



Graphical methods for class prediction using dimension reduction techniques on DNA microarray data

Efstathia Bura^{1,*} and Ruth M. Pfeiffer²

¹Department of Statistics, The George Washington University, 2201 G Street NW, Washington, DC 20052, USA and ²National Cancer Institute, DCEG, 6120 Executive Blvd, EPS/8030, Bethesda, MD 20892-7244, USA

Received on May 20, 2002; revised on November 5, 2002; accepted on January 26, 2003

ABSTRACT

Motivation: We introduce simple graphical classification and prediction tools for tumor status using gene-expression profiles. They are based on two dimension estimation techniques sliced average variance estimation (SAVE) and sliced inverse regression (SIR). Both SAVE and SIR are used to infer on the dimension of the classification problem and obtain linear combinations of genes that contain sufficient information to predict class membership, such as tumor type. Plots of the estimated directions as well as numerical thresholds estimated from the plots are used to predict tumor classes in cDNA microarrays and the performance of the class predictors is assessed by cross-validation. A microarray simulation study is carried out to compare the power and predictive accuracy of the two methods.

Results: The methods are applied to cDNA microarray data on *BRCA1* and *BRCA2* mutation carriers as well as sporadic tumors from Hedenfalk *et al.* (2001). All samples are correctly classified.

Contact: eburag@gwu.edu

INTRODUCTION

DNA microarrays belong to a new class of biotechnologies that allow simultaneous monitoring of expression levels for thousands of genes. The areas of application include cell line classification, understanding the effects of therapeutic agents, and distinguishing tumor types by identifying genes with differentiated expression levels. Two microarray construction technologies dominate the field—oligonucleotide and spotted cDNA microarrays. The statistical analysis problems are similar for both types.

Microarray data are summarized in an $n \times k$ matrix $X = (x_{ij})$. Typically, the genes (k) are the rows and different individuals or subjects (n) correspond to the columns so that the microarray equals X^T . Contrary to

traditional statistical set-ups, the number of subjects n is much smaller than the number of genes k , with k being in the thousands and n below 100. When samples belong to known classes, e.g. in our application, tumor tissue of *BRCA1* or *BRCA2* germline mutation carriers, the data also contain a class label or response Y for each subject in the sample.

Microarray data give rise to two types of problems: class discovery and class prediction. In class prediction, which is the focus of this paper, observations are known to belong to prespecified classes and the task is to build predictors for assigning new observations to these classes. Standard statistical analyses include linear discriminant analysis and diagonal linear discriminant (DLD) classifiers, classification trees, nearest neighbor (NN) and aggregating classifiers. A comprehensive account of such methods and a comparison of their performance is given by Dudoit *et al.* (2002). These methods operate on the variable selection principle as opposed to modeling how classes of genes function to predict the outcome. In effect, DLD classifiers ignore correlations and interactions between predictor variables, i.e. genes. Both correlations and interactions are biologically important and they may have an effect on the classification, or they may yield an insight into the predictive structure of the data. NN classifiers, on the other hand, do take interactions into consideration but in a 'black-box' way which does not aid in understanding the underlying biological process.

The sufficient dimension reduction methods we discuss in this paper, sliced average variance estimation (SAVE) and sliced inverse regression (SIR), capitalize on the correlations among the genes to identify a small number of linear combinations of a subset of genes that can be used to predict cancer tumor genotype. They differ from other approaches in that they do not require the specification of a model in order to estimate linear combinations of genes that contain all the regression information. Once the linear combinations have been identified, they can be used as

*To whom correspondence should be addressed.

input to any classifier in order to predict class membership of a new observation. We will focus on graphical displays of the reduced data to discriminate the different classes and predict the class status of a sample.

