions over thousands of genes are measured over relatively few tissue samples. We consider the classification of a tumor tissue sample on the basis of its expression signature, which is the vector containing the expression levels for very many (possibly thousands) of genes. We focus on the supervised problem (discriminant analysis) and demonstrate the caution that needs to be exercised in assessing the error rates of a discriminant (prediction) rule that has been formed from a relatively small ‘optimal’ subset of the genes. The subset is ‘optimal’ in the sense that it has been chosen to optimize some criterion. The discriminant rule adopted for the purposes of the demonstration is the support vector machine. Recursive feature elimination (RFE) is implemented for the selection of suitable genes for use in the support vector machine in the case of multiple classes. The results are demonstrated on various microarray data sets available in the bioinformatics literature.

*Key words:* Microarrays; Discriminant analysis; Selection bias; External cross-validation; Support vector machine

---

## 1 Introduction

Microarray technology in particular, and large-scale screening approaches in general, lead to the challenging problem of learning from high-dimensional

---

\* Corresponding author.

*Email address:* [gjm@maths.uq.edu.au](mailto:gjm@maths.uq.edu.au) (G.J. McLachlan).

<sup>1</sup> Phone: +61 7 3365 2150, Fax +61 7 3365 1477

data. In a DNA microarray experiment, the expression levels of tens of thousands of genes can be simultaneously measured for a single tissue sample. At present a great challenge in Medicine is to determine how microarrays can be used as diagnostic tools, and in so doing to redefine disease at the level of gene expressions. Clinicians are interested in finding a subset of genes (marker genes) that can improve diagnosis based on standard clinical criteria, for example in determining whether breast cancer patients fall into good- or poor-prognosis groups, as in van 't Veer et al. (2002), and thus aid in therapeutic management. A major obstacle to both supervised and unsupervised methods for determining these marker genes is the limited size of current microarray studies, where at most the number of tissue samples is in the hundreds. The dimension of the feature space (the number of genes) is much greater than the number of observations (the number of tissues), and poses a nonstandard problem in parametric classification (cluster and discriminant analyses) as described in detail below.

Although particular experiments vary considerably in their design, the data generated by microarray experiments can be viewed as a matrix of expression levels. For  $M$  microarray experiments (corresponding to  $M$  tissue samples), where we measure the expression levels of  $N$  genes in each experiment, the results can be represented by the  $N \times M$  matrix. For each tissue, we can consider the expression levels of the  $N$  genes, called its *expression signature*. Conversely, for each gene, we can consider its expression levels across the different tissue samples, called its *expression profile*. The  $M$  tissue samples might correspond to each of  $M$  different patients or, say, to samples from a single patient taken at  $M$  different time points. The expression levels are taken to be the measured (absolute) intensities for oligonucleotide microarrays and the ratios of the intensities for the Cy5-channel (red) images and Cy3-channel (green) images for cDNA microarrays; see, for example, Dudoit et al. (2002b). It is assumed that one starts the clustering process with preprocessed (relative) intensities, such as those produced by RMA (for Affy data), loess-modified log ratios, or differences of logged/generalized-logged data; see, for example, Parmigiani et al. (2003), Huber et al. (2003), Irizarry et al. (2003), Rocke and Durbin (2003), and Speed (2003).

## 2 Unsupervised Classification

Concerning the unsupervised classification of microarray data, there are two distinct clustering problems. One problem concerns the clustering of the tissues on the basis of the genes. The clusters of tissues can play a useful role in the discovery and understanding of new subclasses of diseases. The second problem concerns the clustering of the genes on the basis of the tissues. The clusters of genes obtained can be used to search for genetic pathways or groups of genes

that might be regulated together.

In the past, hierarchical methods have been the primary clustering tool employed to cluster microarray data. The hierarchical algorithms have been mainly applied heuristically to these cluster analysis problems. Further, a major limitation of these methods is their inability to determine the number of clusters. Thus there is a need for a model-based approach to these clustering problems. To this end, McLachlan et al. (2002) developed a mixture model-based algorithm (EMMIX-GENE) for the clustering of tissue samples. Among other work on model-based approaches to the clustering of gene expression data, there are the studies of Yeung et al. (2001, 2003), Ghosh and Chinnaiyan (2002), Medvedovic and Sivaganesan (2002), and Liu et al. (2003). A Bayesian approach is adopted in the latter two papers.

In the sequel, we shall focus exclusively on the supervised classification of microarray data.

### 3 Supervised Classification

As explained by Xiong et al. (2001), there is increasing interest in changing the emphasis of tumor classification from morphologic to molecular. In this context, the problem is to construct a discriminant (prediction) rule  $r(\mathbf{y})$  that can accurately predict the class of origin of a tumor tissue with feature vector  $\mathbf{y}$ , which is unclassified with respect to a known number  $g (\geq 2)$  of distinct tissue types, denoted here by  $C_1, \dots, C_g$ . Here the feature vector  $\mathbf{y}$  contains the expression levels on a very large number  $N$  of genes (features). In applications concerned with the diagnosis of cancer, one class ( $C_1$ ) may correspond to cancer and the other ( $C_2$ ) to benign tumors. In applications concerned with patient survival following treatment for cancer, one class ( $C_1$ ) may correspond to the good-prognosis group and the other  $C_2$  to the poor-prognosis group. Also, there is interest in the identification of marker genes that characterize the different tissue classes. This is the feature selection problem.

