Penalized Discriminant Methods for the Classification of Tumors from Gene Expression Data

Debashis Ghosh

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor,
Michigan 48105, U.S.A.
email: ghoshd@umich.edu

SUMMARY.   Due to the advent of high-throughput microarray technology, it has become possible to develop molecular classification systems for various types of cancer. In this article, we propose a methodology using regularized regression models for the classification of tumors in microarray experiments. The performances of principal components, partial least squares, and ridge regression models are studied; these regression procedures are adapted to the classification setting using the optimal scoring algorithm. We also develop a procedure for ranking genes based on the fitted regression models. The proposed methodologies are applied to two microarray studies in cancer.

KEY WORDS: Cross-validation; Microarrays; Partial least squares; Principal components; Regularization; Ridge regression.

1. Introduction

With the development of large-scale, high-throughput gene expression technology, it has become possible to diagnose and classify disease, particularly cancer, based on these assays (Alizadeh et al., 2001). This has been termed "class prediction" in the microarray literature (Golub et al., 1999).

An example of a microarray experiment in cancer is given by Khan et al. (2001). The goal of this study was to develop a method of classifying childhood cancers to certain diagnostic groupings utilizing the gene expression profiles. For the experiment, 63 training samples, representing various types of small, round blue cell tumors (SRBCTs), were collected; the gene expression profile was analyzed using cDNA microarrays. The authors then used artificial neural networks (ANN) for training and generating a classification model for classification of cancer based on the gene expression profiles. The authors then applied their ANN to a collection of 25 test samples and found that the neural network model correctly classified all 25 of the test cases (although five cases represented non-SRBCTs).

In addition to the example, there have been several investigations utilizing supervised learning methods for the classification of tumors based on microarray data. Golub et al. (1999) utilized a nearest-neighbor classifier method for the classification of acute myeloid lymphoma (AML) and acute leukemia lymphoma (ALL) in children. Dudoit, Fridlyand, and Speed (2002) performed a systematic comparison of several discrimination methods for classification of tumors based on microarray experiments. While they found linear discriminant analysis to perform the best, in order to utilize the method, the number of genes selected had to be drastically reduced from thousands to tens using a univariate filtering criterion. A more recent technique that is popular in computer science, namely, support vector machines, has also been applied to the classification of tumors using microarray data (Yeang et al., 2001). There has also been some work on utilizing latent-factor models for classification (Li and Hong, 2001; West et al., 2001).

One feature of microarray studies is the fact that the number of tumor samples collected tends to be much smaller than the number of genes per chip. The former number tends to be on the order of tens or hundreds, while microarrays typically contain thousands of genes on each chip. In statistical terms, the number of predictor variables is much larger than the number of independent samples. If the scientific question is to see whether or not gene expression profiles can predict tumor type, then from a regression point of view, it makes sense to think of the gene expression profile as the covariates. For these types of problems, it should be obvious that some type of regularization or variable reduction is needed. In most of the previous work described, the authors have used univariate methods for reducing the number of genes under consideration before applying the classification methods. An alternative approach was taken in Khan et al. (2001), where the authors applied principal components analysis to the gene expression data before training the ANN models.

Another field in which the "large p, small n" (West, 2003, to appear) problem exists is chemometrics. Frank and Friedman (1993) proposed using regularized regression models for analyzing chemometric data. However, in these settings, the response of interest is continuous, while for classification problems, the label is a categorical variable.

In this article, we present a methodology that extends the regularized regression models of chemometrics to classification

problems in gene expression studies. This is done using the optimal scoring algorithm of Hastie, Tibshirani, and Buja (1994). This approach seems attractive for a variety of reasons. First, regularized regression models can handle large numbers of correlated predictor variables. Second, we can develop predictive models for classifying tumors, based on the entire gene expression profile, without filtering out any of the genes. Third, this algorithm is quite computationally efficient. Finally, based on the regression output, we can rank the genes for potential follow-up experiments; this constitutes a form of "learning" about the individual genes on the microarray. The format of the article is as follows. In Section 2, we review the methods for regularized regression modelling and optimal scoring. The algorithm for classification of tumors based on gene expression profiles and subsequent ranking of genes is presented in Section 3. We then apply our technique to data from two cancer studies in Section 4. Some concluding remarks are made in Section 5.

2. Methods

2.1 *Notation and Preliminaries*

Let $\mathbf{a}^T$ denote the transpose of the vector $\mathbf{a}$. For the ith sample $(i = 1, \ldots, n)$, we let $\mathbf{X}_i = [X_{i1} \cdots X_{ip}]$ denote the $p \times 1$ gene expression profile vector (i.e., X_{ij} is the gene expression measurement of the jth gene, $j = 1, \ldots, p$). We suppose that the data have already been preprocessed and normalized. In addition, it is assumed that the gene expression data are standardized so that for each chip, the expression profile has mean zero and standard deviation one. Let g_i denote the tumor class for the ith sample $(i = 1, \ldots, n)$; we assume that there are G tumor classes so that g_i takes values $\{1, \ldots, G\}$. In Section 2.2, we assume the existence of a continuous response variable Y_i for the ith sample $(i = 1, \ldots, n)$.

2.2 *Penalized Regression Models*

We focus here on three types of regularized regression models: ridge regression, principal components regression and partial least squares regression. We now briefly outline the model and estimation algorithm associated with each of these procedures.