SYSTEMS AND METHODS

Introduction to Dimension Reduction Methods based on Inverse Regression

A convenient data reduction formulation, that accounts for the correlation among genes, is to assume there exists a $k \times p$, $p \leq k$, matrix $\boldsymbol{\eta}$ so that

$$F(Y|X) = F(Y|\boldsymbol{\eta}^T X) \quad (1)$$

where $F(\cdot|\cdot)$ is the conditional distribution function of the response Y given the second argument. In the microarray data analysis context, Y could be tumor class or survival time, and the predictor vector X comprises the different gene expressions. The statement in (1) implies that the $k \times 1$ predictor vector X can be replaced by the $p \times 1$ predictor vector $\boldsymbol{\eta}^T X$ without loss of information. That is, $\boldsymbol{\eta}^T X$ contains equivalent or *sufficient*, in the statistical sense, information for the regression of Y on X . Most importantly, if $p < k$, then sufficient reduction in the dimension of the regression is achieved. The linear subspace $S(\boldsymbol{\eta})$ spanned by the columns of $\boldsymbol{\eta}$ is a *dimension-reduction* subspace (Li, 1991) and its dimension denotes the number of linear combinations of the components of X needed to model Y . When (1) holds, then it also holds with $\boldsymbol{\eta}$ replaced by any basis for $S(\boldsymbol{\eta})$.

Clearly, knowledge of the smallest dimension-reduction subspace would provide the most parsimonious characterization of Y given X , as it provides the greatest dimension reduction in the predictor vector. Let $S_{Y|X}$ denote the unique smallest dimension-reduction subspace, referred to as the *central subspace* (Cook, 1996). The dimension $d = \dim(S_{Y|X})$ is called the structural dimension of the regression of Y on X , and can take on any value in the $\{0, 1, \dots, k\}$ set.

The estimation of the central subspace is based on finding a kernel matrix \mathbf{M} so that $S(\mathbf{M}) \subset S_{Y|X}$. There have been two main approaches. The first uses first moment methods such as SIR and variations (Li, 1991) with $\mathbf{M} = \text{Cov}(E(X|Y))$, and polynomial inverse regression (Bura and Cook, 2001a) with $\mathbf{M} = E(X|Y)$. The second approach uses second moment methods such as pHd (Li, 1991) with $\mathbf{M} = E((Y - E(Y))XX^T)$, SAVE (Cook and Weisberg, 1991) with $\mathbf{M} = E(\text{Cov}(X) - \text{Cov}(X|Y))^2$, and SIRII (Li, 1991) with $\mathbf{M} = E(\text{Cov}(X|Y) - E(\text{Cov}(X|Y)))^2$. SAVE is the most inclusive among dimension reduction methods as it gains information from both the inverse mean function and the differences of the inverse covariances.

The two conditions needed for all kernel matrices

\mathbf{M} to span subspaces of the central dimension reduction subspace are that $E(X|\boldsymbol{\gamma}^T X)$ be linear, and that $\text{Var}(X|\boldsymbol{\gamma}^T X)$ be constant. Both conditions refer to the marginal distribution of the predictors and are necessarily satisfied when X is a normal vector. The conditions are empirically checked by considering the scatterplot matrix of the predictors. Linearity of $E(X|\boldsymbol{\gamma}^T X)$ can be ascertained if the scatterplots look roughly linear or random, and homogeneity of the variance holds if there are no pronounced fluctuations in data density in the scatterplots. In practice, only substantial departures from both conditions are problematic.

All aforementioned methods can be used to estimate directions in $S_{Y|X}$. We will present SIR and SAVE in more detail in the following section, as they are readily applicable to categorical responses.

IMPLEMENTATION

Without loss of generality one can use standardised predictors $\mathbf{Z} = \boldsymbol{\Sigma}_x^{-1/2}(X - E(X))$ (assuming that $\boldsymbol{\Sigma}_x = \text{Var}(X)$ is nonsingular), since $S_{Y|X} = \boldsymbol{\Sigma}_x^{-1/2}S_{Y|Z}$ so that the columns of the matrix $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\eta}$ form a basis for the central subspace $S_{Y|Z}$ for the regression of Y on Z .

