

# SINGULAR VALUE DECOMPOSITION REGRESSION MODELS FOR CLASSIFICATION OF TUMORS FROM MICROARRAY EXPERIMENTS

DEBASHIS GHOSH

*Department of Biostatistics, University of Michigan  
1420 Washington Heights, Ann Arbor, MI 48109-2029  
ghoshd@umich.edu*

An important problem is the analysis of microarray data in correlating the high-dimensional measurements with clinical phenotypes. In this paper, we develop predictive models for correlating gene expression data from microarray experiments with such outcomes. By using these models, we can achieve three goals. First, we can model the effects of gene expression profile on disease; this leads to the second goal of identifying candidate genes for further followup studies. Finally, clustering of the discriminatory genes can be accomplished in a natural way such that clinical information is incorporated. The regression modelling in this paper is a two-stage procedure. In the first stage, the gene expression measurements are transformed using singular value decomposition. The second stage involves formulating a regression model linking the principal components with the clinical responses. We demonstrate the application of the methodology to data from a breast cancer study.

## 1 Introduction

DNA biochips have the potential of significantly impacting the study of human disease. By simultaneously gauging the expression of thousands of genes in clinical specimens, a wealth of data points is generated coalescing to form a molecular fingerprint of a disease process. Such experiments have been performed on acute leukemias, lymphomas, breast cancers and cutaneous melanomas.<sup>1,2,3</sup> Obtaining large-scale gene expression profiles of tumors, should theoretically allow for the identification of subsets of genes that function as prognostic disease markers or biologic predictors of therapeutic response.

Most primary analyses involve hierarchical clustering techniques.<sup>4</sup> However, in many instances, there is external clinical information (such as survival time or tumor type) that the investigators use in secondary analyses. A clustering analysis that utilized the external clinical data would make better use of the information.

For many molecular profiling studies, the goal is to find candidate genes that successfully discriminate between disease classes. These genes can then be screened for further follow-up studies using immunohistochemical techniques such as tissue microarrays.<sup>5</sup>

Some preliminary work has been put forward correlating gene expression data with clinical outcomes.<sup>6,7</sup> However, the approaches that have been taken so far have been univariate and ignore correlations between genes. A potential problem with joint modelling of gene effects on clinical outcomes is that the number of genes under consideration is typically much larger than the number of samples profiled. In statistical terminology, the number of predictors is much larger than the number of independent samples. Consequently, it is not possible to find regression parameter estimates using traditional statistical procedures.

In this paper, we develop a regression framework for correlating gene expression data with clinical phenotypes. Our goal is threefold. First, we wish to develop predictive models for important clinical outcomes. Second, we wish to identify candidate genes for further experimentation. Third, we want to cluster genes incorporating the clinical phenotype data. While the framework presented here is relatively broad, we are motivated by the specific problem of modelling the association between gene expression profiles with type of tumor. The regression modelling will be performed using singular value decomposition (SVD), a technique that has been applied to other areas of analysis of microarray data.<sup>8,9,10</sup> In the statistical literature, singular value decomposition analysis is known as principal components analysis; we will use the two terms interchangeably throughout the paper. Regression modelling using SVD has been done with great success in other areas of application, such as chemometrics.<sup>11</sup> A complication in the current setting that does not arise in other applications is that the clinical outcome may not be continuous. Our proposal here involves using multinomial logistic regression modelling for associating the gene expression measurements with tumor type.<sup>12</sup> We demonstrate the procedure using data from a recently published breast cancer study.<sup>13</sup>

## 2 Algorithm

Before describing the regression model for correlating gene expression profiles with tumor phenotype, we introduce some notation. Let  $\mathbf{X}_i$  denote the  $p$ -dimensional column vector of gene expression measurements for the  $i$ th subject,  $i = 1, \dots, n$ . Note that  $p$  will typically be much larger than  $n$ . For  $i = 1, \dots, n$ , we define  $Y_i$  to be the tumor type for the  $i$ th individual; this will take values  $0, 1, \dots, J - 1$ , where  $J$  is the number of tumor types. We will assume that the  $\mathbf{X}_i$  are standardized across chips to have mean zero and variance one for each gene.