2.2.1 *Ridge regression.* Suppose we wish to fit the following regression model:

$$E(Y_i \mid \mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}_0, \qquad (2.1)$$

where $\boldsymbol{\beta}_0$ is a $p \times 1$ vector of unknown regression parameters. Because n is smaller than p, the usual ordinary least squares (OLS) estimator will not be well-defined. An alternative is to use the ridge regression estimator of $\boldsymbol{\beta}_0$ in (2.1):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y},$$

where $\mathbf{X}$ is the $n \times p$ matrix whose ith row is $\mathbf{X}_i (i = 1, \ldots, n)$, $\mathbf{I}_p$ is the $p \times p$ identity matrix, λ is a constant, and $\mathbf{Y} \equiv [Y_1 \cdots Y_n]^T$ is a $n \times 1$ vector. The ridge regression approach can also be motivated from a Bayesian viewpoint (Lindley and Smith, 1972).

The parameter λ controls the amount of shrinkage in the data. Setting $\lambda = 0$ yields the ordinary least squares solution, while setting $\lambda > 0$ increases the bias in the estimate of $\boldsymbol{\beta}_0$, but decreases the variance of the parameter estimators. Setting $\lambda = \infty$ yields the parameter estimate $\widehat{\boldsymbol{\beta}} = \mathbf{0}$. One issue involves the choice of λ. A very common approach is to use cross-validation (Stone, 1974).

2.2.2 *Principal components regression.* The method of principal components regression can be traced back to Massy (1965). In this method, we first perform a singular value decomposition of the $p \times n$ matrix $\mathbf{X}^T$:

$$\mathbf{X}^T = \mathbf{UDV},$$

where $\mathbf{U}$ is $p \times n$ matrix, and $\mathbf{D}$ and $\mathbf{V}$ are $n \times n$ matrices. The columns of $\mathbf{U}$ are orthonormal, i.e., $\mathbf{U}^T \mathbf{U} = \mathbf{I}_p$, the $p \times p$ identity matrix. The diagonal matrix $\mathbf{D}$ contains the ordered eigenvalues of $\mathbf{X}$ on the diagonal elements, so that $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$, where $d_1 \geq d_2 \geq d_3 \geq \cdots \geq d_n \geq 0$. We will assume, without loss of generality, that $d_i > 0$ for $i = 1, \ldots, n$. Finally, $\mathbf{V}$ is the $n \times n$ singular value decomposition factor matrix and has both orthonormal rows and columns. Plugging this decomposition into (2.1), we have

$$E(Y_i \mid \mathbf{W}_i) = \mathbf{W}_i^T \boldsymbol{\gamma}_0, \qquad (2.2)$$

where $\mathbf{W}_i$ is the ith row of $\mathbf{W} \equiv \mathbf{DV}$ and $\boldsymbol{\gamma}_0 = \mathbf{U}^T \boldsymbol{\beta}_0$. We can fit the model in (2.2) using ordinary least squares. We can get an estimate of $\boldsymbol{\beta}_0$ by multiplying $\mathbf{U}^T$ with the least squares estimator of $\boldsymbol{\gamma}_0$ in (2.2).

A major issue in principal components regression modelling is determining the number of components (i.e., number of columns of $\mathbf{W}$) to use. There are several possible model selection criteria; the criteria in this article will focus on using methods based on cross-validation (Stone, 1974).

2.2.3 *Partial least squares regression.* A popular regression method in chemometrics is partial least squares (Wold, 1975; Naes and Martens, 1985; Helland, 1988). The partial least squares algorithms attempt to simultaneously find linear combinations of the predictors whose correlation is maximized with the response and which are uncorrelated over the sample. There are several algorithms available for numerical estimation using partial least squares; a nice overview of these methods can be found in Denham (1994).

The model that underlies partial least squares is that $\mathbf{X}_i$ and $Y_i (i = 1, \ldots, n)$ are both linear functions of a set of common latent factors, say, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$, where k is assumed to be less than or equal to the rank of the matrix $\mathbf{X}$. Most partial least squares algorithms involve estimation of both the latent factors and their associated effects on $Y_i (i = 1, \ldots, n)$. For example, here is the pseudocode for one algorithm discussed by Helland (1988):

1. Compute $\mathbf{a}_1 = \mathbf{X}^T y$;
2. Compute $\mathbf{t}_1 = \mathbf{X} \mathbf{a}_1$;
3. Compute $\mathbf{r} = \mathbf{y} - \mathbf{t}_1 (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T \mathbf{y}$;
4. for iteration $l = 2, \ldots, m$, compute

$$\mathbf{a}_l = \mathbf{X}^T \mathbf{r};$$
$$\mathbf{t}_l = \mathbf{X} \mathbf{a}_l;$$
$$\mathbf{r} = \mathbf{y} - \mathbf{T}_l (\mathbf{T}_l^T \mathbf{T}_l)^{-1} \mathbf{T}_l^T \mathbf{y};$$

5. compute $\mathbf{q} = (\mathbf{T}_m^T \mathbf{T}_m)^{-1} \mathbf{T}_m^T \mathbf{y}$;
6. compute $\widehat{\boldsymbol{\beta}} = \mathbf{A}_m \mathbf{q}$,

where $\mathbf{T}_l = [\mathbf{t}_1 \cdots \mathbf{t}_l]$ and $\mathbf{A}_m = [\mathbf{a}_1 \cdots \mathbf{a}_m]$.

While we have also implemented the proposed methodology using the two other algorithms from Martens and Naes (1989), in this article, we focus on using Helland's method for estimation in partial least squares regression models. We can derive an estimator for $\boldsymbol{\beta}_0$ in (2.1) using the partial least squares estimators using a back-transformation similar to that described in 2.2.2.

In contrast to ridge and principal components regression estimation procedures, partial least squares algorithms are nonlinear in the response values. Frank and Friedman (1993) mention that, in practice, partial least squares regression models tend to have good predictive properties.

In practice, the number of latent factors to include in the model must be chosen. The method most often used for choosing this quantity is cross-validation (Stone, 1974).

2.3 *Optimal Scoring*

In the previous section, we have described various penalized regression models that have been used successfully in chemometric applications. However, one aspect that separates those situations from the current one is the fact that the outcome of interest is categorical, and our goal is classification. Thus, one way of applying the regularized regression models of Section 2.2 to classification problems is to re-express the classification problem as a regression problem. This is done using the optimal scoring algorithm, which we describe here; readers who are interested in more details are referred to Hastie et al. (1994) and Hastie, Buja, and Tibshirani (1995). The idea is based on a regression formulation for linear discriminant analysis. We first convert $\mathbf{g} = [g_1 \cdots g_n]^T$ into an $n \times G$ matrix $\mathbf{Z} = [Z_{ij}]$, where $Z_{ij} = 1$ if the class of the ith sample is $g_i = j$, and 0 otherwise.

The point of optimal scoring is to turn the categorical class labels into quantitative variables. Let $\boldsymbol{\theta}_k(g) = [\theta_k(g_1), \ldots, \theta_k(g_n)]^T$ be the $n \times 1$ vector of quantitative scores assigned to $\mathbf{g}$ for the kth class. The optimal scoring problem involves finding the coefficients $\boldsymbol{\beta}_k (k = 1, \ldots, G)$ and the scoring maps θ_k that minimize the following average squared residual:

$$\text{ASR} = n^{-1} \sum_{k=1}^{G} \sum_{i=1}^{n} \left\{ \theta_k(g_i) - \mathbf{X}_i^T \boldsymbol{\beta}_k \right\}^2. \qquad (2.3)$$

Let $\boldsymbol{\Theta}$ be a $G \times J$ matrix, where $J \leq G - 1$, whose kth row are the scores $\theta_k(\cdot)$ for the kth class, $k = 1, \ldots, G$. The parameter J can be chosen by the user; we take $J = G - 1$, to maintain the most flexibility in discrimination boundaries between the three groups. These scores are assumed to be mutually orthogonal and normalized with respect to an inner product; this leads to the constraint on J. Thus, the minimization of (2.3) is subject to the constraint $N^{-1} \| \mathbf{Z} \boldsymbol{\Theta} \|^2 = 1$. As mentioned by Hastie et al. (1994), the minimization of this constrained optimization problem leads to estimates of $\boldsymbol{\beta}_k$ that are proportional to the discriminant variables in linear discriminant analysis (LDA). Hastie et al. (1994) suggest replacing the linear predictor in (2.3) with a more general $f(\mathbf{X}_i)$; they propose the following algorithm for flexible discriminant analysis:

1. Choose an initial score matrix $\boldsymbol{\Theta}_0$ satisfying $\boldsymbol{\Theta}_0^T \mathbf{D}_p \boldsymbol{\Theta}_0 = \mathbf{I}$, where $\mathbf{D}_p = \mathbf{Z}^T \mathbf{Z}/n$. Let $\boldsymbol{\Theta}_0^* = \mathbf{Z} \boldsymbol{\Theta}_0$.

2. Fit a multivariate nonparametric regression model of $\boldsymbol{\Theta}_0^*$ on $\mathbf{X}$, yielding fitted values $\widehat{\boldsymbol{\Theta}}$. Let $\widehat{\mathbf{f}}(\mathbf{X})$ be the vector of fitted regression functions.
3. Obtain the eigenvector matrix $\boldsymbol{\Phi}$ of $\boldsymbol{\Theta}_0^{*T} \widehat{\boldsymbol{\Theta}}$; the optimal scores are then $\boldsymbol{\Theta}* = \boldsymbol{\Theta}_0 \boldsymbol{\Phi}$.
4. Define $\mathbf{f}_{\text{opt}}(\mathbf{x}) = \boldsymbol{\Phi}^T \widehat{\mathbf{f}}(\mathbf{x})$.

In their discussion of the flexible discriminant analysis algorithm, Hastie et al. (1994) focus on using MARS (Friedman, 1991) and BRUTO as possible nonparametric regression algorithms in step 3. They then show through a variety of examples that these procedures tend to have good predictive performance.

3. Algorithms and Comparisons

So far, we have outlined the components necessary for implementation of our procedure. In this section, we develop our algorithm for classification of tumor samples using microarray data. We also discuss associated issues, such as determining the optimal amount of regularization and ranking genes based on the fitted models. In addition, we discuss some analytical comparisons between the three regularized classification approaches as well as comparisons with logistic regression modelling.

3.1 *Penalized Optimal Scoring for Classification*

We propose to use a penalized optimal scoring procedure for classification using regularized regression models. Here is the outline for our method:

1. Choose an initial score matrix $\boldsymbol{\Theta}$ satisfying $\boldsymbol{\Theta}^T \mathbf{D}_p \boldsymbol{\Theta} = \mathbf{I}$, and let $\boldsymbol{\Theta}_0 = \mathbf{Z} \boldsymbol{\Theta}$.
2. Fit a multivariate penalized regression model of $\boldsymbol{\Theta}_0$ on $\mathbf{X}$, yielding fitted values $\boldsymbol{\Theta}_0^*$. Let $\widehat{\mathbf{f}}(\mathbf{X})$ be the vector of fitted regression functions.
3. Obtain the eigenvector matrix $\boldsymbol{\Phi}$ of $\boldsymbol{\Theta}*_0^T \boldsymbol{\Theta}_0$; the optimal scores are $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \boldsymbol{\Phi}$.
4. Define $\mathbf{f}_{\text{opt}}(\mathbf{x}) = \boldsymbol{\Phi}^T \widehat{\mathbf{f}}(\mathbf{x})$.

Our algorithm is similar to that proposed by Hastie et al. (1994), except that we replace a multivariate nonparametric regression model in step 3 with a multivariate penalized regression model. This is equivalent to fitting univariate penalized regression models to each of the columns in $\boldsymbol{\Theta}_0$. One point to note is that when fitting these separate penalized regressions, we use the same amount of shrinkage for each column of $\boldsymbol{\Theta}_0$. We will be focusing on the penalized regression models discussed in Section 2.2.

For a simple situation, let us consider the of $G = 3$ classes. Then $\boldsymbol{\Theta}$ is a 3×2 matrix, and the algorithm proceeds in the following manner. We first find a 3×2 matrix $\boldsymbol{\Theta}$ such that $\boldsymbol{\Theta}^T \mathbf{D}_3 \boldsymbol{\Theta} = \mathbf{I}_2$. Then $\boldsymbol{\Theta}_0 = \mathbf{Z} \boldsymbol{\Theta}$ is a $n \times 2$ matrix of initial optimal scores. We next fit two penalized regression models with the first and second columns of $\boldsymbol{\Theta}_0$ as the responses and $\mathbf{X}$ as the predictor. Based on the estimates, we then have a $n \times 2$ matrix of fitted values $\boldsymbol{\Theta}_0^*$ and an $p \times 2$ matrix of regression coefficients $\hat{\boldsymbol{\beta}}$. At the next step, we eigenanalyze the 2×2 matrix $(\boldsymbol{\Theta}_0^*)^T \boldsymbol{\Theta}_0$ to obtain the eigenvector matrix $\boldsymbol{\Phi}$. Then the 3×2 matrix of optimal scores are given by $\boldsymbol{\Theta}_0 \boldsymbol{\Phi}$, and the regression coefficients are updated as the $p \times 2$ matrix $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} \boldsymbol{\Phi}$. The fitted values are then $\boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta}^*$ with ith row $\boldsymbol{\eta}_i (i = 1, \ldots, n)$.

Once the algorithm has been run, we now have a discriminant rule for classifying future observations. The form of the rule is that of a nearest centroid rule; in particular, we assign a gene expression profile $\mathbf{X}_{\text{new}}$ to the class j that minimizes

$$\delta(\mathbf{X}, j) = \|\mathbf{D}(\mathbf{X}\boldsymbol{\beta}^* - \bar{\boldsymbol{\eta}}^j)\|^2,$$

where

$$\bar{\boldsymbol{\eta}}^j = \frac{\sum_{i:g_i=j} \boldsymbol{\eta}_i}{\sum_{i:g_i=j} I(g_i = j)}.$$

$\mathbf{D}$ is a matrix with diagonal element

$$D_{kk} = \left\{ \frac{1}{\left[\lambda_k^2 \left(1 - \lambda_k^2\right)\right]} \right\}^{1/2},$$

with $\lambda_k (k = 1, \ldots, G - 1)$ being the kth largest eigenvalue calculated in step 3 of the algorithm.

To better understand the algorithm, it is useful to see the criterion function, analogous to (2.3), that we are minimizing. For the ridge regression–based optimal scoring method, we are optimizing the following objective function:

$$\text{ASR}^{\text{rr}} = n^{-1} \sum_{k=1}^{G} \sum_{i=1}^{n} \left\{ \theta_k(g_i) - \mathbf{X}_i^T \boldsymbol{\beta}_k \right\}^2 + \lambda \sum_{k=1}^{G} \boldsymbol{\beta}_k^T \boldsymbol{\beta}_k,$$

subject to $N^{-1}\|\mathbf{Z}\boldsymbol{\Theta}\|^2 = 1$. A variant of the penalized optimal scoring algorithm using ridge regression was studied by Hastie et al. (1995). For the principal components regression-based optimal scoring method, it involves minimizing the following objective function:

$$\text{ASR}^{\text{pcr}} = n^{-1} \sum_{k=1}^{G} \sum_{i=1}^{n} \left\{ \theta_k(g_i) - \mathbf{W}_i^T \boldsymbol{\gamma}_k \right\}^2,$$

where $N^{-1}\|\mathbf{Z}\boldsymbol{\Theta}\| = 1$ and $\boldsymbol{\gamma}_k = \mathbf{U}^T \boldsymbol{\beta}_k$.

Finally, it is algebraically more complicated to write the objective function minimization problem for the partial least squares, but conceptually, the model involves formulating both the class scores and the covariates as linear functions of underlying latent variables. In terms of the algorithmic details, however, the partial least squares algorithm is implemented in the same way as the ridge and principal components regression classification procedures.

If the number of samples in each group is the same, then using the algorithm described above with principal components regression is equivalent to employing principal components analysis followed linear discriminant analysis on the individual components. However, the approach proposed here is more general and can allow for other types of penalized regression models. No clear relationship exists if the number of samples in the various groups is different.

In terms of assessing performance between the methods (ridge regression, principal components regression, and partial least squares regression) from an analytical point, suppose that $\boldsymbol{\Theta}$ is known and there are only $K = 2$ classes. Thus, $J = 1$ and $\boldsymbol{\Theta}$ maps to two points. We can then consider the

following class of objective functions, similar to that considered by Stone and Brooks (1990):

$$\text{Var}(\mathbf{X}^T\boldsymbol{\beta})^2 \text{Cov}(\boldsymbol{\Theta}(g), \mathbf{X}^T\boldsymbol{\beta})^{\alpha/(1-\alpha)-1}, \qquad (3.1)$$

where Var and Cov and shorthand notation for variance and covariance, and α is a number between 0 and 1. In this framework, values of $\alpha = 1/2$ and $\alpha = 1$ correspond to the objective functions maximized by partial least squares (PLS) and principal components regression (PCR), respectively. The value $\alpha = 0$ corresponds to the problem solved by ordinary least squares; because of the "large p, small n" problem, we use ridge regression (RR) instead. In the framework presented here, we can think of PCR, PLS, and RR as corresponding to canonical variance, canonical covariance, and canonical correlation analyses, respectively. Since the classification procedure in the latter two methods involves the correlation between predictor (gene expression profile) and response (tissue/tumor type), one might expect that they would give better classifiers than PCR.

3.2 *Choosing the Optimal Amount of Regularization*

Ridge regression, principal components regression and partial least squares regression models all involve a regularization parameter that must be selected somehow. In ridge regression, the regularization parameter is λ, while for principal components and partial least squares regression, the parameter that needs to be chosen is the number of components to include in the model. This issue is similar to choosing the bandwidth for a kernel in nonparametric regression estimation.

In general, the strategy for choosing the amount of regularization depends on a combination of the size of the study and the amount of computation time required. In the examples presented in Section 4, we use leave-one-out cross validation in the test datasets to determine the optimal shrinkage parameter or the number of components to include in the model.

One other point to note is that the cross-validation procedure for the principal components regression approach leads to selection of those components corresponding to the largest eigenvalues. The discussion at the end of the previous section suggests that the PCR-based classifier will work well if variation in the gene expression explains the difference in class groupings. The principal components in this setting have also been called eigengenes; a biological motivation for these quantities has been given by Alter, Brown, and Botstein (2000).

3.3 *Comparison with Logistic Regression Approaches*

An alternative approach to multigroup classification problems is logistic regression techniques. Methods have been recently developed for logistic regression classification of microarray data (Eilers et al., 2001). In addition, principal components regression (Marx and Smith, 1990) and partial least squares estimation (Marx, 1996) approaches exist for binary data. There is an issue involved in combining the classifiers with multigroup data.

There were a few reasons we chose to pursue the discriminant analysis approach. The first is the superiority of the method in the study of Dudoit et al. (2002). However, while it seemed plausible that differences between tumor types could be characterized in terms of tens of genes, there are

other tumor-differentiation problems that would require more information than is available in tens of genes. This type of discrimination might require on the order of hundreds or thousands of genes. For this technique, regularization of the linear discriminant approach is needed.

In addition, researchers have demonstrated the increase in efficiency of linear discriminant analysis to logistic regression in two-group and multigroup classification problems (Efron, 1975; Bull and Donner, 1987). For example, let us consider the use of PCR for classification. Because we reduce the dimensionality of the predictor space by principal components, the standard results apply, which suggest that use of linear discriminant analysis leads to gains in classification efficiency relative to logistic regression.

3.4 *Gene Selection*

One important scientific goal in microarray studies is to determine what genes are potential candidates for follow-up studies, using validation methods such as quantitative polymerase chain reaction or immunohistochemical techniques. Based on the classification approach we have described in Section 3.1, conditional on the optimal amount of regularization, we can determine a ranking of the most "interesting" genes based on the estimated regression coefficients from step 3 of our algorithm. A similar approach was taken by West et al. (2001).

One point to note about ranking the coefficients is that we will have $G - 1$ sets of estimates of the regression parameters. Thus, the choice of ranking the regression coefficients will depend on which pairwise tumor comparison we are interested in.

It should also be noted that the gene selection process is variable. To assess the stability of the observed rankings, the bootstrap is used. We first fix the regularization parameter and select a prespecified number k. We then sample the chips and rerun the regression procedures, ranking the top k genes. For each of the top genes in the original dataset, we determine the percentage of times they appear in the top k gene list for the bootstrapped datasets. This gives a measure as to the reliability of the selected genes.

3.5 *Software Implementation*

We are in the process of currently developing some software implementing the proposed classification methodology. The suite of functions has been written in R, a freely downloadable statistical software (`http://www.r-project.org/`). The software will be available at the following website: `http://www. sph.umich.edu/~ghoshd/COMPBIO/POPTSCORE/`.