While both SIR and SAVE can accommodate continuous as well as binary outcomes, our objective is to identify predictors (genes) that best predict the binary phenotype Y_i of a sample, or equivalently, $\Pr(Y_i = 1) = E(Y_i)$. Let $\boldsymbol{\mu}_j = E(\mathbf{Z}|Y = j)$, $\boldsymbol{\Sigma}_j = \text{Var}(\mathbf{Z}|Y = j)$, $j = 0, 1$, denote the conditional means and variances, respectively, for the binary response Y , and $\boldsymbol{\nu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$. We denote the subspace spanned by the columns of the two differences, $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}$, by $S(\boldsymbol{\nu}, \boldsymbol{\Delta})$. The main result by Cook and Lee (1999) states that the SAVE kernel matrix $\mathbf{M}_{SAVE} = (\boldsymbol{\nu}, \boldsymbol{\Delta})$ and hence that $S_{SAVE} = S(\boldsymbol{\nu}, \boldsymbol{\Delta})$ contains some or all the sufficient linear combinations that can replace the predictor vector \mathbf{Z} in the regression of Y on \mathbf{Z} , under the moment conditions stated in the previous section. Furthermore, when the conditional distribution of $\mathbf{Z}|Y$ is normal, then $S_{SAVE} = S_{Y|Z}$. Cook and Lee (1999) also showed that $\mathbf{M}_{SIR} = \boldsymbol{\nu}$, that is, $S_{SIR} = S(\boldsymbol{\nu}) \subset S_{SAVE}$.

In implementing the method, $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}$ are replaced by the corresponding sample moments, $\hat{\boldsymbol{\nu}} = \hat{\boldsymbol{\Sigma}}_x^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$ and $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Sigma}}_x^{-1/2}(\hat{\boldsymbol{\Sigma}}_{x|1} - \hat{\boldsymbol{\Sigma}}_{x|0})\hat{\boldsymbol{\Sigma}}_x^{-1/2}$ to yield $\hat{S}_{SAVE} = S(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Delta}})$, a $k \times (k + 1)$ matrix, and $\hat{S}_{SIR} = S(\hat{\boldsymbol{\nu}})$, a $k \times 1$ vector. The latter has obviously dimension of at most 1. The test statistic for dimension is given by $\Lambda_d = n \sum_{l=d+1}^k \hat{\lambda}_l^2$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \hat{\lambda}_k$ are the singular values of the estimated kernel matrix $\hat{\mathbf{M}}_{SAVE} = (\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Delta}})$, or $\hat{\mathbf{M}}_{SIR} = \hat{\boldsymbol{\nu}}$, depending on the method used.

Cook and Lee (1999) showed that the test statistic for SAVE has an asymptotic weighted chi-squared distribu-

tion. In the case of microarray data, asymptotic tests may not be appropriate as the sample size is typically fairly small. We use a permutation test (see Cook and Yin, 2001) to estimate the dimension d of S_{SAVE} . The SIR test statistic for dimension has an asymptotic chi-squared distribution (Li, 1991; Bura and Cook, 2001b). In both cases, the estimation is carried out by performing a series of tests for testing $H_0 : d = m$ against $H_a : d > m$, starting at $m = 0$, which corresponds to independence of Y and Z .

If the dimension is estimated to be, say, m , then the eigenvectors $\mathbf{b}_1, \dots, \mathbf{b}_m$, corresponding to the m largest singular values of \hat{S}_{SAVE} or \hat{S}_{SIR} span a subspace which estimates $S_{Y|Z}$. Back-transforming to $\Sigma_x^{-1/2}\mathbf{b}_1, \dots, \Sigma_x^{-1/2}\mathbf{b}_m$ yields estimates of m basis elements of the central subspace $S_{Y|X}$. The resulting predictors $\hat{\Sigma}_x^{-1/2}\mathbf{b}_1, \dots, \hat{\Sigma}_x^{-1/2}\mathbf{b}_m$ are linear combinations of the original k regressors that contain sufficient information to model the response, Y . For binary classification problems, SIR can estimate at most one basis element of $S_{Y|X}$.