### 2.1 Regression model and estimation

We formulate the effects of gene expression on tumor type using the following multinomial logistic regression model:

$$\frac{P(Y_i = r)}{P(Y_i = 0)} = \beta_{r0}^T \mathbf{X}_i, \quad (1)$$

where  $P(A)$  is the probability of the event  $A$ ,  $\mathbf{a}^T$  is the transpose of the vector or a matrix  $\mathbf{a}$ , and  $\beta_{r0}$  is a  $p$ -dimensional vector of unknown regression coefficients,  $r = 1, \dots, J-1$ . The model is quite general in that separate gene effects are modelled for each of the  $J(J-1)/2$  tumor comparisons. More structure in the model can be imposed by placing constraints on  $\beta_{r0}$  ( $r = 1, \dots, J-1$ ). The constraints imposed on the regression parameters should incorporate the existing biological knowledge of the system under study. However, we will deal with the general model (1) throughout the paper.

In a typical microarray experiment, it is not possible to estimate the parameters in (1) using standard statistical methods because  $p$  is much larger than  $n$ . We propose using the singular value decomposition to reduce the dimension of  $\beta_{r0}$ . If we let  $\mathbf{X}$  denote the  $p \times n$  matrix  $[\mathbf{X}_1 \cdots \mathbf{X}_n]$ , then the singular value decomposition leads to the following decomposition of  $\mathbf{X}$ :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}, \quad (2)$$

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  $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ , where  $d_1 \geq d_2 \geq d_3 \geq \dots \geq d_n \geq 0$ . We will assume without loss of generality that  $d_i > 0$  for  $i = 1, \dots, n$ . Finally,  $\mathbf{V}$  is the  $n \times n$  singular value decomposition factor matrix and has both orthonormal rows and columns. The algorithms used to compute the singular value decomposition are typically iterative and quite computationally efficient.<sup>14</sup>

The effect of the singular value decomposition is to project the high-dimensional gene expression data into a lower dimensional subspace. By plugging (2) into (1), we obtain the following model:

$$\frac{P(Y_i = r)}{P(Y_i = 0)} = \gamma_{r0}^T \mathbf{W}_i, \quad (3)$$

where  $\gamma_{r0}$  ( $r = 0, \dots, J-1$ ) is a  $n \times 1$  vector of regression coefficients and  $\mathbf{W}_i$  ( $i = 1, \dots, n$ ) is the  $i$ th column of the  $n \times n$  matrix  $\mathbf{W} \equiv \mathbf{D}\mathbf{F}$ . It can be shown that  $\beta_{r0}$  in (1) and  $\gamma_{r0}$  in (3) are linked by the following relationship:  $\gamma_{r0} = \mathbf{U}^T\beta_{r0}$ .

By transforming the regression model from (1) into (3), we have reduced the dimension of the space for the predictor variables from  $p$  to  $n$ . This makes the problem computationally tractable, i.e. model (3) can be fit using standard statistical estimation procedures. We use maximum likelihood estimation to estimate  $\gamma_{r0}$  ( $r = 0, \dots, J - 1$ ). Ultimately, we are interested in performing inference about the components of  $\beta_{r0}$  ( $r = 0, 1, \dots, J - 1$ ). We can compute their estimates using the estimated parameters in model (3) and the inverse relationship  $\hat{\beta}_r = \mathbf{U}^T \hat{\gamma}_r$ . The  $\hat{\beta}_r$  ( $r = 0, \dots, J - 1$ ) can then be ranked in order of their estimated absolute magnitude to generate a list of genes that can discriminate between any two of the  $J - 1$  classes. In addition, the variance-covariance matrix of the  $\hat{\beta}_r$  ( $r = 0, \dots, J - 1$ ) can be standardized to yield a correlation matrix, which can then be used as an input in a hierarchical clustering algorithm.<sup>4</sup> The clustering algorithm attempts to find relationships between these discriminating genes and is based on the assumption that mutual coexpression potentially implies a common regulatory mechanism or that the genes might be involved in the same biological pathway.