The algorithms are fairly fast, primarily due to the use of the singular value decomposition in fitting the various types of regularized regression models. This technique avoids the need to have to invert large matrices for the gene expression profiles. The interested reader is referred to Friedman, Hastie, and Tibshirani (2001, Section 3.4.3), Massy (1965), and Denham (1994) for the technical details on applying singular value decomposition to ridge regression, principal components, and partial least squares regression models.

The partial least squares classification methodology has been implemented using the Helland (1988) algorithm described in Section 2.2.3, as well as two algorithms from Martens and Naes (1989). However, in the article, we focus on the Helland approach for performing classification using partial least squares.

4. Numerical Examples

We now illustrate the use of the proposed methodology using two case studies from microarray experiments in the cancer setting. The studies described here use the spotted cDNA technology, which involves measuring gene expression levels on red and green channels.

4.1 *Childhood Cancer Data*

We consider the data from the Khan et al. (2001) study. The goal of this study was to develop a system for classifying four categories of small round blue cell tumors: neuroblastomas (NB), rhabdomyosarcomas (RMS), non-Hodgkin lymphomas (NHL), and Ewing sarcomas (EWS). Their microarrays had spots representing 6567 genes; using a filter based on the red channel intensity, this was reduced down to 2308 genes. For these data, there was a training sample of 63 SRBCTs and a test set of 25 tumors. The test set included five tumors that were non-SRBCTs; we excluded these tumors from the analysis. Because we had a training set available, we built the classification models using the training sample of 63 tumors and estimated classification error rates using the test set.

Using the proposed regression methods, we found that we achieved 100% classification accuracy using ridge regression. This error rate was not sensitive to the choice of λ used. Based on the principal components and partial least squares regression procedures, the optimal classification accuracies we could achieve with either method was 100%. For the principal components regression, this involved including ten components in the model. Using the partial least squares approach, a classification accuracy of 100% was obtained with six components in the model. For comparison, we tried three other approaches. First, we attempted to fit a linear discriminant analysis using the entire gene expression profile and no dimension reduction. Because of instability in the estimated linear discriminant function, the estimated classification probabilities did not converge. Second, we selected the top 40 genes based on the discrimination criterion mentioned in Dudoit et al. (2002) and performed a linear discriminant analysis. This yielded a classification accuracy of 85%. The second method we tried was to utilize principal components analysis, followed by a linear discriminant analysis. Using the test set, we achieved an optimal classification rate of 95%, using 11 principal components. Given our discussion in Section 3.1, it is not surprising that this approach gives similar classification accuracies to those from the proposed method. In addition, we analyzed the data using the random forests method (Breiman, 2001). We grew 100 trees, where at each node, two variables were tried for further splitting. This yielded an out-of-bag error estimate of 0%, corresponding to perfect classification.

Based on the statistical models proposed in the article, we can then generate lists of genes that discriminate between the various classes of SRBCTs. In Tables 1–3 , we summarize the top 20 genes in terms of discrimination between EWS and the other three classes of tumors. The list was generated using the ridge regression method with $\lambda = 1$. We then used the bootstrap to determine the number of times these

Table 1
List of ranked genes for discriminating BL from EWS in childhood cancer data

Clone ID	Gene	Confidence score 1[a]	Confidence score 2[b]
296448[c]	insulin-like growth factor 2 (somatomedin A)	0.99	0.98
207274[c]	human DNA for insulin-like growth factor II (IGF-2) exon 7 and additional ORF	1.00	0.99
784224[c]	fibroblast growth factor receptor 4	1.00	0.98
745343[c]	regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	0.91	0.71
868304[c]	actin, alpha 2, smooth muscle, aorta	0.95	0.75
244618[c]	ESTs	0.97	0.55
882522	argininosuccinate synthetase	0.92	0.76
812965	v-myc avian myelocytomatosis viral oncogene homolog	0.99	0.89
840942[c]	major histocompatibility complex, class II, DP beta 1	0.94	0.74
233721[c]	insulin-like growth factor binding protein 2 (36kD)	0.49	0.09
755145	villin 2 (ezrin)	0.92	0.57
839552	nuclear receptor coactivator 1	0.93	0.24
842806	cyclin-dependent kinase 4	0.70	0.10
298062[c]	troponin T2, cardiac	0.73	0.04
45544	transgelin 2	0.85	0.59
878798	beta-2-microglobulin	1.00	0.89
47475	*Homo sapiens* inducible protein mRNA, complete cds	0.75	0.42
236034	uncoupling protein 2 (mitochondrial, proton carrier)	0.69	0.05
782811	high-mobility group (nonhistone chromosomal) protein isoforms I and Y	0.65	0.20
809603	ESTs, weakly similar to cDNA EST EMBL:M89154 comes from this gene [C.elegans]	0.73	0.05

[a] Proportion of 100 jittered datasets in which gene or EST was one of top 100 ranking genes.

[b] Proportion of 100 jittered datasets in which gene or EST was one of top 20 ranking genes.

[c] Gene selected based on selection criteria in Figure 3(b) of Khan et al. (2001).

genes appeared in the top 100 of ranked genes. This yields the confidence score in the third column on each of the tables. If we were to employ a more strict criteria and determine the number of times the listed genes appeared in the top 20 discriminating genes for the bootstrap samples, then the re-sulting confidence scores are given in the fourth column of the tables. The confidence scores were computed using 100 boot-strap samples. Based on these tables, there are two points to note. First, variable selection using this method can be quite sensitive to the number of ranked genes one uses. For

Table 2
List of ranked genes for discriminating NB from EWS in childhood cancer data

Clone ID	Gene	Confidence score 1[a]	Confidence score 2[b]
1435862[c]	antigen identified by monoclonal antibodies 12E7, F21 and O13	0.99	0.96
377461[c]	caveolin 1, caveolae protein, 22kD	0.98	0.93
770394[c]	Fc fragment of IgG, receptor, transporter, alpha	0.98	0.91
814260[c]	follicular lymphoma variant translocation 1	0.98	0.93
208718[c]	annexin A1	0.97	0.75
302933	nucleolin	0.49	0.11
34357	actin, beta	0.87	0.31
811000[c]	lectin, galactoside-binding, soluble, 33 binding protein (galectin 6 binding protein)	0.90	0.15
866702	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)	0.95	0.79
52076[c]	olfactomedinrelated ER localized protein	0.81	0.43
781097	neurotrophic tyrosine kinase, receptor-related 1	0.90	0.24
491565	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	0.94	0.66
878833	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)	0.87	0.13
627939	cysteine and glycine-rich protein 3 (cardiac LIM protein)	0.88	0.27
241412[c]	E74-like factor 1 (ets domain transcription factor)	0.82	0.35
47475	*Homo sapiens* inducible protein mRNA, complete cds	0.76	0.44
868304[c]	actin, alpha 2, smooth muscle, aorta	0.76	0.63
379708	no label	0.95	0.19
842784	phosphate carrier, mitochondrial	0.45	0.05
50117	glyceraldehyde-3-phosphate dehydrogenase	0.42	0.02

[a,b,c] See Table 1.

Table 3
List of ranked genes for discriminating RMS from EWS in childhood cancer data

Clone ID	Gene	Confidence score 1[a]	Confidence score 2[b]
629896[c]	microtubule-associated protein 1B	1.00	1.00
812105[c]	transmembrane protein	1.00	1.00
810057	cold shock domain protein A	1.00	0.99
878652	postmeiotic segregation increased 2-like 12	1.00	1.00
784224[c]	fibroblast growth factor receptor 4	1.00	0.98
755239	methyltransferase-like 1	0.98	0.65
82225[c]	secreted frizzled-related protein 1	0.97	0.79
135688	GATA-binding protein 2	1.00	0.77
383188	recoverin	1.00	0.61
325182	cadherin 2, N-cadherin (neuronal)	1.00	0.69
855786	tryptophanyl-tRNA synthetase	0.99	0.57
884719	heat shock 70kD protein 10 (HSC71)	0.73	0.12
486110[c]	profilin 2	1.00	0.38
878280[c]	collapsin response mediator protein 1	0.99	0.38
840788	Thymosin, beta 10	0.75	0.06
756405	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein	0.80	0.04
44563[c]	growth associated protein 43	1.00	0.81
1474174	matrix metalloproteinase 2 (gelatinase A, 72kD gelatinase, 72kD type IV collagenase)	0.99	0.78
882522	argininosuccinate synthetase	0.91	0.49
768299	butyrate response factor 1 (EGF-response factor 1)	0.96	0.04

[a,b,c] See Table 1.

example, there are several genes that have a high confidence score when the top 100 ranked genes are used, but which are low when only the top 20 genes are considered. However, if we use the more stringent criteria of confidence score based on top 20 genes, then we find that there are several candidate genes that are worthy of followup studies.

We also compared the rankings in Tables 1–3 with the selection criteria used by Khan et al. for selecting "discriminator" genes. We find that while there is some overlap between the two ranking schemes, there are several genes with high confidence scores that were not found using the Khan et al. (2001) scheme. In addition, several of the discriminator genes found in that article have relatively low confidence scores.

4.2 *Prostate Cancer Data*

We now apply the methodology to data from a prostate cancer study; a subset of the data was analyzed in Dhanasekaran et al. (2001). The goal of the study was to determine if gene expression profiles can be used to classify various types of prostate cancer. While the focus in the article was primarily on the prostate cancer (both local and metastatic) versus the non-prostate-cancer comparison, we will consider three classes: benign prostate hyperplasia (BPH), local prostate cancer (PCA) and metastatic prostate cancer (MET). There are 58 samples: 21 are BPH, 17 represent local prostate cancer, and 20 are metastatic prostate cancer samples. We divided the chips into a training set and a test set: the former consisted of 37 samples, while the remaining samples made up the test set. We estimated the optimal amount of smoothing for the penalized regression approaches using the training set.

The tissue samples were profiled using a spotted microarray chip with 9984 elements. We only considered genes that passed the filtering procedures of Dhanasekaran et al. (2001)

for at least all but one of the experiments. The remaining data were imputed using within-gene median imputation. This left a total of 3495 genes as predictors. Applying the proposed methodology yields an optimal classification accuracy of 81.0% using the ridge regression methods. The principal components regression-based classification scheme gives an optimal classification accuracy of 95.2%, where two principal components were used in the model. The partial least squares classification method gives an optimal accuracy of 95.2%, based on 7 components in the model. For the purposes of comparison, we fit a linear discriminant analysis using the entire gene expression profile; because of the collinearities in the predictors, the estimated class probabilities for the test data failed to converge. A procedure of principal components analysis, followed by linear discriminant analysis, yielded an optimal classification accuracy of 95.2%, based on two components. Because the number of samples in the three tissue classes are almost equal (14 BPH, 11 PCA, and 12 MET), this approach will be similar to the proposed principal components regression-based classification procedure. Applying the criterion of Dudoit et al. (2002), followed by linear discriminant analysis, yielded a classification accuracy of 85.7%. The random-forests method yielded a classification accuracy of 90.5%.

We can now perform a similar analysis for gene selection as was done in Tables 1–3 . Due to space limitations, we have decided to have the gene selection results at our web page, the URL for which was given in Section 3.5.

5. Discussion

There is a great need to adapt classical methods of regression modelling and discriminant analysis to microarray data. In this article, we have presented a general methodology for

classifying tumors based on data from such experiments. The method is fairly simple to implement and utilizes existing penalized regression models that have been used in other applications. Two advantages of this method is that it can incorporate correlations between genes into the analysis and that it can handle the situation where the number of predictors is larger than the number of samples. This method does not require any univariate filtering of predictors to fit the model.

The regression approach proposed in the article gives a direct method for ranking the individual predictor variables. We have also implemented a bootstrap method for assessing the reliability of these rankings. This is scientifically important, as it provides investigators a list of genes that could be studied for followup experiments. It can also be used as a guide in generating hypotheses regarding the biological pathways in cancer. Another use of the ranked gene list, mentioned in Khan et al. (2001), is for designing diagnostic arrays. One point to note is that in the examples presented in Section 4, the ranking of genes is performed conditional on selecting an optimal regularization parameter (i.e., optimal choice of λ or for number of components in model). In practice, it is important to select several values of the regularization parameter to see that the ranking of genes does not change very much.

We found in the numerical examples that we could generally achieve 80–100% classification accuracy based on the methods we have constructed. It might be possible in certain instances to improve this predictive performance using bagging methods (Breiman, 1996). This technique tends to reduce the prediction error for procedures that are inherently unstable. While ridge regression is stable, selecting the number of components for principal components and partial least squares regression is unstable, so bagging might be a useful method for improving the performance of those methods.

In terms of comparisons between ridge regression, PCR, and PLS, our analytical discussion in Section 3.1 suggests that better classifiers would be constructed by PLS, because it is equivalent to a "canonical covariance" analysis. In our experience with real datasets, the PLS-based method tends to work the best as well. However, one computational advantage of ridge regression is that leave-one-out cross-validation can be performed quickly, because the diagonal of the hat matrix can be calculated quickly.

Acknowledgements

Résumé

En raison de l'avancée des technologies de micropuces haut débit, il est devenu possible de développer des systèmes de classifications moléculaires pour différents types de cancer. Dans cet article, nous proposons une méthode fondée sur les modèles de régression régularisée pour la classification des tumeurs dans les expériences de micropuces. Les performances des modèles en composantes principales, des moindres carrés partiels et de la ridge régression sont étudiées; Ces procédures de régression sont adaptées aux techniques de classification utilisant l'algorithme de score optimal. Nous développons également une méthode de classement des gènes fondée sur les modèles de régression ajustée. Les méthodes proposées sont appliquées à deux études de cancérologie utilisant les micropuces.

References

Alizadeh, A. A., Ross, D. T., Perou, C. M., and van de Rijn, M. (2001). Towards a novel classification of human malignancies based on gene expression patterns. *Journal of Pathology* **195,** 41–52.

Alter, O., Brown P. O., and Botstein D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97,** 10101–10106.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24,** 123–140.

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

Bull, S. and Donner, A. (1987). The efficiency of multinomial logistic regression compared with multigroup normal discriminant analysis. *Journal of the American Statistical Association* **82,** 1118–1122.

Denham, M. C. (1994). Implementing partial least squares. *Statistics and Computing* **5,** 191–202.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412,** 822–826.

Dudoit, S., Fridyland, J. F., and Speed, T. P. (2002). Comparison of discrimination methods for tumor classification based on microarray data. *Journal of the American Statistical Association* **97,** 77–87.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70,** 113–121.

Eilers, P. H. C., Boer, J. M., van Ommen, G.-J., and van Houwelingen, H. C. (2001). Classification of microarray data with penalized logistic regression. *Proceedings of SPIE* **4266,** 187–198.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics* **35,** 109–143.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19,** 1–141.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning.* New York: Springer.

Golub, T. R., Slonim, D. K., Tamayo, P., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* **89,** 1255–1270.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics* **23,** 73–102.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics—Simulation and Computation* **17,** 581–607.

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

Li, H. and Hong, F. (2001). Cluster-Rasch models for microarray data. *Genome Biology* **2,** 1–13.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society*, *Series B* **34,** 1–40.

Martens, H. and Naes, T. (1989). *Multivariate Calibration*. New York: Wiley.

Marx, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* **38,** 374–381.

Marx, B. D. and Smith, E. P. (1990). Principal components estimation for generalized linear regression. *Biometrika* **77,** 23–31.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association* **60,** 234–246.

Naes, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data. *Communications in Statistics—Simulation and Computation* **14,** 545–576.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction (with discussion). *Journal of the Royal Statistical Society*, *Series B* **36,** 111–147.

Stone, M. and Brooks, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *Journal of the Royal Statistical Society*, *Series B* **52,** 237–269.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7,** to appear.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, 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.

Wold, H. (1975). Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett,* J. Gani (ed). London: Academic Press.

Yeang, C.-H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J., and Golub, T. (2001). Molecular classification of multiple tumor types. *Bioinformatics* **17,** 316S–322S.