REMARKS. (a) Even though the discussion so far has concentrated on binary response regressions, both SIR and SAVE can be applied to problems with multinomial or multi-valued responses. (b) When X is normally distributed, SIR is equivalent to linear discriminant analysis in the sense that they both estimate the same discriminant linear combinations of the predictors. Unfortunately, in binary regression both LDA and SIR estimate at most one direction in the central dimension reduction subspace and may miss important relevant information that could improve the accuracy of classification. Chiaromonte and Martinelli (2002) were the first to apply SIR on microarrays. (c) When X is a normal vector, SAVE is equivalent to Quadratic Discriminant Analysis (Cook and Yin, 2001). (d) It is not required X be normal for either SIR or SAVE to yield directions in the central subspace.

Class prediction via minimal sufficient summary plots and numerical thresholds

Sufficient dimension reduction techniques estimate the dimension as well as the directions in the central subspace without requiring the specification of any underlying parametric model. Thus, they constitute a pre-processing tool for the data that aims to facilitate and guide further analysis. For the class prediction problem that we consider in this paper, the directions obtained via any dimension reduction method can be subsequently used in a discriminant function for class prediction. This approach is also advocated by Flury *et al.* (1997).

Instead of using a classifier function, we use graphical displays obtained by both the SIR and SAVE directions directly to classify samples. To predict class membership, the class label Y is plotted against the estimated SIR and

SAVE linear combinations $\hat{\eta}^T X$, where X is the vector of gene expressions. As this is a sufficient summary plot, the classes are expected to be completely separated by $\hat{\eta}^T X$. These plots can be also used to estimate a numerical threshold that separates the classes. In one-dimensional problems this threshold will be a single point, in two-dimensional problems it will be a line, and in general, in a d -dimensional problem it will be a $d - 1$ -dimensional separating hyperplane. For example, in the case of SAVE applied to a 2D problem, the separating line is estimated by first identifying the best discriminating view of the 3D plot of Y versus the two SAVE predictors, and projecting it to a 2D view. The line that has the maximum distance from both classes is the separating line.

Software

The data analysis was carried out in *Arc*, a regression analysis software that includes SAVE, SIR and 3D dynamic graphics developed by Cook and Weisberg (1999). *Arc* can be downloaded freely from <http://www.stat.umn.edu/arc/>. The dimension reduction software is also available as package *dr* in R from <http://cran.r-project.org/>.

RESULTS

Simulation study

We compare the performance of SAVE and SIR on simulated data of structural dimension 2. The data were generated by adapting the approaches used by Kepler *et al.* (2002) and Cook and Lee (1999).

We assumed two classes with labels $Y = 0$ and $Y = 1$. For each class we generated 50 independent samples of 1000 gene expressions (X) as follows: The data $X|Y = 0$ were simulated from a multivariate normal distribution, with mean 0 and covariance matrix Σ . For the $Y = 1$ group, we selected 1% of the genes to be differentially expressed. They were generated from a mixture of two multivariate normal distributions, with means μ_1 and μ_2 , respectively, and same covariance structure Σ . The mixing probability was chosen to be 1/2. The means μ_1 and μ_2 correspond to the log of the fold changes, or log of the ratios of expression level between the two groups. The non-differentially expressed 99% of the $X|Y = 1$ values followed the same $N(0, \Sigma)$ distribution as the $X|Y = 0$ genes. Cook and Lee (1999) showed that for the set of differentially expressed genes, the dimension of the central subspace equals two. The covariance matrix $\Sigma = (\sigma_{ij})$ had a block structure with $\sigma_{ij} = 0.2$ for $|j - i| \leq 5$ and zero otherwise.

Five hundred replications of the experiment were produced. First, we identified differentially expressed groups of genes using the standard two-sample t -test at level $\alpha = 0.05/(2 \times 1000) = 0.000025$. We then applied SAVE and SIR to the differentially expressed genes. SIR was

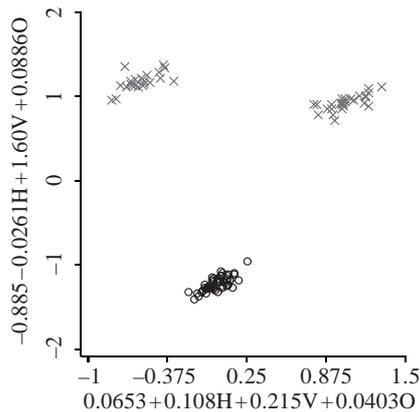


Fig. 1. Simulation study: 2D view of the 3D plot of the two SAVE predictors. The symbol \times indicates $Y = 1$. The axes coordinates are also given.

of course unable to identify the second dimension. SAVE had power equal to 1 for identifying the first direction and 0.994 for the second. The level of the test was empirically estimated as 0.038, slightly smaller than the nominal 0.05. Ten plots were randomly selected among the 500 in order to assess the discriminating power of the SIR and SAVE directions. SAVE was 100% accurate in separating the two classes. Three of the 10 SIR summary plots did not separate the classes successfully. Additionally, the SAVE sufficient summary plots were able to pick up the underlying structure of the data. Figure 1 shows a 2D view of the 3D graph of the class response versus the two SAVE directions. The three data clouds, of which the top two correspond to class label 1, are clearly distinguished and reveal the mixture structure of the data.

cDNA microarrays on breast cancer

The data consist of 22 cDNA microarrays, each representing 5361 genes based on biopsy specimens of primary breast tumors of seven patients with germ-line mutations of *BRCA1*, eight patients with germ-line mutations of *BRCA2*, and seven with sporadic cases. These data were first presented and analysed by Hedenfalk *et al.* (2001). Information on the data can be found in <http://www.nejm.org> and <http://www.nhgri.nih.gov/DIR/Microarray>. The analysis focuses on identifying groups of genes that can be used to predict class membership to the two *BRCA* mutation carrier groups. The problem consists of two parts: the classification of *BRCA1* carriers, and the classification of *BRCA2* carriers. The class label $Y_i^{(1)} = 1$ if the i th subject carries a *BRCA1* mutation, and it is 0 otherwise. The *BRCA2* mutation carrier class is defined by $Y^{(2)}$ similarly.

Hedenfalk *et al.* (2001) predict class membership using a compound covariate predictor, $c_i = \sum_j t_j x_{ij}$, where

t_j is the t -statistic for the two group comparison of classes with respect to gene j , x_{ij} is the log-ratio measured in tumor sample i for gene j , and the sum is over all differentially expressed genes. The choice of t -statistics as the coefficients of the genes is reasonable, but nevertheless, a rather ad-hoc approach.

Hedenfalk *et al.* first identified differentially expressed groups of genes using the standard two-sample t -test at level $\alpha = 0.0001$. Each sample was then classified by comparing its compound covariate score of the differentially expressed genes with a classification threshold, taken to be the midpoint of the means of the compound covariates for the two classes. Their prediction algorithm resulted in one misclassification for the two *BRCA1* groups and four misclassifications for the two *BRCA2* groups. Their classification process involved leave-one-out cross validation at each step so that different sets of differentially expressed genes were identified for each sample. Their sizes varied from 4 to 15 for the *BRCA1*-mutation-positive, and from 3 to 14 for the *BRCA2*-mutation-positive groups. The gene list was made available to us by personal communication. We cross validated our procedures in the same manner, by first cross-validating the pre-selection of the genes and, secondly, the SAVE and SIR classifiers.

We considered all scatterplot matrices of the 44 sets of predictors for the *BRCA1*-mutation-positive and the *BRCA2*-mutation-positive groups. They all indicated that both the linearity and the constant variance assumptions hold. Then, both SIR and SAVE were applied to the 22 different sets of genes for the *BRCA1* and *BRCA2* groupings. In the case of *BRCA1* and for all 22 samples, SIR identified one significant direction, and SAVE identified two, except for case 7 for which both estimated the dimension as 1. We next plotted the binary response Y versus the SIR direction, and the two SAVE directions (linear combinations of the gene expressions). In all 22 cases, both SIR and SAVE directions perfectly discriminate the two classes graphically. SAVE correctly classifies all 22 samples, and SIR misclassified only specimen 6. For illustration, the plot of Y versus the SIR direction with the third sample excluded and subsequently added is depicted in Figures 2 and 3, respectively. Also for the third sample, Figures 4 and 5 show 2D views of the two SAVE predictors where the *BRCA1* mutation status classes are clearly separated with the third observation excluded and included, respectively. The six genes that were used as input to SAVE and SIR, for the third observation, are (their image clone ID is parenthesized): sigma (258), Peroxisomal acyl-coenzyme A oxidase [human, liver, mRNA, 3086 nt] (659), CDC28 (809), keratin (1008)—this gene appeared in most of the 22 gene sets, ESTs (1859), ESTs (1999), very (2423), minichromosome (2734).

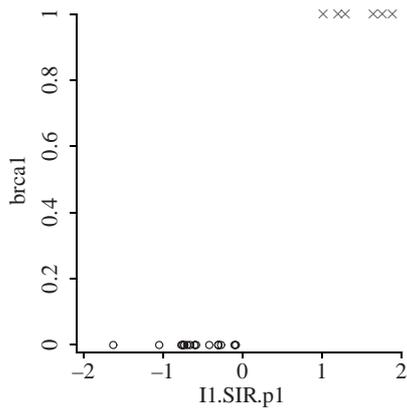


Fig. 2. *BRCA1* status label versus SIR direction without the third sample. The symbol \times indicates *BRCA1* mutation present.

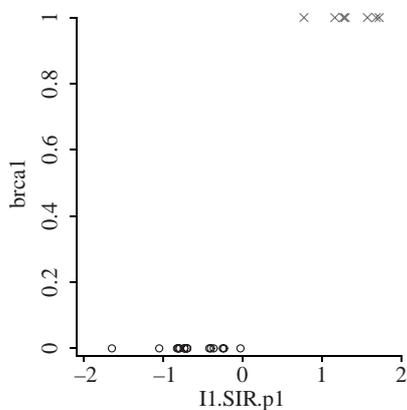


Fig. 3. *BRCA1* status label versus SIR direction with the third sample. The symbol \times indicates *BRCA1* mutation present.

For the *BRCA2*-mutation-positive and *BRCA2*-mutation-negative classes, SIR identified one significant direction, and SAVE identified two, except for case 7 for which both estimated the dimension as 1. Next, the binary response Y was plotted versus the SIR direction, and the two SAVE directions (linear combinations of the gene expressions). In all 22 cases, both SIR and SAVE directions perfectly discriminate the two classes graphically and correctly classify all specimens.

Our analysis confirms the results of Hedenfalk *et al.* that these sets of genes can be used to predict *BRCA1* and *BRCA2* mutation status. SAVE estimated the number of linear combinations of the covariates (genes) needed for the prediction to be two as opposed to a single covariate SIR and also Hedenfalk *et al.* used. The slight improvement of accuracy of the SAVE class prediction plot, no misclassifications, over the SIR summary plot, one misclassification for the two *BRCA1* classes, points

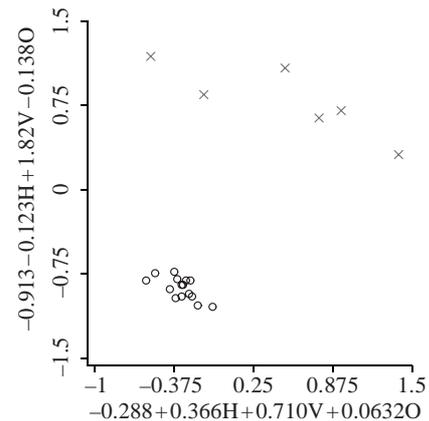


Fig. 4. 2D view of the 3D plot of the two SAVE predictors without third observation. It has *BRCA1* positive (1) status. The symbol \times indicates *BRCA1* mutation present. The axes coordinates are also given.