There are several advantages to using models (1) and (3). First, we are able to jointly model the gene effects on tumor types, thereby incorporating correlations between genes. Univariate statistical methods, on the other hand, ignore this correlation. Second, by using a multicategorical response, we can model the correlations that exist between various tumor types. Third, if our ultimate interest is in doing clustering to determine if there are relationships that exist among the genes, this procedure allows us to incorporate external information (here, tumor type) into the clustering.

## 2.2 Initial variable selection

Typically in microarray experiments, the number of potential predictor genes will be on the order of thousands. For sufficiently large numbers of genes, the singular value decomposition (2) becomes computationally infeasible. We therefore utilize an initial preprocessing of the dataset in order to filter out a subset of the original set of genes. We fit an analysis of variance (ANOVA) model of gene expression measurement versus tumor class individually for each gene. For each ANOVA model, we calculate an overall F-statistic; this yields a set of  $p$  F-statistics. We then take the  $M$  genes with the largest F-statistics as the potential predictor variables in the model. The effect of this variable selection is to eliminate genes whose power in discriminating between tumor types is not significantly above their experimental variability in the gene expression measurements. Heuristically, we are selecting genes that have a relatively high signal-to-noise ratio. An empirical study of the effect of  $M$  on the perfor-

mance on the singular value decomposition regression modelling is given in the application to the breast cancer data.

### *2.3 Choosing number of principal components*

A major issue in the application of singular value decomposition regression modelling to high-dimensional data is determining how many principal components to use in model (3). There are many ways of performing this variable selection.<sup>11</sup> We have employed leave-one-out cross-validation.<sup>15</sup> In this procedure, one sample is removed from the dataset at a time. For a fixed number of principal components, say  $k$ , the regression model is fit to the remaining data. Based on the estimated model, the model is used to predict the class of the left out sample. An error measure is then calculated based on Hamming distance. We repeat this training procedure, leaving out each of the other samples from the dataset one at a time; summing the error measure over the samples yields an estimate of the classification error rate. We do this for every possible value of  $k$  and choose the value of  $k$  that yields the smallest classification error rate. Leave-one-out cross-validation is a popular method in situations with small samples where no test data are available.

It is also possible to combine the procedures in the two previous paragraphs. In this instance, one sample is held out of the experiment. We then perform the initial variable selection based on the ANOVA F-statistics using the remaining data, followed by a singular value decomposition regression model for each possible value of  $k$ . We then use the estimated model to predict the class for the held out sample and compute an error score using Hamming distance. In this leave-one-out cross-validation procedure, we take into account the variability in the initial selection of genes.

## **3 Application**

In this section, we apply the proposed methodology to data from a study of BRCA1- and BRCA2-positive tumors.<sup>13</sup> In this study, 22 biopsy specimens of primary breast tumors were collected. Seven had BRCA1 germ-line mutations, and eight had BRCA2 germ-line mutations. In addition, another seven samples were collected that had neither BRCA1 nor BRCA2 germ-line mutations; these were treated as sporadic cases of breast cancer. The goal of the study was to determine if there were differences in global gene expression profiles that could be used to discriminate the three classes of cancer (BRCA1, BRCA2 and sporadic).

While we will not go into the details of the analysis performed by Hedenfalk

et al., we do wish to make two points. First, they mostly used univariate statistical methods in order to determine the ability of genes to discriminate between the tumor types. Second, they actually subdivided the analysis of the data into two subgroup analyses. The first subgroup comparison was between BRCA1-positive and sporadic tumors; the second involved comparing BRCA2-positive and sporadic tumors. Given the scientific goal of the study, it seems more natural to jointly consider the three tumor classes simultaneously as opposed to doing the subgroup analyses.