In order to train the discriminant (prediction) rule, there are available training data  $\mathbf{t}$  consisting of  $n = M$  tissue samples of known classification. These data are obtained from  $M$  microarrays, where the  $j$ th microarray experiment gives the expression levels of the  $p = N$  genes in the  $j$ th tissue sample  $\mathbf{y}_j$  of the training set. Here  $\mathbf{y}_j$  is the gene expression signature vector for the  $j$ th tissue. The class of origin of the  $j$ th tissue sample  $\mathbf{y}_j$  is denoted by the  $g$ -dimensional vector of zero-one class labels  $\mathbf{z}_j$  ( $j = 1, \dots, n$ ). We write the sample rule formed from the training data  $\mathbf{t}$  as  $r(\mathbf{y}; \mathbf{t})$  to show its dependence on  $\mathbf{t}$ . In the sequel, we shall refer to the tissue sample  $\mathbf{y}_j$  simply as a tissue, since in statistics the collection of the  $n$  tissue samples from  $\mathbf{y}_1, \dots, \mathbf{y}_n$  that belong

to a given class would be referred to as a sample.

In a standard discriminant analysis, the number of training observations  $n$  is usually much larger than the number of feature variables  $p$ . But in the present context of microarray data, the number of tissue samples ( $n = M$ ) is typically between 10 and 100, and the number of genes ( $p = N$ ) is in the thousands. This presents a number of problems. Firstly, the prediction rule  $r(\mathbf{y}; \mathbf{t})$  may not be able to be formed using all  $p$  available genes. For example, the pooled within-class sample covariance matrix  $\mathbf{S}$  required to form Fisher's linear discriminant function is singular if  $n < g + p$ . Secondly, even if all the genes can be used as, say, with the nearest-centroid rule or a support vector machine (SVM), the use of all the genes may allow the noise associated with genes of little or no discriminatory power, to inhibit and degrade the performance of the rule  $r(\mathbf{y}; \mathbf{t})$  in its application to unclassified data. That is, although the apparent error rate  $A$  (the proportion of the training tissues misallocated by  $r(\mathbf{y}; \mathbf{t})$ ) will decrease as it is formed from more and more genes, its error rate in classifying tissues outside of the training set will eventually increase. That is, the generalization error of  $r(\mathbf{y}; \mathbf{t})$  will be increased if it is formed from a sufficiently large number of genes. Hence, in practice, consideration has to be given to implementing some procedure for reducing the dimension of the feature vector of genes to be used in constructing the rule  $r(\mathbf{y}; \mathbf{t})$ .

## 4 Reducing the Dimension of the Feature Space of Genes

### 4.1 *Principal Components*

A common approach is to carry out a principal component analysis (PCA) and work with the leading components. The PCA can be implemented via a singular value decomposition as in West et al. (2001) and Liu et al. (2003). The disadvantages of this approach are that the PCA does not take into account the class structure of the genes, and genes that show a large variation across the tissues may not be differentially expressed. Also, as the principal components are linear combinations of the original number of genes, biological interpretation of the components is not straightforward.

### 4.2 *Partial Least Squares*

One method that does take into account the class structure of the tissue samples in reducing the dimension of the feature space is partial least squares. However, it still suffers from the same interpretation difficulties as with prin-

principal components, as the components are linear combinations of all the genes.

Partial least squares for the supervised classification of tissue samples has been considered by Nguyen and Rocke (2001, 2002a, 2002b), using logistic discrimination and the normal-based quadratic rule. The response vector is taken to consist of the class-indicator variables. Nguyen and Rocke (2002a) demonstrated in their study that if the top genes for discrimination purposes were selected before performing the principal component analysis, then it would give similar results to partial least squares.

### 4.3 Ranking of Genes

One common way of approaching the gene selection problem is to perform a preliminary ranking of genes on the basis of a fast computable criterion and then arbitrarily select a number of the best-ranked genes. Then either a discriminant rule is formed on the basis of these selected genes or further selection is undertaken before constructing the rule.

A commonly used criterion for ranking the individual genes  $y_v = (\mathbf{y})_v$  ( $v = 1, \dots, p$ ) is the ratio of the between-class sum of squares to the within-class sum of squares on their degrees of freedom,

$$F_v = (\mathbf{B})_{vv}/(\mathbf{S})_{vv}, \quad (1)$$

where

$$\mathbf{B} = \sum_{i=1}^g n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T / (g - 1) \quad (2)$$

and

$$\mathbf{S} = \sum_{i=1}^g \sum_{j=1}^n z_{ij} (\mathbf{y}_j - \bar{\mathbf{y}}_i)(\mathbf{y}_j - \bar{\mathbf{y}}_i)^T / (n - g), \quad (3)$$

respectively, and where

$$\bar{\mathbf{y}}_i = \sum_{j=1}^n z_{ij} \mathbf{y}_j / n_i \quad (4)$$

and

$$\bar{\mathbf{y}} = \sum_{j=1}^n \mathbf{y}_j / n. \quad (5)$$

In (3) and (4), the  $z_{ij}$  denote the class labels, where  $z_{ij}$  is one or zero, according as  $\mathbf{y}_j$  belongs or does not belong to class  $C_i$  ( $i = 1, \dots, g$ );  $n_i = \sum_{j=1}^n z_{ij}$ .

Under the null hypothesis that the  $v$ th gene has the same variance in each class, the statistic  $F_v$  has an  $F$ -distribution with  $g - 1$  and  $n - g$  degrees of freedom. The use of (1) is equivalent to the likelihood ratio statistic  $-2 \log \lambda$  for the test of no differences between the means of the classes under the assumption of equal covariance matrices for the class-conditional distributions of the genes. Also, in the case of  $g = 2$  classes, it is equivalent to the usual two-sample (pooled) Studentized  $t$ -statistic.