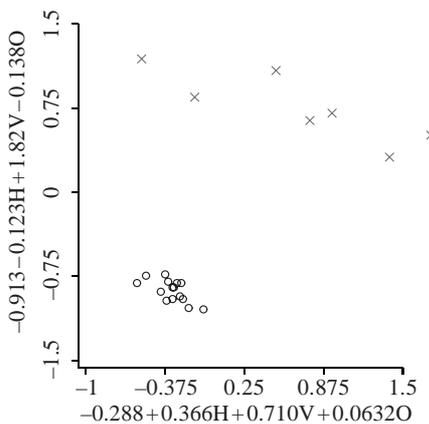


Fig. 5. 2D view of the 3D plot of the two SAVE predictors with third observation. It has *BRCA1* positive (1) status. The symbol \times indicates *BRCA1* mutation present. The axes coordinates are also given.

to the fact that one linear combination may be inadequate. Nevertheless, the less complex SIR performed remarkably well.

We also estimate numerical classification thresholds from the summary plots. For example, from the SIR summary plots, zero is estimated to be the separating point of the two classes for both *BRCA1* and *BRCA2*. As mentioned previously, for SAVE the separating line is estimated by projecting the best discriminating view of the 3D plot, Y versus the two SAVE predictors, to a 2D view using *Arc*. *Arc* supports 3D dynamic graphics and also computes their 2D projections. For the full data set the separating line was estimated as $y = -x + .25$.

In an additional analysis we used the odds from logistic regression models as a predictor function to predict class membership for both *BRCA1* and *BRCA2* classes. We also analyzed the data as a three class problem—*BRCA1*, *BRCA2*, and sporadic—using the 51 genes that were found to be differentially expressed in all three groups. Details on both analyses can be found at <http://home.gwu.edu/~ebura/publications.html>.

In summary, our analysis points to the existence of multiple subsets of genes that can be used for class prediction with similar accuracy. While this may at first seem an unattractive attribute of these data, it is not surprising, as the number of cases is too small to capture unique differences in gene expressions. Our finding agrees with the criticism for the work of Hedenfalk *et al.* (2001) expressed in the correspondence section of the article by Sudbø, Reith and Lindgærde. Results of applying SAVE and SIR to a second example on oligonucleotide arrays on acute leukemia first presented by Golub *et al.* (1999) are available at <http://home.gwu.edu/~ebura/publications.html>.

DISCUSSION

We introduced graphical displays based on dimension reduction methods such as SIR and SAVE, as tools for classification of tumor tissue using gene expression profiles. While the methods apply to data that belong to multiple classes we present the implementation of the two-class analysis in detail. Two properties of the distribution of the predictors are required, namely the linearity of the conditional expectation and constant variance of the marginal distribution of the predictors. These conditions are trivially satisfied when the predictors are normally distributed. The logarithmic transformation of the dye intensity ratios in cDNAs appears to frequently induce data that comply with both conditions, for example the breast cancer data we consider in this paper.

Dimension reduction methods such as SIR and SAVE estimate both the dimension of the regression of the binary response on a set of genes and the linear combinations of the genes that contain sufficient information for class prediction. Effectively, they reduce the dimension of the prediction problem so that classifiers with known statistical properties can be used for the allocation of the samples to the two classes. In our example, we used a simple graphical approach to classifying the specimens.

The most appealing and unique feature of this formulation for dimension reduction is that it allows formal statistical inference on dimension without imposing a particular model for the functional relationship between Y and X , as is the case with most other dimension reduction methods, e.g. *additive, generalized additive and projection pursuit models* (Friedman and Stuetzle, 1981; Hastie

and Tibshirani, 1990), *partially linear or spline models, single- and multi-index models* (Green and Silverman, 1994). Both SAVE and SIR become increasingly powerful with increasing sample sizes, allowing for the inclusion of more genes in the dimension reduction phase. However, the only limitation to the number of genes used in the analysis is numerical instability induced by having to invert the sample covariance matrix Σ_x . As an alternative to using t -tests to preselect genes, one could also use a generalized inverse which corresponds to replacing the original X vector with its principal components.

Among the two methods, SAVE is the most comprehensive as it captures a larger section of the central subspace (Cook and Critchley, 2000), without any restrictions in the dimension it can estimate. It should be noted though that, in particular when the sample size is very small, SIR may have a computational advantage over SAVE as it requires the estimation only of the inverse mean vector.

ACKNOWLEDGEMENTS

We would like to thank the associate editor and four anonymous referees for their comments that helped improve the paper significantly. The first author would like to thank the National Science Foundation and Ingram Olkin (DMS-9631278) for giving her the opportunity to visit the Department of Statistics at Stanford University, where part of this work was carried out. Her research was also supported by National Science Foundation grant DMS-0204563.

REFERENCES

- Bura,E. and Cook,R.D. (2001a) Estimating the structural dimension of regressions via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B*, **63**, 393–410.
- Bura,E. and Cook,R.D. (2001b) Extending SIR: the weighted Chi-squared test. *J. Am. Stat. Assoc.*, **96**, 996–1003.
- Chiaromonte,F. and Martinelli,J. (2002) Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.*, **176**, 123–144.
- Cook,R.D. (1998) *Regression Graphics*. Wiley, New York.
- Cook,R.D. (1996) Graphics for regressions with a binary response. *J. Am. Stat. Assoc.*, **91**, 983–992.
- Cook,R.D. and Critchley,F. (2000) Identifying outliers and regression mixtures graphically. *J. Am. Stat. Assoc.*, **95**, 781–794.
- Cook,R.D. and Lee,H. (1999) Dimension reduction in binary response regression. *J. Am. Stat. Assoc.*, **94**, 1187–1200.
- Cook,R.D. and Nachtsheim,C.J. (1994) Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Stat. Assoc.*, **89**, 592–599.
- Cook,R.D. and Weisberg,S. (1999) *Applied Regression Including Computing and Graphics*. Wiley, New York.
- Cook,R.D. and Weisberg,S. (1991) Discussion of ‘Sliced inverse regression’ by K.C.Li. *J. Am. Stat. Assoc.*, **86**, 328–332.
- Cook,R.D. and Yin,Z. (2001) Dimension reduction and visualization in discriminant analysis. *Aust. N. Z. J. Stat.*, **43**, 147–199.

- DeRisi,J., Penland,L., Brown,P.O. Bittner,M.L. *et al.* (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.*, **14**, 457–460.
- Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumours using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
- Flury,L., Boukai,B. and Flury,B. (1997) The discrimination subspace model. *J. Am. Stat. Assoc.*, **92**, 758–766.
- Friedman,J.H. and Stuetzle,W. (1981) Projection pursuit regression. *J. Am. Stat. Assoc.*, **76**, 817–823.
- 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.
- Green,P.J. and Silverman,B.W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall, London.
- Hastie,T.J. and Tibshirani,R. (1990) *Generalized Additive Models*. Chapman and Hall, London.
- Hedenfalk,M.S., Duggan,D. Chen,Y. *et al.* (2001) Gene expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.
- Kepler,T.B., Crosby,L. and Morgan,K.T. (2002) Normalization and analysis for DNA Microarray data by self-consistency and local regression. *Genome Biol.*, **3**, RESEARCH0037.
- Li,K.C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, **86**, 316–342.
- Lipshutz,R.J., Fodor,S.P.A., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**, 20–24.
- Schena,M. (ed.) (1999) *DNA Microarrays: A Practical Approach*. Oxford University Press.
- Schena,M. (ed.) (2000) *Microarray Chip Technology*. Eaton.