We first focus on the performance of the principal components regression modelling in terms of the classification error rate. In this study, we also focus on the effects of  $M$  and variable selection on the predictive performance of the models. The results are summarized in Figures 1 and 2. The difference between the two graphs is that in Figure 2, the variable selection procedure is performed in the leave-one-out cross-validation, while it was not done in Figure 1. There are several points to note from these plots. Based on Figure 1, we can obtain low misclassification rates with a sufficient number of principal components. For example, with using  $M = 25$ , we have one misclassification using the singular value decomposition procedure with 11 principal components in the model. Comparable optimal misclassification rates can be obtained using  $M = 1500$  and  $M = 3226$ , although it does appear that for larger  $M$  the classification error rate tends to increase for most values of  $k$ . This suggests that models with fewer numbers of initial genes tend to have higher discriminative power. Comparing Figures 1 and 2 demonstrates that including the variable selection in the cross-validation procedure tends to worsen the predictive performance of the singular value regression models. This shows that variable selection for microarray experiments using the procedure described in Section 2.2 is an unstable procedure. The instability of the gene selection scheme here is a result of the fact that the magnitude of  $p$  is much larger than  $n$ . One potential way to improve the stability of the variable selection method of Section 2.2 is to use bagging methods.<sup>16</sup> We do not pursue that option here but leave it as an open question for future work.

The genes are now ranked using the singular value regression modelling. For this analysis, we take  $M = 100$  and focus on the comparison between BRCA1-positive tumors and sporadic tumors. The optimal number of principal components for  $M = 100$  where the variable selection is included in the cross-validation step is  $k = 2$ . We fit model (3) using two principal components. Based on fitting the model and the appropriate back-transformation, we can rank the genes in terms of their ability to discriminate between these three classes of tumors. A ranking of the top 20 genes from the subset of  $M = 100$ , based on their estimated absolute effects, is given in Table 1. Many of the

genes on this list overlap with the discriminatory genes found by Hedenfalk et al.

Finally, we wish to examine potential relationships between the genes in Table 1. One way to do this would be to simply cluster the genes using average linkage hierarchical clustering.<sup>4</sup> This yields the dendrogram in Figure 3. If we now use the estimated variance matrix from the SVD regression model based on two principal components as the basis of the hierarchical clustering, this yields the dendrogram in Figure 4. In particular, we find that there is better separation of two clusters with the second dendrogram. However, there is a loss of structure for larger numbers of clusters in Figure 4 relative to Figure 3. This is because the estimated regression coefficients from the singular value decomposition model are highly correlated. Both dendrograms offer complementary information about the relationships between the genes.

Further details of the analysis can be found at the following URL:  
<http://www.sph.umich.edu/~ghoshd/COMPBIO/SVD>.

#### 4 Discussion

In this article, we have developed a singular value decomposition regression modelling strategy for correlating gene expression profiles with tumor class in microarray settings. This methodology is important for determining the diagnostic and predictive ability of microarray technology in clinical settings. While we have focused mainly on a categorical response (tumor type), the ideas in this article can be applied to other types of clinical phenotypes, such as censored failure times, using different regression models in lieu of (1).

The major advantage of this regression modelling approach is that it provides a unified framework for accomplishing three goals in microarray experiments. First, predictive models for discrimination between tumor classes based on gene expression profiles are developed. Second, candidate genes can be selected using singular value decomposition regression modelling for follow-up experiments. Finally, clustering can be performed on the selected genes or on their associated covariance matrix in order to determine relationships that might exist between the genes.

As was mentioned in the Introduction, singular value decomposition regression models have been applied in other disciplines; one unique challenge here is that the outcome measure is not continuous. A major advantage of this method is that it can accommodate the scenario where the number of predictors is larger than the number of independent samples. However, other predictive modelling methods exist in this setting, such as partial least squares and ridge regression.<sup>11</sup> It would be very useful to compare these methods in

terms of their predictive modelling capabilities and is a current area of focus of our research.