Another criterion is to use the apparent error rate  $A_v$  of the rule  $r(\mathbf{y}_v; \mathbf{t}_v)$ , where the latter is formed using just the training data on the  $v$ th gene (Braganeto et al., 2004). Alternatively, we may use the (leave-one-out) cross-validated error rate  $A_v^{(CV)}$ . A further criterion is to rank the genes on the basis of the absolute values of their coefficients in the linear form of  $r(\mathbf{y}; \mathbf{t})$  for an SVM formed with linear kernel. This is to be discussed further in the next section. There are also rules where the ranking is being done implicitly in their construction; for example, nearest-shrunken centroids to be considered later on.

#### 4.4 Grouping of Genes

Another way to handle the problem of having to form a discriminant rule from a very large number of feature variables (genes) is to put the genes into groups either by some clustering method or by some supervised selection procedure that makes use of their known class labels. There is now a variety of ways proposed in the literature for the grouping of the genes. Having so grouped the genes, a discriminant rule can be formed from the genes (metagenes) selected to represent each group. Recent papers that make use of this approach include Dettling and Bühlmann (2002), Liu et al. (2002), Díaz-Uriarte (2003), Hastie et al. (2001a), and Goh, Kasabov and Song (2004).

## 5 SVM with Recursive Feature Elimination (RFE)

There are many techniques available for supervised learning; see, for example, Hastie, Tibshirani and Friedman (2001), Hand, Mannila and Smyth (2001), McLachlan (1992), and Ripley (1996). However, most of these techniques are not likely to work “off the shelf”, as expression data present special challenges. The difficulty is that the number of feature variables (genes) is large compared with the number of observations (tissue samples), and some of the genes tend

to be highly correlated.

In this paper, we concentrate on the use of support vector machines, which are becoming increasingly popular classifiers in many areas, including microarrays (Brown et al., 2000; Furey et al., 2000; Guyon et al., 2002; Ramaswamy et al., 2001). Advantages of an SVM in the present context, where the number of feature variables (genes)  $p$  is so large relative to the sample size  $n$ , are that it is able to be fitted to all the genes and that its performance appears not to be too affected by using the full set of genes. However, in practice, some form of gene selection would generally be contemplated. Another advantage of the SVM (with a linear kernel) is that gene selection can be undertaken fairly simply using the vector of weights as the criterion.

For an SVM with linear kernel, the rule  $r(\mathbf{y}; \mathbf{t})$  can be written as

$$r(\mathbf{y}; \mathbf{t}) = \text{sign}(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \mathbf{y}), \quad (6)$$

where  $\hat{\beta}_v = (\hat{\boldsymbol{\beta}})_v$  denotes the coefficient of the expression level  $y_v$  for gene  $v$ .

Vapnik (1998) considered support vector machines (SVMs) in the case of  $g = 2$  classes. The method aims to find the separating hyperplane

$$\beta_0 + \boldsymbol{\beta}^T \mathbf{y} = 0$$

that is maximally distant from the training data of the two classes. When the classes are linearly separable, the hyperplane is located so that it has maximal margin (that is, so that there is maximal distance between the hyperplane and the nearest point in any of the classes), which should lead to better performance on test data. When the data are not separable, there is no separating hyperplane; in this case, we still try to maximize the margin but allow some classification errors subject to the constraint that the total error (distance from the hyperplane on the wrong side) is less than a constant. Support vector machines have since been generalized to handle multiple classes as described, for example, in McLachlan et al. (2004, Chapter 6).

As shown by Guyon et al. (2002), a good guide to the relative importance of the genes in this SVM is given by the relative size of the absolute values of their fitted coefficients  $\hat{\beta}_v$  (that is, the weights). Hence a ranking of the discriminatory power of the genes can be given by ranking the genes from top to bottom on the basis of the absolute values of the weights  $\hat{\beta}_v$ .

We consider here the selection procedure of Guyon et al. (2002), who used a backward selection procedure, which they termed recursive feature elimination (RFE). It considers initially all the available genes, which are ranked according to their weights and the bottom-ranked genes discarded. The SVM is then

refitted to the remaining genes, which are then reranked according to their new weights. Again, the bottom-ranked genes are discarded, and so on.

In the applications to follow on microarray data, we first discarded enough bottom-ranked genes so that the number retained was the greatest power of 2 (less than the original number of genes). We then proceeded sequentially to discard half the current number of genes on each subsequent step. Initially, the error rate usually falls as genes are deleted, but generally, it will start to rise once a sufficiently large number of genes have been deleted.

The error rate at any stage can be assessed by undertaking an external cross-validation. An alternative is to use the 0.632+ estimator of Efron and Tibshirani (1997). As discussed in detail later, there may be a considerable selection bias present in the apparent error rate when a reduced number of genes is selected from a very large number. Thus it is not sufficient to use an ordinary (internal) cross-validation, as employed by several authors in the past, including Guyon et al. (2002). With an external cross-validation of the error rate at a given stage of the selection process, at each split of the original training data into training and validation subsets, the selection process has to be implemented from the beginning on the basis of the training subset. That is, with the present selection process of RFE, the SVM has to be fitted to all the genes and then the process continued until the present stage of the selection process has been reached. Then the rule for this newly selected subset of genes is applied to the validation subset.

## 6 Comparison of SVM with Nearest-Shrunken Centroids

For discriminant rules formed from a limited number of training data of very high dimension, it seems that the selection method and the number of selected genes are more important than the classification method for constructing a reliable prediction rule. To illustrate this point, we compare the SVM with nearest-shrunken centroids as proposed by Tibshirani et al. (2002, 2003). For high-dimensional data, they considered a modification to the nearest-centroid rule. They termed this approach nearest-shrunken centroids. It is directly applicable in the present context of the supervised classification of tissue samples. With this approach the usual estimates of the class means  $\bar{y}_i$  are shrunk toward the overall mean  $\bar{y}$  of the data.

The nearest-centroid rule is given by

$$r(\mathbf{y}; \mathbf{t}) = \arg \min_i \sum_{v=1}^p (y_v - \bar{y}_{iv})^2 / s_{iv}^2 - \log(\hat{\pi}_i), \quad (7)$$

where  $y_v$  is the  $v$ th element of the feature vector  $\mathbf{y}$  and  $\bar{y}_{iv} = (\bar{\mathbf{y}})_{iv}$ . In the definition (7) of the nearest-centroid rule, we replace the sample mean  $\bar{y}_{iv}$  of the  $v$ th gene by its shrunken estimate

$$\bar{y}_{iv}^* = \bar{y}_{iv} + m_i s_i d_{iv}^* \quad (i = 1, \dots, g; v = 1, \dots, p), \quad (8)$$

where

$$d_{iv}^* = \text{sign}(d_{iv})(|d_{iv}| - k)_+ \quad (9)$$

and

$$d_{iv} = \frac{\bar{y}_{iv} - \bar{y}_v}{m_i s_v}, \quad (10)$$

and where  $m_i = (n_i^{-1} - n^{-1})^{\frac{1}{2}}$ . In (10),  $s_v$  is the pooled within-class sample standard deviation of gene  $v$ ; that is,  $s_v^2$  is the  $v$ th diagonal of the pooled (bias-corrected) sample covariance matrix  $\mathbf{S}$ . Also, in (9), the subscript plus means *positive part*; that is,  $a_+ = a$ , if  $a > 0$ , and zero otherwise. The denominator  $m_i s_v$  in (10) is the standard error of the numerator  $\bar{y}_{iv} - \bar{y}_v$ .

An attractive property of this shrunken approach is that many of the genes are eliminated as far as their contribution to the sample rule if the threshold  $k$  is chosen sufficiently large. For if  $k$  causes  $d_{iv}$  to shrink to zero, then  $\bar{y}_{iv}^*$  is the same as the overall mean  $\bar{y}_v$ , and so gene  $v$  does not contribute to the class decision based on (8).

Tibshirani et al. (2003) use ten-fold cross-validation to choose the value of the threshold  $k$ . It is external in the sense that for a given value of  $k$  the selection of genes is carried out separately on each of the ten cross-validation trials. They also consider adaptive choice of thresholds, soft versus hard thresholding, and ways to capture heterogeneity in the class training data.

## 6.2 Application to Alon Data

We apply the SVM and the nearest-shrunken centroids methods to the colon cancer data of Alon et al. (1999). In this study Affymetrix oligonucleotide arrays were used to measure the expressions of over 6,500 human genes in 40 tumor and 22 normal colon tissue samples. These samples were taken from 40 different patients, so that 22 patients supplied both a tumor and a normal tissue sample. Alon et al. (1999) focused on the 2,000 genes with highest minimal intensity across the samples, and the data set as analyzed here consists of  $N = 2,000$  genes on a total of  $M = 62$  tissues, made up of  $n_1 = 40$  colon tissues and  $n_2 = 22$  normal tissues.

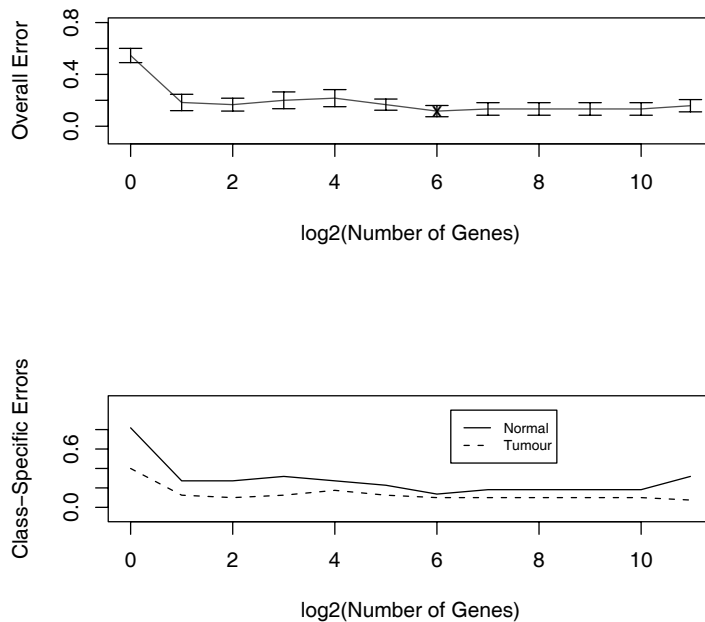


Fig. 1. Overall and class-specific error rates for SVM with RFE applied to Alon data.

On comparing the results in Figures 1 and 2, it can be seen that the SVM with the genes selected by RFE has a similar error rate over the genes to nearest shrunken-centroids.

As noted in Guyon et al. (2002) on the use of RFE with SVM, removing one variable at a time is more accurate than removing chunks of variables at a time. So we reapplied the SVM with the RFE implemented as before down to 128 genes, but from then on eliminating only one gene at a time. It led to a slightly better error rate of 10% for 10 to 18 genes on the Alon data.

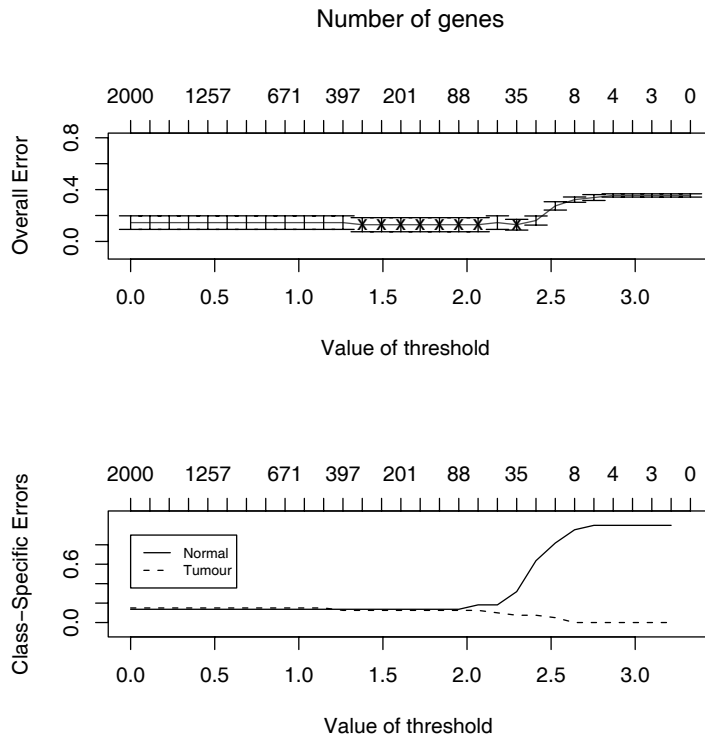


Fig. 2. Plot of overall and class-specific error rates for nearest-shrunken centroids applied to Alon data.

## 7 Selection Bias: SVM with RFE

We now consider the selection bias associated with forming a discriminant rule from thousands of genes, using the results of Ambroise and McLachlan (2002) obtained for the support vector machine. The size of the selection bias has also been investigated by Nguyen and Rocke (2002b) for multiclass discrimination via the logistic and the normal quadratic discriminant rules, using partial least squares.

Ambroise and McLachlan (2002) investigated the magnitude of the selection bias and its correction for an SVM (with linear kernel) and Fisher’s linear discriminant function in their application to two cancer data sets; the colon data of Alon et al. (1999) as described previously, and also the leukemia data of Golub et al. (1999). Here we report the results from the Alon data, though similar results were obtained for the Golub data set, which comprised  $M = 72$  tissues made up of  $n_1 = 47$  patients with acute lymphoblastic leukemia (ALL) and  $n_2 = 25$  patients with acute myeloid leukemia (AML).

To illustrate the size of the selection bias for the Alon data set, Ambroise and McLachlan (2002) split it into a training set and a test set, each of size 31, by sampling without replacement from the 40 tumor and 22 normal

tissues separately, so that each set contained 20 tumor and 11 normal tissues. The training set is used to carry out gene selection and to form the apparent error rate  $A$ , the (leave-one-out) cross-validated error rate  $A^{(CV)}$  using just internal validation, and the external ten-fold cross-validated rate  $A^{(CV10E)}$  for a selected subset of genes. An unbiased error-rate estimate is given by the test error ( $T$ ), equal to the proportion of tissues in the test set misallocated by the rule. They calculated these quantities for 50 such splits of the colon data into training and test sets.

The average values of the error-rate estimates are plotted in Figure 3, where the apparent error  $A$ , the (leave-one-out) cross-validated error  $A^{(CV)}$ , the external ten-fold cross-validated error  $A^{(CV10E)}$ , the 0.632+ bootstrap error estimate  $B^{(0.632+)}$ , and the test error are denoted by  $A$ ,  $CV$ ,  $CV10E$ ,  $B.632+$ , and  $T$ , respectively. It can be seen from Figure 3 that the true prediction error rate as estimated by  $T$  is not negligible, being above 15% for all selected subsets. The lowest value of 17.5% occurs for a subset of  $2^6$  genes.

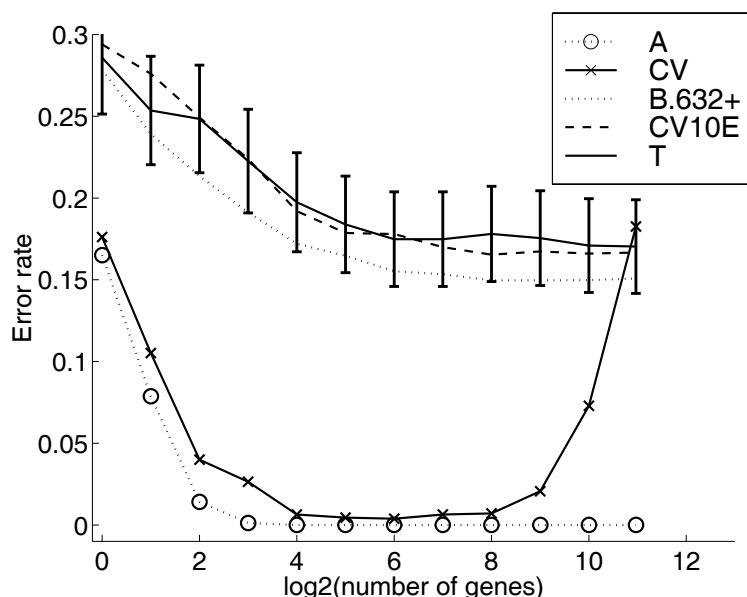


Fig. 3. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each.

Concerning the estimation of the prediction error by external ten-fold cross-validation and the bootstrap, it can be seen that  $A^{(CV10E)}$  has little bias for both data sets.

It can be seen from Figure 3 that the estimated prediction rate according to  $B^{(0.632+)}$  and  $A^{(CV10E)}$  remains essentially constant as genes are deleted in the SVM, until around about 64 or so genes when these estimates start to rise sharply. The internal cross-validated error  $A^{(CV)}$ , which is uncorrected for selection bias, also starts to rise then. Hence feature selection provides

essentially little improvement in the performance of the SVM rule for the two considered data sets. But it does show that the number of genes can be greatly reduced without increasing the prediction error.

## 8 Selection Bias: Noninformative Data

To further illustrate this selection bias, Ambroise and McLachlan (2002) generated a no-information training set by randomly permuting the class labels of the colon tissue samples. For each of 20 no-information sets so obtained, an SVM rule was formed by selecting genes by the RFE method and the apparent error  $A$  and the leave-one-out cross-validated error  $A^{(CV)}$  were calculated. The average values of these two error rates and the no-information error  $\gamma$  over the 20 sets are plotted in Figure 4, with the average value of the  $A^{(CV10E)}$  and  $B^{(.632+)}$  error estimates that correct for the selection bias. The no-information error  $\gamma$  is the error rate that would apply if the distribution of the class-membership label of the  $j$ th feature vector did not depend on its feature vector  $\mathbf{y}_j$ . It is estimated by

$$\gamma = \sum_{i=1}^g p_i(1 - q_i), \quad (11)$$

where  $p_i$  is the proportion of the (original) training data from the  $i$ th class and  $q_i$  is the proportion of them assigned to the  $i$ th class by  $r(\mathbf{y}; \mathbf{t})$ .

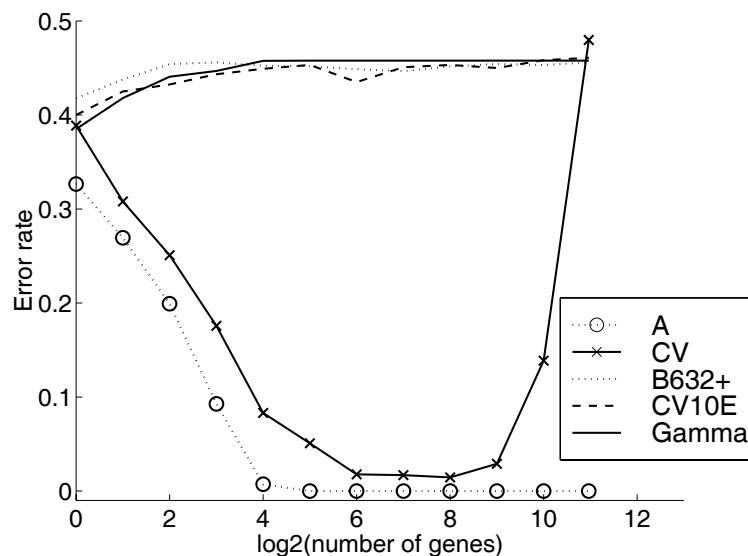


Fig. 4. Error rates of the SVM rule averaged over 20 noninformative samples generated by random permutations of the class labels of the colon tumor tissues.

It can be seen that, although the feature vectors have been generated independently of the class labels, we can form an SVM rule that has not only an average zero apparent error  $A$ , but also an average  $A^{(CV)}$  error close to zero for a subset of 128 genes and around 20% for only eight genes in the selected subset.

## 9 Application of SVM with RFE

We consider here the breast cancer microarray data of van de Vijver et al. (2002). Previously, van 't Veer et al. (2002) had used supervised methods to identify 70 marker genes with expression profiles associated with the risk of early metastasis in breast cancer patients. They used oligonucleotide arrays to measure the expressions of 24,881 genes in 78 sporadic (non-BRCA carrier) breast tumors, where the sample group was selected on the basis of patient outcome (34 patients developed distant metastases within 5 years, while 44 remained disease-free after 5 years). The study of van de Vijver et al. involved a larger group of sporadic breast cancer patients. They ran oligonucleotide arrays as above for tumor samples taken from 295 patients, of which 61 of the original 78 were included, as well as 234 patients not chosen on the basis of outcome. Based on the expressions of just the 70 marker genes, they classified tumors (patients) into either good- or poor-prognosis categories. For this larger data set, only the expression levels of the  $N=70$  marker genes for the  $M=295$  tissues have been made publicly available, and it is these expression values which we used for the SVM. Note that given the way these training data have been compiled, they are not really the observed outcomes of random samples drawn from the two classes.

The SVM with RFE was applied to the data set of van de Vijver, with half of the genes being eliminated at each stage of the RFE procedure. The external ten-fold cross-validated error rate  $A^{(CV10E)}$  was calculated as the number of genes is reduced from 70 to 64, and then successively halved; see Table 1. It can be seen that  $A^{(CV10E)}$  starts to increase as the number of genes is reduced, but only slightly down to 32 genes. So we decided to adopt as our prediction rule the SVM based on 32 genes.

To provide some guide to the relative importance of the genes in the formation of this rule, we noted how many times a gene was selected in the ten subset of size 32 selected for each of the ten splits of the training data. The frequencies are reported in Table 2 for the 44 genes that were selected for at least one of the ten splits.

In Table 2, the gene number refers to its numbering in the data set as supplied at the supplementary website (<http://www.rii.com/publications/2002/>

Table 1  
 Results of SVM with RFE Applied to 295 Breast Cancer Tissue Samples on 70 Genes in the van de Vijver Data

Number of Genes	Overall Error Rate
1	0.298
2	0.203
4	0.203
8	0.138
16	0.118
32	0.105
64	0.108
70	0.105

nejm.htm) in van de Vijver et al. (2002). This gene number is the rank of the gene when the 70 marker genes are ordered on the basis of the magnitude of their correlation with the class label for the good-prognosis class (that is, the  $F$ -ratio), using the 78 tissue samples in van 't Veer et al. (2002).

For the highly selected genes, we searched for functional annotation in order to link possible biological mechanisms with their apparent role in early tumor metastasis. Out of the 24 genes selected 10 times in Table 2, we found 12 with gene products of known function. These included cell nucleus proteins: CENPA, PRC1, RAMP, NUSAP1, and MELK (many of which seem to be involved in chromatin structure), and also the protein, ORC6L, essential for the initiation of replication, suggesting a role for these genes in tumor cell division. In addition, we identified cell signaling proteins; including IGFBP5 (which appears twice), a known potent inhibitor of growth of breast cancer cells in vitro

Table 2  
 Selection Frequencies of Genes in External Ten-Fold Cross-Validation of SVM with RFE Applied to 295 Breast Cancer Tissue Samples on 70 Genes in the van de Vijver Data

Gene No.	Frequency of Selection on Ten-Fold Validation
66 65 64 60 57 55 52 51 49 46 42 40 39 38 33 22 21 15 9 8 7 4 3 2	10
63 59 23	9
68 13	8
1	6
61 25 20 10	4
43	3
54 27 14	2
53 50 48 44 34 12	1

and in vivo, as well as others involved in cell cycle control and tumorigenesis (FGF18 and TGFB3), and also CEGP1, a secreted protein expressed in vascular endothelium, suggesting a possible role for this gene in angiogenesis as part of tumor metastasis. Finally, we identify the protein AP2B1, involved in the cellular processes of endocytosis and Golgi assembly as it forms part of the clathrin coat assembly.

## 10 Selection Bias Working with the Top 70 Genes

We return now to the breast cancer data of van 't Veer et al. (2002), where the full data set for the 24,881 gene expressions measured on the 78 sporadic breast tumors was available for download. The top 70 marker genes were chosen on the basis of the gene expressions in the 78 tissues. We show here the bias that can be incurred when applying a rule to the top genes as provided by some other method, if one does not actually treat them as the top genes during the cross-validation. We initially applied a filtering step to the full gene set, following the same protocol as in van 't Veer, and retained 5,422 genes for the 78 tumors. We applied an SVM with RFE to the 78 tissue samples on the top 70 genes.

At each stage of the feature elimination process with the SVM, we estimated the overall error rate using ten-fold cross-validation. We performed the latter, using both internal and external cross-validation. For internal cross-validation, the top 70 genes were fixed during the validation process, and so it ignores the selection bias in working with the top 70 genes from the set of 5,422 genes.

In the external cross-validation, this bias is corrected for by going back to the full set of 5,422 genes and selecting the top 70 genes on the training subset at each stage of cross-validation. Then the SVM with RFE is applied to this selected set of top 70 genes, which may have little in common with the original set of top 70 genes.

The results for the internal ten-fold cross-validated overall error rate  $A^{(CV10)}$  and the corresponding external rate  $A^{(CV10E)}$  are listed in Table 3 and plotted in Figure ???. It can be seen from Table 3 and Figure ??? that the selection bias in ignoring the fact that the SVM is being applied to the top 70 genes from a total of 5,422 is approximately 12%. We have also listed in Table 3 the external cross-validated error for the SVM with RFE, starting with the full set of 5,422 genes. It can be seen that it is similar to that of the external cross-validated error rate of the rule starting with the top 70 genes.

Table 3  
Number of Genes and Error Rates with and without Corrections for Selection Bias

Number of Genes	Error Rate	Error Rate	Error Rate
	for Top 70 Genes	for Top 70 Genes	for 5,422 Genes
	(without Correction	(with Correction	(with Correction
	for Selection Bias	for Selection Bias	for Selection Bias)
	as Top 70)	as Top 70)	
1	0.50	0.53	0.56
2	0.32	0.41	0.44
4	0.26	0.40	0.41
8	0.27	0.32	0.43
16	0.28	0.31	0.35
32	0.22	0.35	0.34
64	0.20	0.34	0.35
70	0.19	0.33	—
128	—	—	0.39
256	—	—	0.33
512	—	—	0.34
1,024	—	—	0.33
2,048	—	—	0.37
4,096	—	—	0.40
5,422	—	—	0.44

## 11 Discussion

From the examples presented in Sections 7 and 8, it can be seen that it is important to recognize that a correction for the selection bias be made in estimating the prediction error of a rule formed using genes selected from a very large set of available genes. It is also important to note that if a test set is used to estimate the prediction error, then there will be a selection bias if this test set was used also in the gene-selection process. Thus the test set must play no role in the feature selection process for an unbiased estimate to be obtained.

Given that there are usually only a limited number of tissue samples available for the training of the prediction rule, it is not practical for a subset of tissue samples to be so put aside for testing purposes. However, we can correct for the selection bias either by cross-validation or by the bootstrap, as implemented above in the examples. Concerning the former approach, an internal cross-validation does not suffice. That is, an external cross-validation must be performed whereby at each stage of the validation process with the deletion of a subset of the observations for testing, the rule must be trained on the retained subset of observations by performing the same feature selection procedure used to train the rule in the first instance on the full training set.

The example in Section 10 serves to make the point that care must be exercised in comparing the error rates of two discriminant rules formed from the same tissue samples of different sets of genes. For example, one rule  $r_1$  may be formed from a training set of  $n = M$  tissue samples of  $p = N$  genes, while another rule  $r_2$  might be formed using a subset of these  $N$  genes, say, the top 100 genes. If a fair comparison is to be made between the error rates of these two rules, then the error rate of the second rule  $r_2$  should not be estimated by just working with the top 100 genes during the cross-validation. Rather, one should start initially with the full set of  $N$  genes and select the top 100 genes on each stage of the training of  $r_2$  in the cross-validation trials.

## References

- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* 96, 6745-6750.
- Ambroise, C. and McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences* 99, 6562-6566.
- Braga-Neto, U.M. and Dougherty, E., 2004. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20, 374-480.
- Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., and Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* 97, 262-267.
- Dettling, M. and Bühlmann, P., 2002. Supervised clustering of genes. *Genome Biology* 3, research0069.1-0069.15.
- Díaz-Uriarte, R., 2003. A simple method for finding molecular signatures from gene expression. *Bioinformatics Unit Technical Report 004*, Madrid, Spain: Spanish National Cancer Center (CNIO).
- Efron, B. and Tibshirani, R., 1997. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* 92, 548-560.
- Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906-914.
- Ghosh, D. and Chinnaiyan, A.M., 2002. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* 18, 275-286.
- Goh, L., Kasabov, N., and Song, Q., 2004. A novel feature selection method to improve classification of gene expression data. In *Proceedings of the Second Asia-Pacific Bioinformatics Conference (APBC2004)*, Dunedin, New

- Zealand, CRPIT, 29, Y.-P. P. Chen (Ed.). Australian Computer Society, pp. 161–166.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeck, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46, 389–422.
- Hand, D.J., Mannila, H., and Smyth, P., 2001. *Principles of Data Mining*. Cambridge, MA: MIT Press.
- Hastie, T., Tibshirani, R., and Friedman, J., 2001. *The Elements of Statistical Learning*. New York: Springer-Verlag
- Hastie, T., Tibshirani, R., Botstein, D., and Brown, P., 2001a. Supervised harvesting of expression trees. *Genome Biology* 2, research0003.1–0003.12.
- Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M., 2003. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology* 2(1), Article 3.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.
- Liu, A., Zhang, Y., Gehan, E., and Clarke, R., 2002. Block principal component analysis with application to gene microarray data application. *Statistics in Medicine* 21, 3465–3474.
- Liu, J.S., Zhang, J.L., Palumbo, M.J., and Lawrence, C.E., 2003. Bayesian clustering with variable and transformation selections. *Bayesian Statistics*, Vol. 7, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West (Eds.). Oxford: Oxford University Press, pp. 249–275.
- McLachlan, G.J., 1992. *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- McLachlan, G.J., Bean, R.W., and Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18, 413–422.
- McLachlan, G.J., Do, K.-A., Ambrose, C., 2004. *Analyzing Microarray Gene Expression Data*. Wiley.
- Medvedovic, M. and Sivaganesan, S., 2002. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* 18, 1194–1206.
- Nguyen, D.V. and Rocke, D.M., 2001. Classification of acute leukemia based on DNA microarray gene expressions using partial least squares. In *Methods of Microarray Data Analysis*, S.M. Lin and K.F. Johnson (Eds.). Dordrecht, The Netherlands: Kluwer, pp. 109–124.

- Nguyen, D.V. and Rocke, D.M., 2002a. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18, 39–50.
- Nguyen, D.V. and Rocke, D.M., 2002b. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18, 1216–1226.
- Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L.(Eds.), 2003. *The Analysis of Gene Expression Data*. New York: Springer-Verlag.
- Ripley, B.D., 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
- Rocke, D.M. and Durbin, B., 2003. Approximate variance-stabilizing transformations for a gene-expression microarray data. *Bioinformatics* 19, 966–972.
- Speed, T. (Ed.), 2003. *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton, FL: Chapman & Hall/CRC.
- Tibshirani, R.J., Hastie, T., Narasimhan, B., and Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99, 6567–6572.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science* 18, 104–117.
- van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., and Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.
- van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., and Bernards, R., 2002. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 347, 1999–2009.
- Vapnik, V., 1998. *Statistical Learning Theory*. New York: Wiley.
- West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Sprang, R., Zuzan, H., Olson, J.A., Marks, J.R., and Nevins, J.R., 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* 98, 11462–11467.
- Xiong, M., Li, W., Zhao, J., Jin, L., and Boerwinkle, E., 2001. Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism* 73, 239–247.
- Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L., 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987.
- Yeung, K.Y., Medvedovic, M., and Bumgarner, R.E., 2003. Clustering gene-expression data with repeated measurements. *Genome Biology* 4 No. 5, Article R34.