For the SVD regression modelling procedure, a major question involves choosing the number of components to use in the regression model. While we used leave-one-out cross-validation, an alternative strategy would be to use a Bayesian model selection techniques. This would entail placing a prior on the number of principal components and then combining it with the likelihood given by (3) in order to calculate a posterior distribution for the number of principal components.

## References

1. Golub T. R. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
2. A. A. Alizadeh *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.
3. M. Bittner *et al.* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**: 536–540.
4. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–14868.
5. J. Kononen *et al.* (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine* **4**, 844–847.
6. P. J. Park, M. Pagano and M. Bonetti. A nonparametric scoring algorithm for identifying informative genes from microarray data. In PSB, 2000.
7. V. G. Tusher, R. Tibshirani and G. Chu (2001) Significance analysis of microarrays applied to ionizing response. *Proceedings of the National Academy of Sciences* **98**, 5116–5121.
8. S. Raychaudhuri, J. M. Stuart and R. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. In PSB, 2000.
9. N. S. Holter *et al.* (2000). *Proceedings of the National Academy of Sciences* **97**, 8409–8414.
10. O. Alter, P. O. Brown and D. Botstein (2000). *Proceedings of the National Academy of Sciences* **97**, 10101–10106.
11. I. E. Frank and J. H. Friedman (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics* **35**, 109–135.
12. A. Agresti. *Categorical Data Analysis*. (1990). New York: John Wiley

and Sons.

13. I. Hedenfalk *et al.* (2001) Gene expression profiles in hereditary breast cancer. *NEJM***244**, 539–548.
14. G. H. Golub and C. F. van Loan. *Matrix Computations*. (1996). Baltimore: John Hopkins University Press.
15. M. Stone (1974). Cross-validation choice and assessment of statistical predictions. *J. R. Statist. Soc. Ser. B* **36**, 111–147.
16. L. Breiman (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

Figure 1: Cross-validation estimated classification error rates versus number of principal components. Solid line:  $M = 25$ ; dashed line,  $M = 1500$ ; dotted line,  $M = 3226$ . Variable selection was not performed in the cross-validation procedure.

Figure 2: Cross-validation estimated classification error rates versus number of principal components. Solid line:  $M = 25$ ; dashed line,  $M = 1500$ ; dotted line,  $M = 3226$ . Variable selection was performed in the cross-validation procedure.

Figure 3: Dendrogram based on average linkage hierarchical clustering of genes in Table 1. Dissimilarity matrix based on correlation between gene expression measurements.

Figure 4: Dendrogram based on average linkage hierarchical clustering of genes in Table 1. Dissimilarity matrix based on correlation between estimated regression coefficients from singular value decomposition regression model.

Table 1: List of Ranked Genes for Discriminating BRCA1-positive tumors from sporadic breast cancer tumors.

Clone	Gene
823775	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3
364840	ESTs, Moderately similar to mouse Dhml protein [M.musculus]
44180	alpha-2-macroglobulin
32231	KIAA0246 protein
81518	apelin; peptide ligand for APJ receptor
417124	APEX nuclease (multifunctional DNA repair enzyme)
839594	ribosomal protein L38
239958	DKFZP586G1822 protein
234150	myotubularin related protein 4
73531	nitrogen fixation cluster-like
204897	phospholipase C, gamma 2 (phosphatidylinositol-specific)
725860	transcription factor AP-2 gamma (activating enhancer-binding protein 2 gamma)
246524	CHK1 (checkpoint, S.pombe) homolog
429135	suppression of tumorigenicity 13 (colon carcinoma) (Hsp70-interacting protein)
307843	ESTs
22230	collagen, type V, alpha 1
50413	armadillo repeat gene deletes in velocardiofacial syndrome
81331	fatty acid binding protein 5 (psoriasis-associated)
341130	retinoblastoma-like 2 (p130)
810551	low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor)