Prediction by supervised principal components

Eric Bair*

Trevor Hastie†

Debashis Paul‡

and Robert Tibshirani§

September 15, 2004

Summary

In regression problems where the number of predictors greatly exceeds the number of observations, conventional regression techniques may produce unsatisfactory results. We describe a technique called supervised principal components that can be applied to this type of problem. Supervised principal components is similar to conventional principal components analysis except that it uses a subset of the predictors that are selected based on their association with the outcome. Supervised principal components can be applied to regression and generalized regression problems such as survival analysis. It compares favorably to other techniques for this type of problem, and can also account for the effects of other covariates and help identify which predictor variables are most important. We also provide asymptotic consistency results to help support our empirical findings. These methods could become important tools for DNA microarray data, where they may be used to more accurately diagnose and treat cancer.

*Dept. of Statistics, Stanford Univ., CA 94305. ebair@stat.stanford.edu

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

‡Depts. of Statistics Stanford Univ., CA 94305. debashis@stat.stanford.edu

§Depts. of Health, Research & Policy, and Statistics, Stanford Univ, tibs@stat.stanford.edu

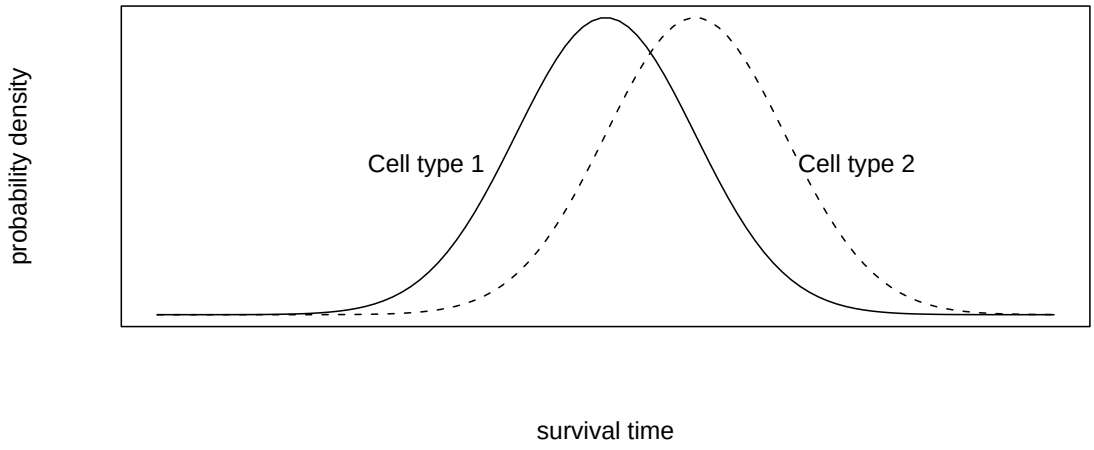

Figure 1: *Underlying conceptual model: there are two cell types and patients with the good cell type live longer on the average. However there is considerable overlap in the two sets of survival times. Hence it could be advantageous, to try to uncover the cell types and use these to predict survival time, rather than to predict survival time directly.*

1. Introduction

In this paper we study a method for predicting an outcome variable Y from a set of predictor variables $X_1, X_2, \ldots X_p$, measured on each of N individuals. In the typical scenario that we have in mind, the number of measurements p is much larger than N. In the example that motivated our work, $X_1, X_2, \ldots X_p$ are gene expression measurements from DNA microarrays. The outcome Y might be a quantitative variable, that we might assume to be normally distributed. More commonly in microarray studies, Y is a survival time, subject to censoring.

One approach to this kind of problem would be a "fully supervised" method. For example we could use a form of regression applicable when $p > N$; partial least squares (Wold 1975) would be one reasonable choice, as would be ridge regression(Hoerl and Kennard 1970). However Figure 1 illustrates why a more semi-supervised approach may be more effective.

We imagine that there are two cell types, and patients with the good cell type live longer on the average. However there is considerable overlap in the two sets of survival times. We might think of

survival time as a "noisy surrogate" for cell type. A fully supervised approach would give the most weight to those genes having the strongest relationship with survival. These genes are partially, but not perfectly, related to cell type. If we could instead discover the underlying cell types of the patients, often reflected by a sizable signature of genes acting together in pathways, then we would do a better job of predicting patient survival.

Now we can extract information about important cell types from both the relationship between Y and $X_1, X_2, \ldots X_p$, and the correlation among the predictors themselves. Principal components analysis is a standard method for modeling correlation. Regression on the first few principal components would seem like a natural approach, but this might not always work well. The fictitious data in Figure 2 illustrates the problem (if we were we to use only the largest principal component). It is a heatmap display with each gene represented by a row, and each column gives the data from one patient on one microarray. Gene expression is coded from blue (low) to yellow (high). In this example, the largest variation is seen in the genes marked A, with the second set of 10 patients having higher expression in these genes than the first 10. The set of genes marked B show different variation, with the 2nd and fourth blocks of patients having higher expression in these genes. The remainder of the genes show no systematic variation. At the bottom of the display, the red points are the first two singular vectors u_1, u_2 (principal components) of the matrix of expression values. In microarray studies, these are sometimes called "eigengenes" (Alter et al. 2000). (The broken lines represent the "true" grouping mechanism that generated the data in the two groups). Now if the genes in A are strongly related to the outcome Y, then Y will be highly correlated with the first principal component. In this instance we would expect that a model that uses u_1 to predict Y will be very effective. However the variation in genes A might reflect some biological process that is unrelated to the outcome Y. In that case, Y might be more highly correlated with u_2 or some higher order principal component.

The "supervised principal component" technique that we describe in this paper is designed to uncover such structure automatically. This technique was described in a biological setting in Bair and Tibshirani (2004), in the context of a related method known as "supervised clustering". The supervised principal component idea is simple: rather than perform principal component analysis using all of the genes in a data set, we use only those genes with the strongest estimated correlation

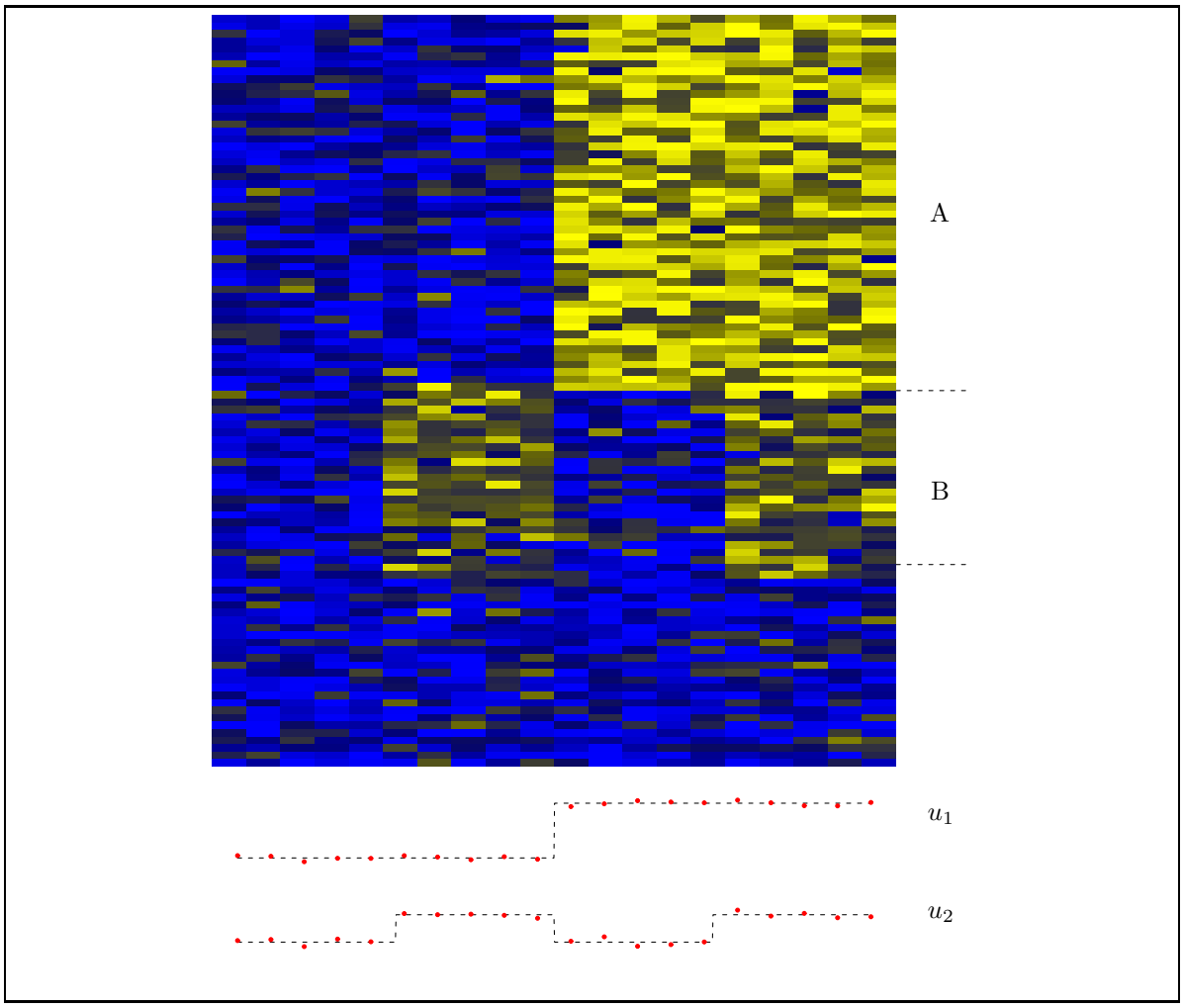

Figure 2: *Fictitious microarray data for illustration. A heatmap display with each gene represented by a row, and each column giving the data from one patient on one microarray. Gene expression is coded from blue (low) to yellow (high). The largest variation is seen in the genes marked A, with the second set of 10 patients having higher expression in these genes. The set of genes marked B show different variation, with the 2nd and fourth blocks of patients having higher expression in these genes. At the bottom of the display are shown the first two singular vectors (principal components) of the matrix of expression values (red points), and the actual grouping generators for the data (dashed lines). If the outcome is highly correlated with either principal component, the supervised principal component technique will discover this.*

with Y. In the scenario of Figure 2, if Y were highly correlated with the second principal component u_2, the genes in block B would have the highest correlation with Y. Hence we would compute the first principal component using just these genes, and this would yield u_2.

As this example shows, the use of principal components helps to uncover groups of genes that express together. Biologically, one or more cellular processes, accompanied by their cadre of expressing genes, determine the survival outcome. This same model underlies other approaches to supervised learning in microarray studies, including supervised gene shaving (Hastie et al. 2000) and tree harvesting (Hastie, Tibshirani, Botstein and Brown 2001). The supervised principal component procedure can be viewed as a simple way to identify the clusters of relevant predictors by (a) selection based on scores to remove the irrelevant sources of variation, and b) application of principal components to identify the groups of co-expressing genes.

In the next section we define the supervised principal components procedure. Section 3 shows an example from a lymphoma study, while in section 4 we show how the procedure can be generalized to allow adjustment for covariates. Section 5 describes alternative approaches to semi-supervised prediction, including "gene shaving", and in section 6 we present a simulation study comparing the various methods. In section 7 we summarize the results of supervised principal components on some survival studies. Section 8 shows that the standard principal components regression is not consistent as the sample size and number of features grow, while supervised principal components is consistent under appropriate assumptions. We conclude with some discussion in section 8.

2. Supervised principal components

2·1. Description

We assume there are p features measured on N observations (e.g. patients). Let $\mathbf{X}$ be an N times p matrix of feature measurements (e.g. genes), and y the N-vector of outcome measurements. We assume that the outcome is a quantitative variable; below we discuss other types of outcomes such as censored survival times. Here in a nutshell is the supervised principal component proposal:

Supervised principal components

1. Compute (univariate) standard regression coefficients for each feature
2. Form a reduced data matrix consisting of only those features whose univariate coefficient exceeds a threshold θ in absolute value (θ is estimated by cross-validation)
3. Compute the first (or first few) principal components of the reduced data matrix
4. Use these principal component(s) in a regression model to predict the outcome

We now give details of the method. Assume that the columns of $\mathbf{X}$ (variables) have been centered to have mean zero. Write the singular value decomposition of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{1}$$

where $\mathbf{U}, \mathbf{D}, \mathbf{V}$ are $N \times m$, $m \times m$ and $m \times p$ respectively, and $m = \min(N-1, p)$ is the rank of $\mathbf{X}$. $\mathbf{D}$ is a diagonal matrix containing the singular values d_j; the columns of $\mathbf{U}$ are the principal components $u_1, u_2, \ldots u_m$; these are assumed to be ordered so that $d_1 \geq d_2 \geq \ldots d_m \geq 0$.

Let s be the p-vector of standardized regression coefficients for measuring the univariate effect of each gene separately on y:

$$s_j = \frac{x_j^T y}{||x_j||}, \tag{2}$$

with $||x_j|| = \sqrt{x_j^T x_j}$. Actually, a scale estimate $\hat{\sigma}$ is missing in each of the s_j, but since it is common to all, we can omit it. Let C_θ be the collection of indices such that $|s_j| > \theta$. We denote by $\mathbf{X}_\theta$ the matrix consisting of the columns of $\mathbf{X}$ corresponding to C_θ. The SVD of $\mathbf{X}_\theta$ is

$$\mathbf{X}_\theta = \mathbf{U}_\theta \mathbf{D}_\theta \mathbf{V}_\theta^T \tag{3}$$

Letting $\mathbf{U}_\theta = (u_{\theta,1}, u_{\theta,2}, \ldots u_{\theta,m})$, we call $u_{\theta,1}$ the first supervised principal component of $\mathbf{X}$, and so on. We now fit a univariate linear regression model with response y and predictor $u_{\theta,1}$,

$$\hat{y}^{\mathrm{spc},\theta} = \bar{y} + \hat{\gamma} \cdot u_{\theta,1}. \tag{4}$$

Note that since $u_{\theta,1}$ is a left singular vector of $\mathbf{X}_\theta$, it has mean zero and unit norm. Hence $\hat{\gamma} = u_{\theta,1}^T y$, and the intercept is $\bar{y}$, the mean of y (expanded here as a vector of such means).

We use cross-validation to estimate the best value of θ. In most examples in this paper we consider only the first supervised principal component; in the examples of section 7, we allow the possibility of using more than one component.

Note that from (3),

$$\begin{aligned} \mathbf{U}_\theta &= \mathbf{X}_\theta \mathbf{V}_\theta \mathbf{D}_\theta^{-1} \\ &= \mathbf{X}_\theta \mathbf{W}_\theta. \end{aligned} \tag{5}$$

So, for example, $u_{\theta,1}$ is a linear combination of the columns of $\mathbf{X}_\theta$: $u_{\theta,1} = \mathbf{X}_\theta w_{\theta,1}$. Hence our linear regression model estimate can be viewed as a restricted linear model estimate using *all* the predictors in $\mathbf{X}_\theta$:

$$\hat{y}^{\mathrm{spc},\theta} = \bar{y} + \hat{\gamma} \cdot \mathbf{X}_\theta w_{\theta,1} \tag{6}$$

$$= \bar{y} + \mathbf{X}_\theta \hat{\beta}_\theta, \tag{7}$$

where $\hat{\beta}_\theta = \hat{\gamma} w_{\theta,1}$. In fact, by padding $w_{\theta,1}$ with zeros (corresponding to the genes excluded by C_θ), our estimate is linear in all p genes.

Given a test feature vector x^*, we can make predictions from our regression model as follows:

1. We center each component of x^* using the means we derived on the training data: $x_j^* \leftarrow x_j^* - \bar{x}_j$.
2. $\hat{y}^* = \bar{y} + \hat{\gamma} \cdot x_\theta^{*T} w_{\theta,1} = \bar{y} + x_\theta^{*T} \hat{\beta}_\theta$,

where x_θ^* is the appropriate sub-vector of x^*.

In the case of uncorrelated predictors, it is easy to verify that the supervised principal components procedure has the desired behavior: it yields all predictors whose standardized univariate coefficients exceed θ in absolute value.

Our proposal is also applicable to generalized regression settings, for example survival data, classification problems, or data typically analyzed by a generalized linear model. In these cases we use a score statistic in place of the standardized regression coefficients in (2) and use a proportional hazards or appropriate generalized regression in (4). Let $\ell_j(\beta)$ be the log-likelihood or partial

likelihood relating the data for a single predictor X_j and the outcome y, and let $U_j(\beta_0) = d\ell/d\beta|_{\beta=\beta_0}$, $I_j(\beta_0) = -d^2\ell_j/d\beta^2|_{\beta=\beta_0}$. Then the score statistic for predictor j has the form

$$s_j = \frac{U_j(0)^2}{I_j(0)}. \tag{8}$$

Of course for the Gaussian log-likelihood, this quantity is equivalent to the standardized regression coefficient (2).

2·2. An underlying model

We now consider a model to support the supervised principal component method. Suppose we have a response variable Y which is related to an underlying latent variable U by a linear model

$$Y = \beta_0 + \beta_1 U + \varepsilon. \tag{9}$$

In addition, we have expression measurements on a set of genes X_j indexed by $j \in \mathcal{P}$, for which

$$X_j = \alpha_{0j} + \alpha_{1j} U + \epsilon_j, \quad j \in \mathcal{P}. \tag{10}$$

We also have many additional genes X_k, $k \notin \mathcal{P}$ which are independent of U. We can think of U as a discrete or continuous aspect of a cell type, which we do not measure directly. $\mathcal{P}$ represents a set of genes comprising a pathway or process associated with this cell type, and the X_j are noisy measurements of their gene expression. We would like to identify $\mathcal{P}$, estimate U and hence fit the prediction model (9). This is a special case of a latent structure model, or single-component factor analysis model (Mardia et al. 1979).

The supervised principle component algorithm (SPCA) can be seen as a method for fitting this model:

1. The screening step estimates the set $\mathcal{P}$ by $\hat{\mathcal{P}} = C_\theta$;
2. Given $\hat{\mathcal{P}}$, the SVD of $\mathbf{X}_\theta$ estimates U in (10) by the largest principal component $u_{\theta,1}$;
3. finally the regression fit (4) estimates (9).

Step (1) is natural, since under assumption Y is correlated with U, and hence through U, each of the X_j, $j \in \mathcal{P}$ is correlated with Y. Step (2) is natural if we assume the errors ϵ_j have

a Gaussian distribution, with the same variance. In this case the SVD provides the maximum likelihood estimates for the single factor model (Mardia et al. 1979). The regression in (3) is an obvious final step.

In fact, given $\mathcal{P}$, the model defined by (9) and (10) is a special structured case of an *errors-in-variables* model (Miller 1986, Huffel and Lemmerling 2002). One could set up a joint optimization criterion

$$\min_{\beta_0,\beta_1,\{\alpha_{0,j},\alpha_{1,j}\},u_1,\ldots,u_N} \frac{\sum_{i=1}^N (y_i-\beta_0-\beta_1 u_i)^2}{\sigma_Y^2} + \sum_{j\in\mathcal{P}} \frac{\sum_{i=1}^N (x_{ij}-\alpha_{0j}-\alpha_{1j}u_i)^2}{\sigma_X^2} \tag{11}$$

Then it is easy to show that (11) can be solved by an augmented and weighted SVD problem. In detail, we form the augmented data matrix

$$\mathbf{X}_a = (y : \mathbf{X}), \tag{12}$$

assign weight $\omega_1 = \sigma_X^2/\sigma_Y^2$ to the first column, and $\omega_j = 1$ to the rest. Then with

$$v_0 = \begin{pmatrix} \beta_0 \\ \alpha_{0j_1} \\ \vdots \\ \alpha_{0j_q} \end{pmatrix}, \quad v_1 = \begin{pmatrix} \beta_1 \\ \alpha_{1j_1} \\ \vdots \\ \alpha_{1j_q} \end{pmatrix}, \tag{13}$$

(with $q = |\mathcal{P}|$) the rank-1 *weighted* SVD $\mathbf{X}_a \approx 1v_0^T + uv_1^T$ fits model (11). While this approach might seem more principled than our two-step procedure, SPCA has a distinct advantage. $\hat{u}_{\theta,1} = \mathbf{X}_\theta w_{\theta,1}$, and hence can be defined for future x^* data and be used for predictions. In the errors-in-variables approach, $\hat{u}_{EV} = \mathbf{X}_A w_{EV}$, which involves y as well, and leaves no obvious estimate for future data. We return to this model in Section 5.

This latent-variable model can be extended easily to accommodate multiple components $U_1, \ldots, U_m$. One way is to assume

$$Y = \beta_0 + \sum_{m=1}^M \beta_m U_m + \varepsilon \tag{14}$$

$$X_j = \alpha_{0j} + \sum_{m=1}^M \alpha_{1jm} U_m + \epsilon_j, \quad j \in \mathcal{P}. \tag{15}$$

Fitting this model proceeds as before, except now we extract M rather one principal component from $\mathbf{X}_\theta$. We study this model more deeply in Section 8.

2·3. An example

The SPCA model anticipates other sources of variation in the data, unrelated to the response. In fact these sources can be even stronger than those driving the response, to the extent that principle components would identify them first. By guiding the principal components, SPCA extracts the desired components.

We simulated data from a scenario like that of Figure 2. We used 1000 genes and 40 samples, all with base error model being Gaussian with unit variance. We then defined the mean vectors μ_1 and μ_2 as follows. We divide the samples into consecutive blocks of 10, denoted by the sets (a, b, c, d). Then

$$\mu_{1i} = \begin{cases} -2 & \text{if } i \in a \cup b \\ +2 & \text{otherwise} \end{cases} \tag{16}$$

$$\mu_{2i} = \begin{cases} -1 & \text{if } i \in a \cup c \\ +1 & \text{otherwise} \end{cases} \tag{17}$$

The first 200 genes have the mean structure μ_1:

$$x_{ij} = \mu_{1i} + \epsilon_{ij},\ j = 1, \ldots, 200; \quad i = 1, \ldots, 40. \tag{18}$$

The next 50 genes had mean structure μ_2:

$$x_{ij} = \mu_{2i} + \epsilon_{ij},\ j = 201, \ldots, 250; \quad i = 1, \ldots, 40. \tag{19}$$

In all cases, $\epsilon_{ij} \sim N(0, 1)$, which is also how the remaining 750 genes are defined. Finally the outcome is generated as $y_i = \alpha \cdot \mu_{1_i} + (1 - \alpha) \cdot \mu_{2i} + \varepsilon_i$ where ε_i is $N(0, 1)$. The first two principal components of $\mathbf{X}$ are approximately μ_1 and μ_2 (see Figure 2).

We tried various values of $\alpha \in [0, 1]$ as shown in Figure 3. Plotted is the correlation of the supervised principal component predictor with an independent (test set) realization of y, as θ in the screening process $|s_j| > \theta$ is varied. The number of genes surviving the screening is shown on the horizontal axis. The extreme right end of each plot represents standard principal components regression. When $\alpha = 0$, so that the outcome is correlated with the 2nd principal component, supervised PC easily improves upon principal components regression. When α reaches 0.5, the advantage disappears, but supervised PC does no worse than principal components regression.

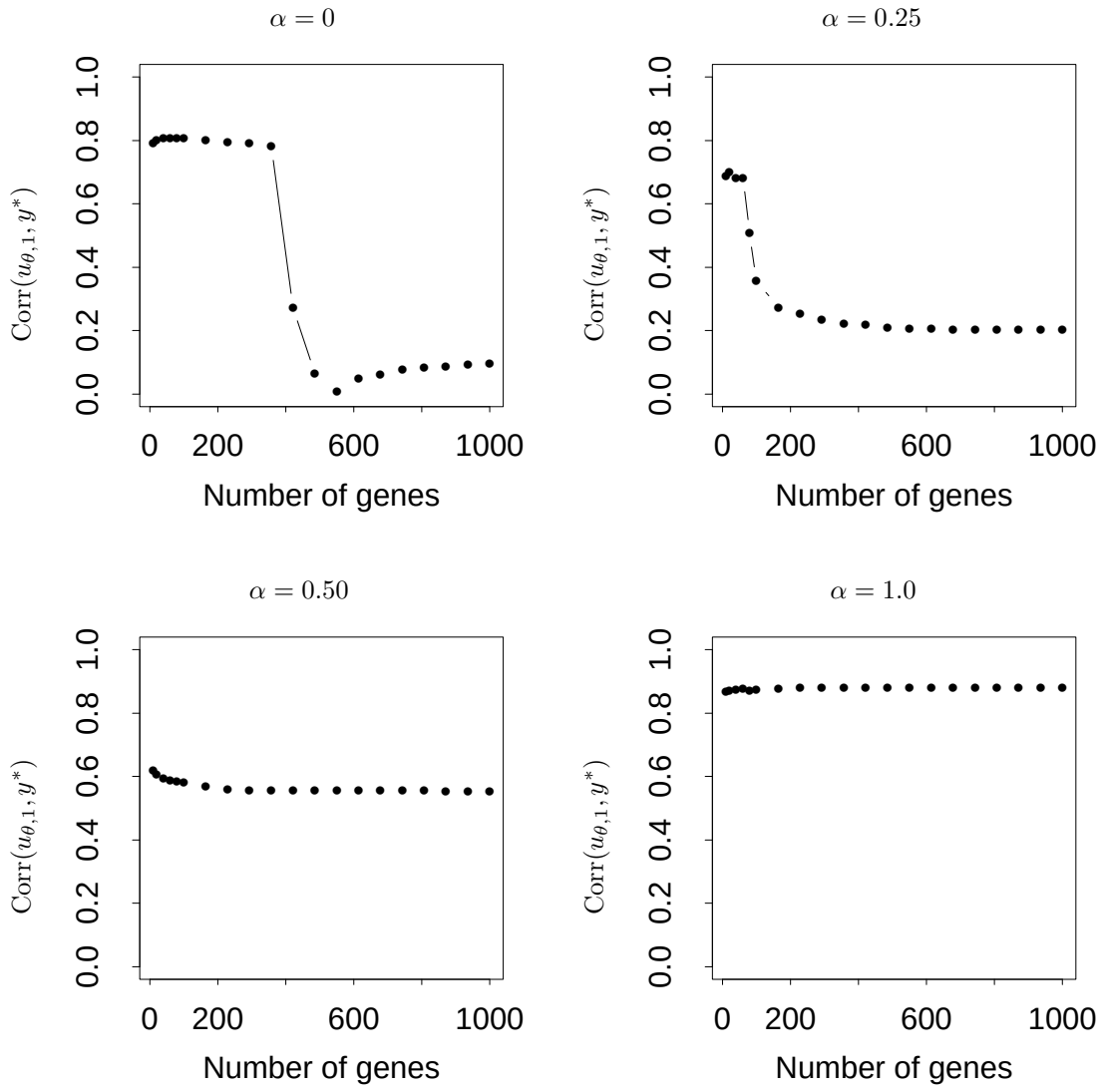

Figure 3: *Correlation between the first supervised principal component $u_{\theta,1}$ and a test outcome y, as the weight α given to the first principal component in the data generation is varied. The number of genes used by the procedure is shown on the horizontal axis in each panel. The sharp switch in the first two panels corresponds to the point at which the order of the principal components is reversed.*

2·4. *Importance scores and a reduced predictor*

Having derived the predictor $u_{\theta,1}$, how do we assess the contributions of the p individual features? It is not true that the features that passed the screen $|s_j| > \theta$ are necessarily important or are the only important features. Instead, we compute the *importance score* as the correlation between each feature and $u_{\theta,1}$:

$$\text{imp}_j = \text{cor}(x_j, u_{\theta,1}) \tag{20}$$

Features j with large values of $|\text{imp}_j|$ contribute most to the prediction of y.

Typically all p genes will have non-zero importance scores. We can take this idea further, and look for a reduced predictor that performs as well as $u_{\theta,1}$. We define

$$\hat{u}_{\text{red}} = \sum s(\text{imp}_j, \gamma) \cdot x_j \tag{21}$$

where $s(x,t)$ is the soft-threshold function $\text{sign}(x)(|x| - t)_+$, $+$ indicating positive part. With $\gamma = 0$ most or all features will get non-zero weight; for larger values of γ, features with lower values of imp_j get zero weight. We illustrate this idea in the next section.

The ability of supervised principal components to build a model based on only a small number of inputs is very important for practical applications. For example, a predictor that requires expression measurements for a few thousand genes is not likely to be useful in a everyday clinical settings: microarrays are too expensive and complex for everyday use, and simpler assays like RT-PCR can only measure 50 or 100 genes at a time. In addition, isolation of a smaller gene set could aid in biological understanding of the disease.

3. Example: survival of lymphoma patients

This data is taken from Rosenwald et al. (2002), consisting of 240 samples from patients with diffuse large B-cell lymphoma (DLBCL), with gene expression measurements for 7399 genes. The outcome was survival time, either observed or right censored. We randomly divided the samples into a training set of size 160 and a test set of size 80. The results of various procedures are shown in Table 1. We used the genes with top 25 Cox scores (cutoff of 3.53) in computing the first supervised principal component. Although PLS (described in Section 5) provides a strong predictor of survival,

Method	Z-score	P-value
1st principal component	-1.04	0.2940
Partial least squares	2.49	0.0130
1st supervised principal component (25 genes)	-2.92	0.0023

Table 1: *Lymphoma data: Test sets results for the various methods*

the supervised principal component is even stronger. This performance is typical, as shown in the studies of Section 6 and 7.

The left panel of Figure 4 shows the importance scores for each gene, plotted against the raw Cox score. The two sets of scores are correlated, but are far from identical. The right panel shows the average *test set* Cox score for the top k genes, when ranked on the training set by either supervised principal components or Cox score. Remarkably, genes ranked by the first supervised principal component exhibit higher test set Cox scores, than those obtained using the Cox score itself to do the ranking.

Note that methods like "significance analysis of microarrays" (Tusher et al. 2001*b*) use the Cox score as the basis for determining the genes that are strongly related to survival time. Figure 4 suggests that loading of each gene on the first supervised principal component might provide a better measure of significance than the Cox score.

Figure 5 shows the training set Cox score for the reduced predictor (21). We see that the best predictions can be obtained using as few as about 50 genes. The test set p-value from this reduced model is about the same as the raw model in the first line of Table 1. Figure 6 shows the top 50 genes and their loadings. Details are given in the figure caption.

4. Adjustment for covariates

Typically there may be covariates measured on each of the cases, and it might be of interest to adjust for these. For example in gene expression survival studies, in addition to the predictors $X_1, X_2, \ldots X_p$, we might have available covariates $z = (z_1, z_2, \ldots z_k)$ such as tumor stage and tumor type. There might be interest in finding gene expression predictors that work independently of stage

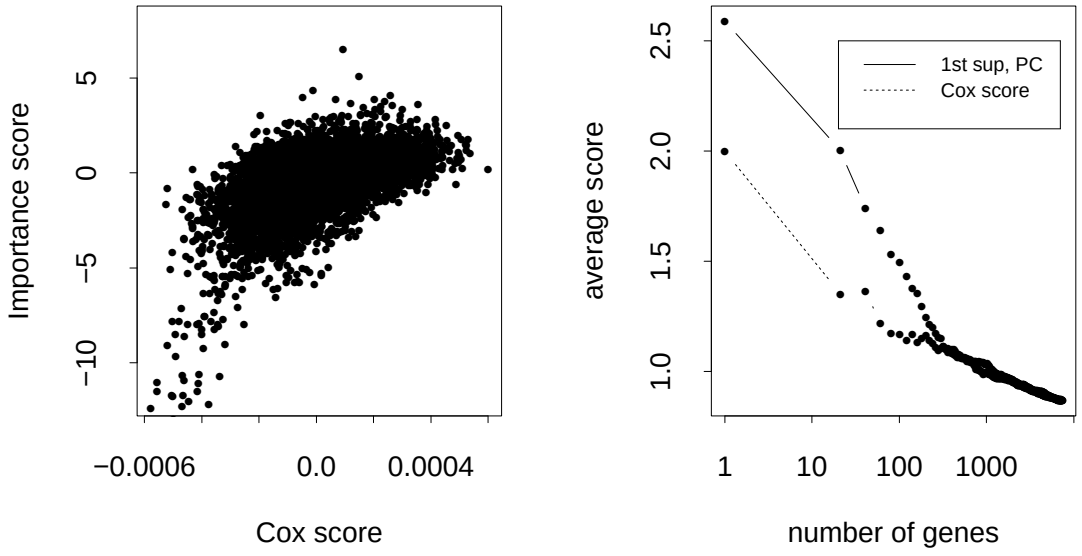

Figure 4: *Lymphoma data: left panel shows the importance scores for each gene, plotted against the raw Cox score. The two sets of scores are correlated, but are far from identical. The right panel shows the average* test set *Cox score for the top k genes, when ranked on the training set by either supervised principal components or Cox score.*

and tumor; that is, having adjusted for these factors, the gene expression predictor is still strongly related to survival.

The supervised principal component procedure can be easily generalized to incorporate covariate adjustment. Obviously we want to include the covariates z in the final regression (4). However we can potentially do better by adjusting the predictor scores to account for the covariates z. In the case of the linear regression model, these adjusted scores are the usual t-statistics associated with the coefficient for x_j in a joint linear-model fit of y on z and x_j.

For likelihood models, we use the Rao score statistics, which generalize these (and are the square of the t-statistics above). In detail, we denote by $\ell_j(\beta_1, \beta_2)$. the log-likelihood for the model relating the outcome to predictor j and z, and $U_j(\beta_1, \beta_2)$ and $I_j(\beta_1, \beta_2)$ the corresponding gradient and information matrix. Here parameters β_1 and β_2 relate to predictor j and z respectively. Then in

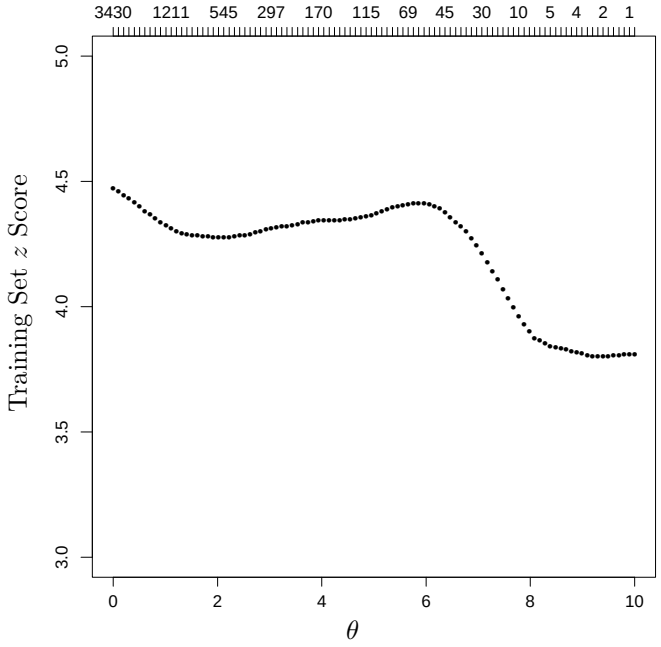

Figure 5: *Lymphoma data: training set Cox score for the reduced predictor (21), as the threshold θ is varied. The corresponding number of genes used in the reduced predictor is shown along the top of the figure.*

place of the log-likelihood score (8) we use the adjusted score

$$s'_j = U_j(0, \hat{\beta}_2)^T I_j(0, \hat{\beta}_2) U_j(0, \hat{\beta}_2) \tag{22}$$

where $\hat{\beta}_2$ is the maximum likelihood estimator of β_2. Large values of s'_j correspond to predictors that are strongly related to the outcome, having adjusted for z. We then use these scores to define an adjusted first principal component score.

Figure 7 shows the result of applying this procedure to the DLBCL data, with z chosen to be the International Prognostic Index (IPI), a standard clinical risk score, graded as "low", "medium" or "high". In the left column we have stratified the (unadjusted) first principal component score into two groups, by cutting at its median. The survival curves in each IPI group are shown. The right column displays the survival curves using the adjusted first principal component. The gap between the pairs of curves is similar for the low and medium IPI groups, but is clearly larger for the high IPI group using the adjusted score. These curves are computed on the training set and so could be the result of over-fitting. In the test set there were very few patients in the high IPI group and we

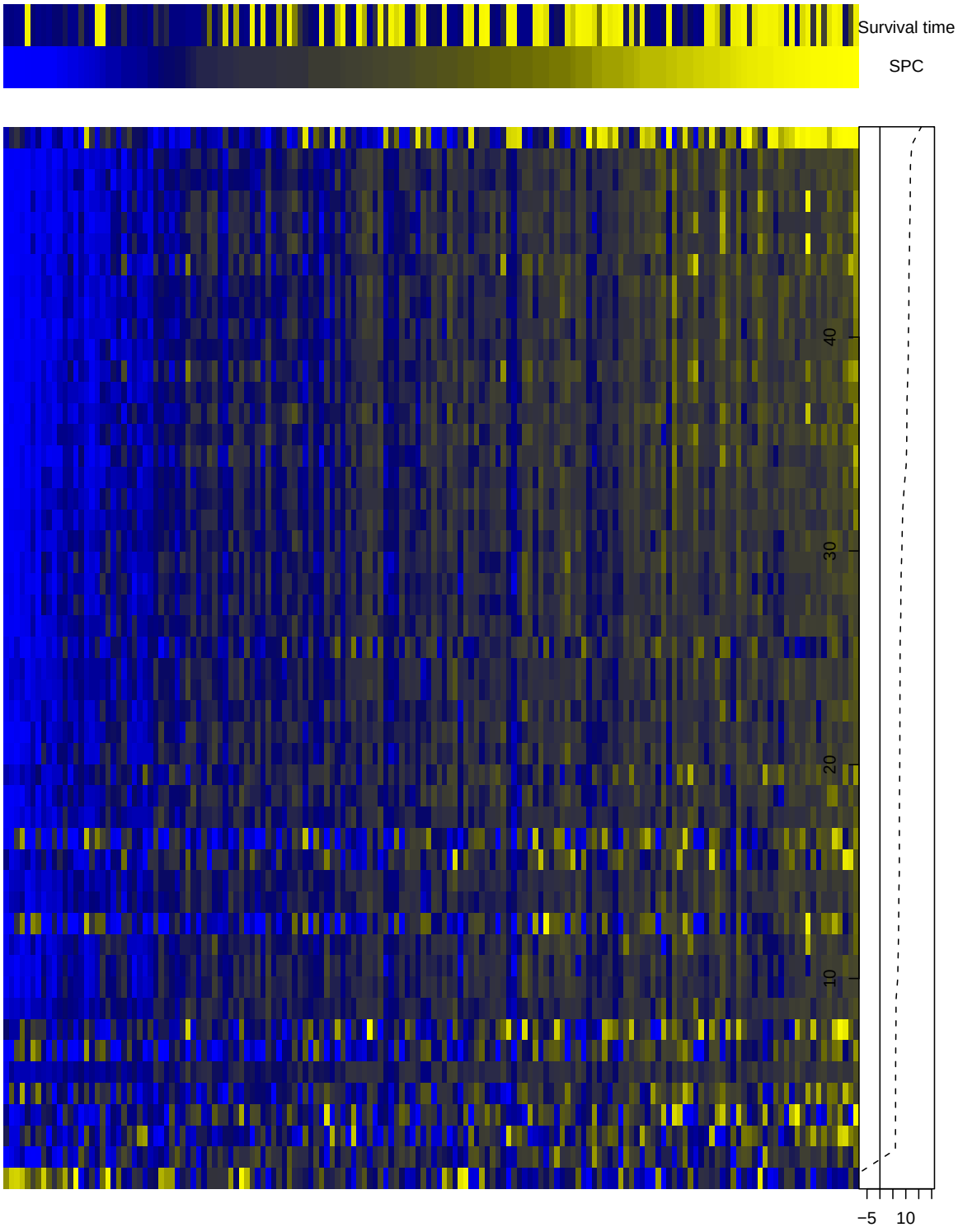

Figure 6: *Lymphoma data: heatmap display of the top 50 genes. The top two rows of the figure show the observed survival times and first supervised principal component (SPC) $u_{\theta,1}$; for survival times T censored at time c, we show $\hat{\mathrm{E}}(T|T \geq c)$ based on the Kaplan-Meier estimator. All columns have been sorted by increasing value of $u_{\theta,1}$. On the right of the heatmap the "loadings" $w_{\theta,1}$ are shown (see (6)); the genes (rows) are sorted by decreasing value of their loading. All genes but the last one have positive loadings.*

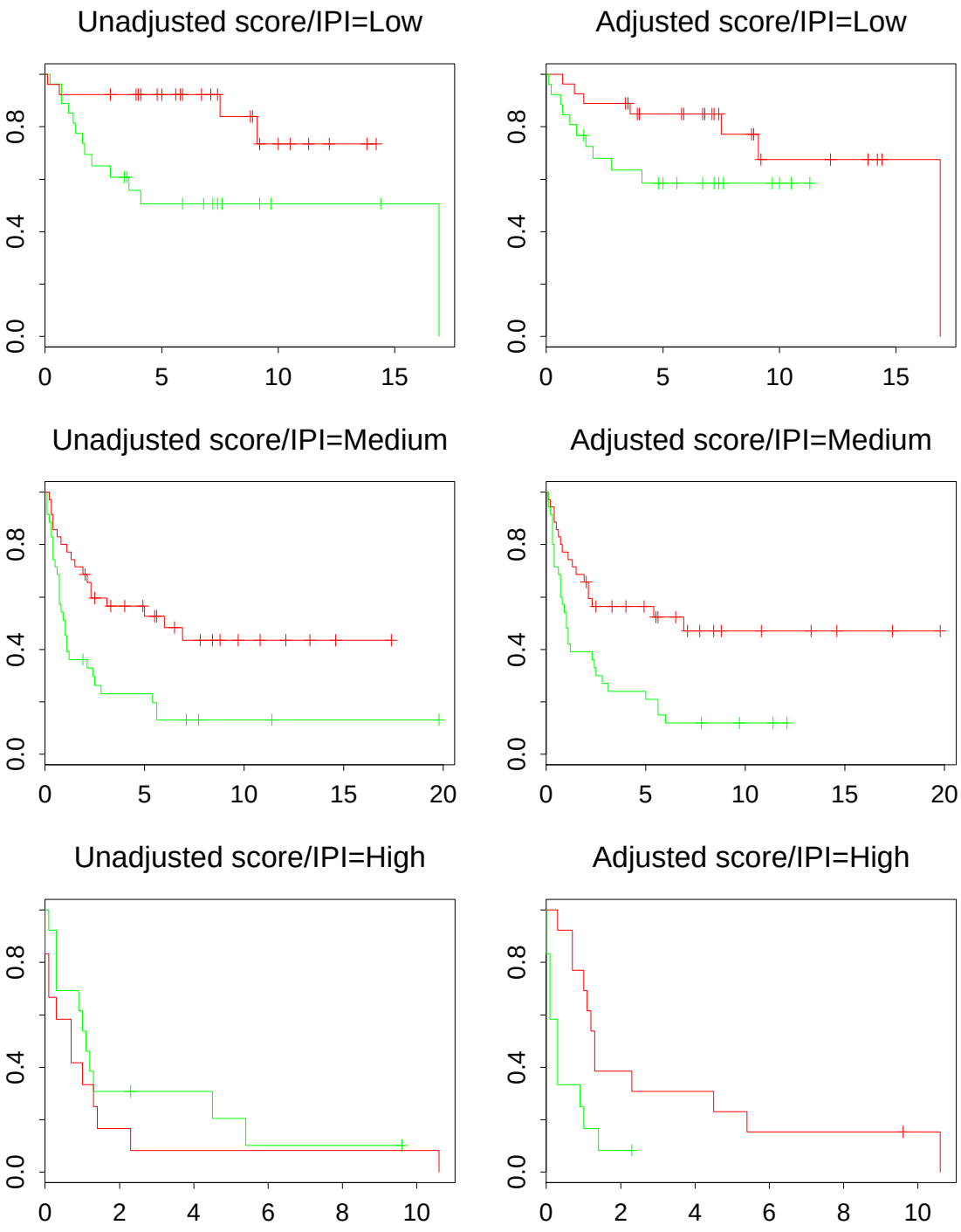

Figure 7: *Adjusting for IPI in the lymphoma data. In the left column we have stratified the (unadjusted) first principal component score into two groups, by cutting at its median. The survival curves in each IPI group are shown. The right column displays the survival curves using the adjusted first principal component.*

found no real differences between the adjusted and unadjusted analysis.

5. Some alternative approaches

In this section we discuss some alternative approaches to this problem: some classical, and some reflecting other approaches we have explored.

Ridge regression

Ridge regression (Hoerl and Kennard 1970) is a classical regression procedure when there are many correlated predictors, and one that could reasonably be applied in the present setting. Ridge regression fits the full linear regression model, but manages the large number of predictors in these genomic settings by regularization (Hastie and Tibshirani 2003). Ridge regression solves

$$\min_{\beta} ||y - \beta_0 - \mathbf{X}\beta||^2 + \lambda||\beta||^2, \tag{23}$$

where the second term shrinks the coefficients toward zero. The regularization parameter λ controls the amount of shrinkage, and for even the smallest $\lambda > 0$, the solution is defined and is unique. It can also be shown that this form of regularizations shrinks the coefficients of strongly correlated predictors toward each other, an attractive property in this setting.

Using the singular value representation (1), the fitted values from a ridge regression have the form

$$\begin{aligned} \hat{y}^{RR} &= \bar{y} + \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}y \\ &= \bar{y} + \sum_{j=1}^{m} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y. \end{aligned} \tag{24}$$

Ridge regression is like a smooth version of principal components regression; rather than retaining the first k principal components and discarding the rest, it weights the successive components by a factor that decreases with decreasing eigenvalue d_j^2. Note that ridge regression is a linear method, i.e $\hat{y}^{RR}$ is a linear function of y; in contrast, SPCA is non-linear, because of the initial gene-selection step.

We now consider several approaches to supervised principal components that modify the optimization criterion behind PCA in a supervisory fashion.

Partial least squares

Partial Least Squares (PLS) is one such approach, with a long history (Wold 1975, Frank and Friedman 1993, Hastie, Tibshirani and Friedman 2001). PLS works as follows

1. Standardize each of the variables to have zero mean and unit norm, and compute the univariate regression coefficients $w = \mathbf{X}^T y$.

2. define $u_{\rm PLS} = \mathbf{X}w$, and use it in a linear regression model with y.

Although PLS goes on to find subsequent orthogonal components, one is sufficient for our purposes here. PLS explicitly uses y in estimating its latent variable. Interestingly, it can be shown that the (normalized) w in PLS solves (Frank and Friedman 1993)

$$\max_{||w||=1} \mathrm{Corr}^2(y, \mathbf{X}w)\mathrm{Var}(\mathbf{X}w), \tag{25}$$

a compromise between regression and PCA. We include PLS among the competitors in our comparisons in the next sections.

Mixed variance-covariance criterion

The largest principal component is that normalized linear combination $z = \mathbf{X}v$ of the genes with the largest sample variance. Another way to supervise this would be to seek a linear combination $z = \mathbf{X}v$ having both large variance and a large (squared) covariance with y, leading to the compromise criterion

$$\max_{||v||=1} (1-\alpha)\mathrm{Var}(z) + \alpha\mathrm{Cov}(z, y)^2 \quad \text{s.t. } z = \mathbf{X}v. \tag{26}$$

This is equivalent to

$$\max_{||v||=1} (1-\alpha)v^T\mathbf{X}^T\mathbf{X}v + \alpha v^T\mathbf{X}^T y y^T\mathbf{X}v \tag{27}$$

If y is normalized to unit norm, then the second term in (27) is a regression sum of squares (regressing z on y), and has the interpretation "the variance of z explained by y". The solution v can be efficiently computed as the first right singular vector of the augmented $(N+1) \times p$ matrix

$$\mathbf{X}_a = \begin{pmatrix} (1-\alpha)^{\frac{1}{2}}\mathbf{X} \\ \alpha^{\frac{1}{2}} y^T\mathbf{X} \end{pmatrix} \tag{28}$$

By varying the mixing parameter α we control the amount of supervision. Although the mixed criterion can guide the sequence of eigenvectors, all genes have non-zero loadings which adds a lot of variance to the solution.

Supervised gene shaving

Hastie et al. (2000) proposed "gene shaving" as a method for clustering genes. The primary focus of their method was to find small clusters of highly correlated genes, whose average exhibited strong variance over the samples. They achieved this through an iterative procedure, which repeatedly computed the largest principal component of a subset of the genes, but after each iteration "shaved" away a fraction of the genes with small loadings. This produces a sequence of nested subsets of gene clusters, with successively stronger pairwise correlation and variance of the largest principal component.

They also proposed a supervised version of gene shaving, which uses precisely a mixed criterion of the form (27). Although this method has two tuning parameters, α and the subset size, here we fix α to the intermediate value of 0.5 and focus attention on the subset size. As in SPCA, for each subset the largest principal component is used to represent its genes.

This method is similar in flavor SPCA; it produces principal components of subset of genes, where the choice of subset is supervised. Simultaneously searching for sparse components with high variance and correlation with y is an attempt to omit features that might slip through the SPCA screening step. Our experiments in the next section show that shaving can exhibit very similar performance to SPCA, the latter having the advantage of being simpler to define, and only one tuning parameter to select.

Another mixed criterion

The largest normalized principal component u_1 is the largest eigenvector of $\mathbf{X}\mathbf{X}^T$. This follows easily from the SVD (1) and hence $\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$. Intuitively, since

$$u_1^T \mathbf{X}\mathbf{X}^T u_1 = \sum_{j=1}^{p} \langle u_1, x_j \rangle^2, \tag{29}$$

we are seeking the vector u_1 closest on average to each of the x_j. A natural supervised modification is to perturb this criterion in a manner that encourages the leading eigenvector to align with y:

$$\max_{u_1, ||u_1||=1} (1-\alpha) \sum_{j=1}^{p} \langle u_1, x_j \rangle^2 + \alpha \langle u_1, y \rangle^2 \tag{30}$$

Solving (30) amounts to finding the largest eigenvector of

$$C(y; \alpha) = (1-\alpha)\mathbf{X}\mathbf{X}^T + \alpha y y^T. \tag{31}$$

Equivalently, one could form an augmented matrix $\mathbf{X}_a$ with y in the $(p+1)$st column. If we assign weights α to this row and $(1-\alpha)$ to first p rows, then a weighted singular value decomposition of $\mathbf{X}_a$ is equivalent to an eigen-decomposition of (30). We note that this is exactly the situation described in the errors-in-variables model (11)–(13) in Section 2·2. As mentioned there, the estimate u_1 involves y as well as the x_j, so cannot be used directly with test data. This approach was not pursued further.

Discussion of methods

Figure 8 illustrates the methods discussed above on a simulation example with $N = 100$ samples and $p = 5000$ features. The data are generated according to the latent-variable model (34), where there are 4 dominant principal components, and the one associated with the response is ranked number 3 (when estimated from the data). The methods are identified in the figure caption. The leftmost M point corresponds to principal component regression using the largest PC. SPCA and shaving do much better than the other methods.

Figure 9 gives us a clue to what is going on. Shown are the first 1000 of 5000 feature loadings for two of the methods demonstrated in Figure 8 (chosen at the best solution points). Both methods correctly identified the important component (the one related to y involving the first 50 features). In a regular SVD of $\mathbf{X}$ this important component was dominated by two other components. In detail, the training data from model (34) has four *built-in* components, with singular values computed to be 99.9, 88.3, 80.9 and 80.5 respectively. Empirically, we verified that component three is identified with the response mechanism, but its singular value is just above the noise level (the fifth singular value was 79.2). However, the mixed criterion also brings with it noisy coefficients, somewhat smaller,

for ALL the other variables, while SPCA sets most of the other loadings to zero. The coefficients for shaving show a very similar pattern to SPCA, while those for ridge and PLS are very similar to the mixed criterion, and are not shown here.

Our experience on many similar examples is much the same, although the shaving method occasionally gets the wrong component completely. SPCA tends to be more reliable, is simpler to define, and hence our method of choice. The simulations in the next section support this choice as well.

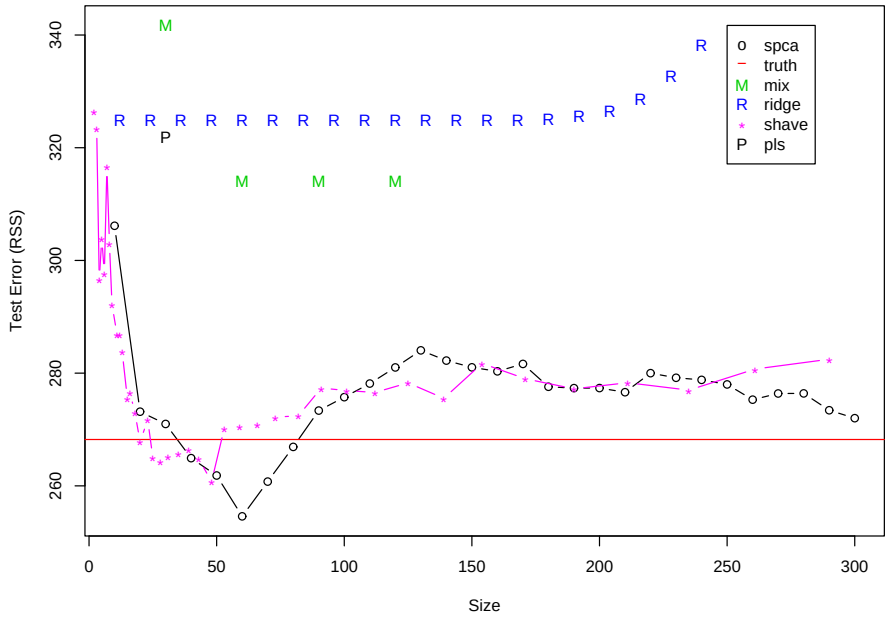

Figure 8: *One simulation example illustrating typical behavior of the different methods. The data are generated according to the model (34) described in the next section, with* $N = 100$ *and* $p = 5000$*. Ridge regression, PLS, and the mixed criterion all suffer from the very high dimensions. Although not shown, the regularization parameter* λ *for the ridge points increases to the right, as does the* α *for the mixed criterion, the leftmost value being 0. Both shaving and SPCA are indexed by subset size. The line labeled "truth" uses the known linear combination of 50 features as the regression predictor.*

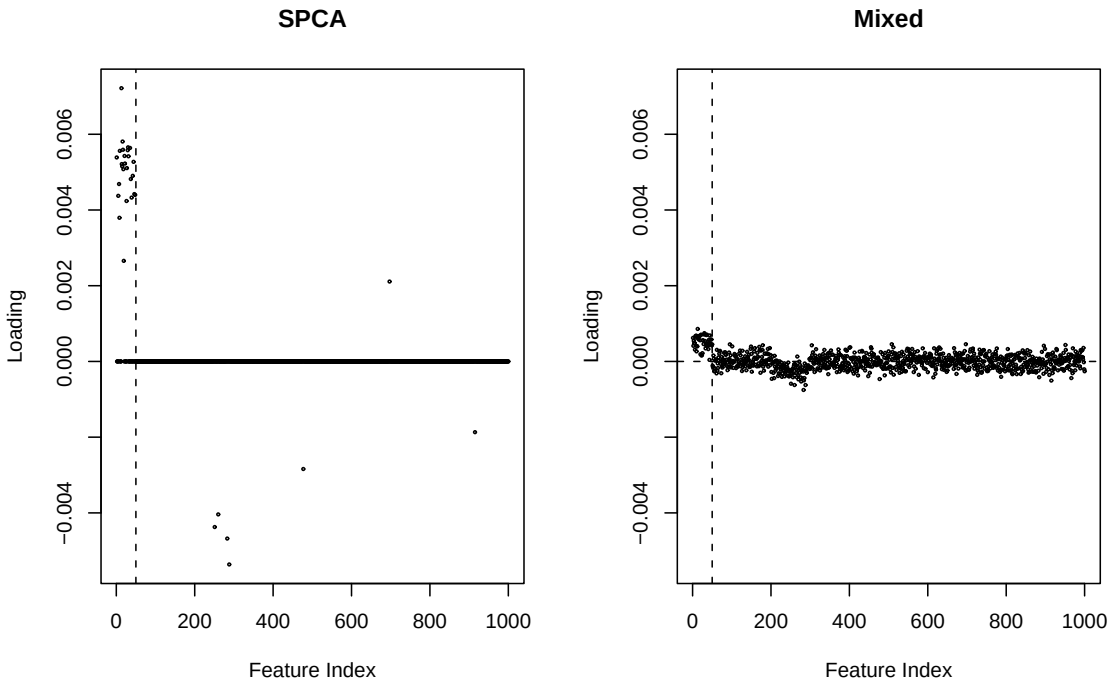

Figure 9: *Feature loadings w for SPCA (left plot), and the mixed criterion (27) (right plot). The first 1000 of 5000 are shown, at the "best" solution point. The vertical line indicates that the first 50 variables generated the response. While both these methods were able to overwhelm the first two dominant principal components (which were unrelated to y), SPCA is able to ignore the majority of variables, while the mixed criterion gives them all weight, albeit more weight to the first 50.*

6. Simulation studies

We performed two simulation studies to compare the performance of the methods that we have considered. Each simulated data set $\mathbf{X}$ consisted of 5000 "genes" (rows) and 100 "patients" (columns). Let x_{ij} denote the "expression level" of the ith gene and jth patient. In the first study we generated the data as follows:

$$x_{ij} = \begin{cases} 3 + \epsilon_{ij} & \text{if } i \leq 50,\ j \leq 50 \\ 4 + \epsilon_{ij} & \text{if } i \leq 50,\ j > 50 \\ 3.5 + \epsilon_{ij} & \text{if } i > 51 \end{cases} \tag{32}$$

where the ϵ_{ij} are independent normal random variables with mean 0 and variance 1. We also let

$$y_j = \frac{\sum_{i=1}^{50} x_{ij}}{25} + \epsilon_j \tag{33}$$

where the ϵ_j's are independent normal random variables with mean 0 and standard deviation 1.5.

We designed this simulation so that there are two "tumor subclasses." Patients 1 through 50 belong to "class 1," and patients 51 through 100 belong to "class 2." The first 50 genes have slightly lower average expression levels in the patients with tumor class 1. Furthermore, since y is proportional to the sum of the expression level of the first 50 genes, y is slightly lower for patients with tumor class 1. The other 4950 genes are unrelated to y.

We applied seven methods to this simulated data set: (1) principal components regression, (2) principal components regression using only the first principal component, (3) partial least squares (one direction), (4) ridge regression (see e.g. Hastie, Tibshirani and Friedman (2001), chapter 3), (5) supervised principal components, (6) mixd variance-covariance, and (7) gene shaving. We trained each of these models using a simulated data set generated as we described above. We select the optimal value of the tuning parameters for each method using 10-fold cross-validation. Then we used the same procedure to generate an independent test data set and used the models we built to predict y on the test data set. We repeated this procedure 10 times and averaged the results. Table 2 shows the errors produced by each model.

We see that gene shaving and supervised principal components produced smaller cross-validation and test errors than any of the other methods, with the former holding a small edge. Principal com-

Method	CV Error	Test Error
PCR	290.5 (10.18)	227.0 (7.36)
PCR-1	315.3 (12.43)	252.1 (8.76)
PLS	284.9 (10.04)	219.8 (7.43)
Ridge regression	291.3 (10.94)	226.3 (7.89)
Supervised PC	242.0 (10.32)	184.6 (7.36)
Mixed var-cov.	282.5 (11.07)	221.6 (7.24)
Gene shaving	219.6 (8.34)	163.0 (4.34)

Table 2: *Results of the simulation study based on the "easy" simulated data. Each entry in the table represents the squared error of the the test set predictions averaged over 10 simulations. The standard error of each error estimate is in parentheses. The prediction methods are: (1) principal components regression, (2) principal components regression restricted to using only one principal component, (3) partial least squares, (4) ridge regression, (5) supervised principal components, (6) mixed variance-covariance, and and (7) gene shaving.*

ponents regression and partial least squares gave comparable results (although principal components regression performed slightly worse when restricted to one component).

Next, we generated a "harder" simulated data set. In this simulation, we generated each x_{ij} as follows:

$$x_{ij} = \begin{cases} 3 + \epsilon_{ij} & \text{if } i \leq 50,\ j \leq 50 \\ 4 + \epsilon_{ij} & \text{if } i \leq 50,\ j > 50 \\ 3.5 + 1.5 \cdot I(u_{1j} < 0.4) + \epsilon_{ij} & \text{if } 51 \leq i \leq 100 \\ 3.5 + 0.5 \cdot I(u_{2j} < 0.7) + \epsilon_{ij} & \text{if } 101 \leq i \leq 200 \\ 3.5 - 1.5 \cdot I(u_{3j} < 0.3) + \epsilon_{ij} & \text{if } 201 \leq i \leq 300 \\ 3.5 + \epsilon_{ij} & \text{if } i > 301 \end{cases} \tag{34}$$

Here the u_{ij} are uniform random variables on $(0, 1)$ and $I(x)$ is an indicator function. For example, for each of the genes 51–100, a single value u_{1j} is generated for sample j; if this value is larger than 0.4, then all the genes in that block get 1.5 added. The motivation for this simulation is that there

Method	CV Error	Test Error
PCR	301.3 (14.47)	303.3 (8.30)
PCR-1	318.7 (14.90)	329.9 (10.79)
PLS	308.3 (15.12)	300.0 (7.22)
Ridge regression	312.5 (14.50)	303.5 (8.30)
Supervised PC	231.3 (11.07)	255.8 (6.68)
Mixed var-cov.	301.6 (14.05)	303.1 (5.53)
Gene shaving	228.9 (12.02)	258.5 (11.35)

Table 3: *Results of the simulation study based on the "hard" simulated data. Each entry in the table represents the squared error of the the test set predictions averaged over 10 simulations. The standard error of each error estimate is in parentheses. The prediction methods are the same as in the previous table.*

are other clusters of genes with similar expression patterns that are unrelated to y. This is likely to be the case in real microarray data, since there are pathways of genes (that probably have similar expression patterns) that are not related to y. Figures 8 and 9 illustrate some of the methods applied to a realization from this model.

We repeated the experiment described above using (34) to generate the data sets instead of (32). The results are given in Table 3. Most of the methods performed worse in this "harder" experiment. Once again gene shaving and supervised principal components produced smaller errors than any of the competing methods; gene shaving shows much more variability than supervised principal components in this case.

7. Application to various survival studies

Here we compare several methods for performing survival analysis on real DNA microarray data sets.[Some of these results are also reported in Bair and Tibshirani (2004)]. We applied the methods to four different data sets. First, we examined a microarray data set consisting of diffuse large B-cell lymphoma (DLBCL) patients (Rosenwald et al. 2002). There are 7399 genes, 160 training patients and 80 test patients in this data set. Second, we considered a breast cancer data set (van 't Veer

et al. 2002). There were 4751 genes and 97 patients in this data set. We partitioned this data set into a training set of 44 patients and a test set of 53 patients.

Next, we examined a lung cancer data set (Beer et al. 2002). There were 7129 genes and 86 patients, which we partitioned into a training set of 43 patients and a test set of 43 patients. Finally, we considered a data set of acute myeloid leukemia (AML) patients (Bullinger et al. 2004). It consisted of 6283 genes and 116 patients. This data set was partitioned into a training set of 59 patients and a test set of 53 patients.

In addition to supervised principal components, we examined the following methods: principal components regression, partial least squares, and two other methods that we call "median cut," and "clustering-Cox", described in (Bair and Tibshirani 2004). Both of these methods turn the problem into a two-class classification problem and then apply the nearest shrunken centroid classifier of Tibshirani et al. (2001). The median cut method stratifies the patients into high or low risk, depending on whether they survived past the median survival time. The "clustering-Cox" method is like supervised principal components, using 2-means clustering applied to the genes with highest Cox scores.

For methods (3), (4) and (5), we allowed the possibility of using more than component, and chose this number by cross-validation. The results are shown in Table 4. Supervised principal components again performs better than the competing methods.

8. Consistency of supervised principal components

In this section we show that the standard principal components regression is not consistent as the sample size and number of features grow, while supervised principal components is consistent under appropriate assumptions.

8·1. Setup

Suppose that the rows of $\mathbf{X}$ are independent and identically distributed. Then one can formulate a population model as follows. Denoting the rows by X_i^T $(i = 1, ..., N)$ we have the model :

$$X_i \overset{i.i.d.}{\sim} N_p(\mu, \Sigma)$$

	(a) DLBCL			(b) Breast Cancer		
Method	R^2	p-val	NC	R^2	p-val	NC
(1) Median Cut	0.05	0.047		0.13	0.0042	
(2) Clustering-Cox	0.08	0.006		0.21	0.0001	
(3) SPCA	0.11	0.003	2	0.27	2.1×10^{-5}	1
(4) PC Regression	0.01	0.024	2	0.22	0.0003	3
(5) PLS	0.10	0.004	3	0.18	0.0003	1
	(c) Lung Cancer			(d) AML		
Method	R^2	p-val	NC	R^2	p-val	NC
(1) Median Cut	0.15	0.0016		0.07	0.0487	
(2) Clustering-Cox	0.07	0.0499		0.08	0.0309	
(3) SPCA	0.36	1.5×10^{-7}	3	0.16	0.0013	3
(4) PC Regression	0.11	0.0156	1	0.08	0.0376	1
(5) PLS	0.18	0.0044	1	0.07	0.0489	1

Table 4: *Comparison of the different methods on four different datasets from cancer studies. The methods are (1) Assigning samples to a "low-risk" or "high-risk" group based on their median survival time. (2) Using 2-means clustering based on the genes with the largest Cox scores. (3) Using the supervised principal components method. (4) Using principal components regression. (5) Using partial least squares regression. Table lists the R^2 (proportion of log-likelihood explained) and p-values for the test set predictions as well as the number of components used.*

where Σ $(p \times p)$ is the covariance matrix. Without loss of generality we shall assume $\mu = 0$, since it can be quite accurately estimated from data.

Suppose $\mathbf{X}$ is partitioned as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ is $N \times p_1$ and $\mathbf{X}_2$ is $N \times p_2$ with $p_1 + p_2 = p$. We assume that the corresponding partition of Σ is given by

$$\Sigma = \begin{bmatrix} \Sigma_1 & O \\ O & \Sigma_2 \end{bmatrix} \tag{35}$$

Suppose further that we can represent Σ_1 $(p_1 \times p_1)$ as

$$\Sigma_1 = \sum_{k=1}^{M} \lambda_k \theta_k \theta_k^T + \sigma^2 I \tag{36}$$

where θ_k $(k = 1, ..., M)$ are mutually orthonormal eigenvectors and the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_M > 0$. $\sigma^2 > 0$ represents the contribution of (isotropic) "background noise" that is unrelated to the interactions among genes. This model can be described as a covariance model for gene expressions that is an M-rank perturbation of identity. Here $1 \leq M \leq p_1 - 1$.

Our assumption is that $\mathbf{X}_1$ is the matrix containing all the columns whose variations are related to the variations in y. First we assume that the selection procedure is such that it selects $\mathbf{X}_1$ with probability 1. In section 8·5 we consider the more realistic scenario in which we estimate this subspace from the data. Our key assumptions regarding the matrix Σ_1 are the following.

A1 The eigenvalues of Σ_1 satisfy the *identifiability condition* :

$$\lambda_1 > \ldots > \lambda_M > 0$$

and M is a fixed positive integer.

A2 $p_1 \to \infty$ as N increases to infinity in such a way that $\frac{p_1}{N} \to 0$.

Note that p_1 could as well be fixed, but if not, our result still holds provided that **A2** holds. We may also let σ^2 and λ_k's vary with N. However, then we need to replace conditions **A1** and **A2** by

A1' The eigenvalues are such that $\frac{\lambda_k}{\lambda_1} \to \rho_k$ for $k = 1, \ldots, M$ with $1 = \rho_1 > \rho_2 > \ldots > \rho_M > 0$ and $\lambda_1 \to c > 0$ as $N \to \infty$. Moreover, $\sigma^2 \to \sigma_0^2 \in [0, \infty)$ as $N \to \infty$.

A2' p_1 varies with N in such a way that $\frac{\sigma^2 p_1 (\log N)^2}{N \lambda_1} \to 0$ as $N \to \infty$.

8·2. The underlying regression model

In this setting we can denote the rows of $\mathbf{X}_1$ by X_i^T, $i = 1, \ldots, N$ and express them as

$$X_i = \sum_{k=1}^{M} \eta_{ik} \sqrt{\lambda_k} \theta_k + \sigma w_i, \qquad i = 1, \ldots, N \tag{37}$$

where η_{ik} are i.i.d. $N(0,1)$ random variables, w_i are i.i.d. $N_p(0, I)$ random vectors and η's and w's are independent. Let η_k denote the vector $(\eta_{1k}, \ldots, \eta_{Nk})^T$.

Observe that this model and the model for response (described in (39) below) are closely related to the latent variables model considered in Section 2.2 (cf. equations (9), (9), (14) and (15). However for identifiability purposes we now require that η_k's be uncorrelated, which was not necessary for the analogous variables U_m's in eqn. (15). In essence our results are stronger than we might need in actual practice: we prove the consistency of not just the latent variance subspace but also that of its separate components.

Remark : Notice that conditions **A1'** and **A2'** make allowance for the possibility that $\frac{\lambda_1}{\sigma^2} \to \infty$ and $\frac{p_1}{N}$ converges to a positive limit. This particular fact becomes relevant when we try to connect to the scenario that we describe now. Suppose $M = 1$, and as before let $\mathcal{P}$ be the genes forming the columns of matrix $\mathbf{X}_1$. Then $|\mathcal{P}| = p_1$, and

$$X_{ij} = \sqrt{\lambda_1} \theta_{j1} \eta_{i1} + \sigma w_{ij}, \qquad j \in \mathcal{P}$$

represent the expression for the genes in the set $\mathcal{P}$ of i-th array (replicate), $i = 1, \ldots, N$. (Compare with eqn. (10) in Section 2.2). In this case, if $\sqrt{\lambda_1}\theta_{j1}$ is of roughly the same magnitude for all $j \in \mathcal{P}$, then $\lambda_1 \sim p_1$ as $p_1 \to \infty$. Even otherwise, it is reasonable to think that the "signal-to-noise ratio" $\frac{\lambda_1}{\sigma^2}$ is going to ∞ as $p_1 \to \infty$, since the presence of larger number of genes associated with a common latent factor yields a greater amount of information.

Suppose the singular value decomposition of $\mathbf{X}_1$ is given by

$$\mathbf{X}_1 = UDV^T \qquad \text{where } U \text{ is } N \times m,\ D \text{ is } m \times m \text{ and } V \text{ is } p_1 \times m, \text{ with } m = \min(N, p_1). \tag{38}$$

Here N is the number of observations (patients) and p_1 is the dimension (number of genes). Let $u_1, \ldots, u_m$ denote the columns of U and $v_1, \ldots, v_m$ denote the columns of V. For obvious reasons we set $\widehat{\theta}_k = v_k,\ k = 1, \ldots, M$. Also, we denote the diagonal elements of D by $d_1 > \ldots > d_m$.

The model for the outcome is:

$$y = \beta_0 \frac{1}{\sqrt{N}}\mathbf{1} + \sum_{k=1}^{K} \beta_k \frac{1}{\sqrt{N}}\eta_k + Z \tag{39}$$

where $K \leq M$, $\mathbf{1}$ is the vector with 1 in each coordinate, and $Z \sim N_N(0, \frac{\tau^2}{N} I)$ independent of $\mathbf{X}$ for some $\tau \in [0, \infty)$.

Remark : Note that we could as well have described the model in terms of similar quantities for the full data set, i.e. $\mathbf{X}$ (correspondingly Σ). There are two difficulties associated with this formulation. First, it is not at all likely that all the systematic variation in the gene expressions is associated with the variation in the response. So even if model (35) and (36) were true, there is no guarantee that the largest K eigenvalues of Σ are the largest K eigenvalues of Σ_1. This will result in spurious (i.e., unrelated to the response y) components being added to the model.

The second issue is to do with the accuracy of estimation. Since typically p is very large, in fact much larger than, or at least comparable to, the sample size N, it is almost never going to be the case that assumption **A2'** is satisfied (with p_1 replaced by p). But the assumption for p_1 is reasonable since only a few genes are expected to be associated with a certain type of disease. Violation of this condition results in an inconsistency in the estimates of θ_k. (Details in the next section). So the procedure of selecting the genes before performing the PCA regression is not only sensible, it is in effect necessary.

8·3. Results on estimation of θ_k and λ_k

In order to discuss consistency of the eigenvectors θ_k we consider the quantity $dist(\widehat{\theta}_k, \theta_k)$ where *dist* is a distance measure between two vectors on the p_1-dimensional unit sphere. One can either choose $dist(a, b) = \angle(a, b)$ or $dist(a, b) = \| a - sign(a^T b) \cdot b \|_2$ for $a, b \in \mathbb{S}^{p_1}$.

First suppose we perform PCA on the full data set $\mathbf{X}$ and estimate θ_k by $\widetilde{\theta}_k$, the restriction of the k-th right singular vector of $\mathbf{X}$ to the coordinates corresponding to the set $\mathbf{X}_1$. Then the following result asserts that if p is very large, then we may not have consistency.

Theorem 1 (Lu 2002), (Johnstone and Lu 2004): *Suppose* **A1** *holds (and assume that* σ^2 *and*

λ_k's are fixed) and $\frac{p}{N} \to \gamma \in (0, \infty)$ as $N \to \infty$. Then

$$dist(\widetilde{\theta}_k, \theta_k) \not\to 0 \qquad \text{in probability as } \; N \to \infty$$

i.e., the usual PCA based estimate of θ_k is inconsistent.

Under the same conditions as in *Theorem 1*, the sample eigenvalues are also inconsistent estimates for the populations eigenvalues. However, the behavior is rather complicated.

From now onwards we treat exclusively the singular value decomposition of $\mathbf{X}_1$. We denote the PCA-based estimate of the k-th largest eigenvalue of Σ_1 by $\widehat{\ell}_k$, $k = 1, 2, \ldots, m$. Observe that $\widehat{\ell}_k = \frac{1}{N} d_k^2$. The corresponding population quantity is $\ell_k := \lambda_k + \sigma^2$.

A natural estimator of λ_k is $\widehat{\lambda}_k = \max\{\widehat{\ell}_k - \sigma^2, 0\}$ if σ^2 is known. However, if σ^2 is unknown one can estimate this by various strategies. One approach is to use the median of the diagonal elements of $\frac{1}{N}\mathbf{X}_1^T\mathbf{X}_1$ as a (usually biased) estimate of σ^2 and then define $\widehat{\lambda}_k = \max\{\widehat{\ell}_k - \widehat{\sigma}^2, 0\}$.

Next we establish consistency for principal components analysis restricted to the matrix $\mathbf{X}_1$.

Theorem 2 (Paul 2004):

- *Suppose that conditions* **A1'** *and* **A2'** *hold. Then*

$$dist(\widehat{\theta}_k, \theta_k) = O_P\left(\sqrt{\frac{\sigma^2 p_1}{N\lambda_1}}\right) \qquad as \;\; N \to \infty$$

  If moreover, $\frac{\lambda_1}{\sigma^2} \to \infty$, then $\widehat{\ell}_k = \lambda_k(1 + o_P(1))$.

- *If σ^2 and λ_k's are fixed and* **A1** *and* **A2** *hold then*

$$dist(\widehat{\theta}_k, \theta_k) = O_P\left(\sqrt{\frac{p_1}{N}}\right) \qquad and \qquad \widehat{\ell}_k \xrightarrow{P} \ell_k = \lambda_k + \sigma^2 \qquad as \;\; N \to \infty$$

8·4. Estimation of β_k

In this section we discuss the estimation of the parameters β_k, $k = 1. \ldots, K$. To simplify matters we shall treat σ^2 and λ_k's to be fixed and assume that **A1** and **A2** hold.

Suppose either σ^2 is known or a consistent estimate $\widehat{\sigma}^2$ is available. Then define $\widehat{\lambda}_k = \max\{\widehat{\ell}_k - \sigma^2, 0\}$. Let u_k be as before and define $\widetilde{u}_k$ as $\frac{1}{\sqrt{\widehat{\lambda}_k}} \frac{1}{\sqrt{N}} \mathbf{X}_1 v_k$ if $\widehat{\lambda}_k > 0$, and as any fixed unit vector (say

$(1,0,\ldots,0)^T$) otherwise. Define an estimate of β_k (for $1 \le k \le K$) as $\widetilde{\beta}_k = \widetilde{u}_k^T y$. One can compare its performance with another estimate $\widehat{\beta}_k = u_k^T y$ with u_k as before. Also, define $\widehat{\beta}_0 = \widetilde{\beta}_0 = \frac{1}{\sqrt{N}} \sum_{j=1}^N y_j$.

Observe that

$$u_k = \frac{1}{d_k}\mathbf{X}_1 v_k = (\widehat{\ell}_k)^{-1/2}\frac{1}{\sqrt{N}}\mathbf{X}_1\widehat{\theta}_k = (\widehat{\ell}_k)^{-1/2}\left[\sum_{l=1}^M \sqrt{\lambda_l}(\theta_l^T\widehat{\theta}_k)\frac{1}{\sqrt{N}}\eta_l + \frac{\sigma}{\sqrt{N}}W\widehat{\theta}_k\right]$$

where W is the $N \times p_1$ matrix whose rows are w_i^T $(i = 1,\ldots,N)$. Then since $\widehat{\theta}_k = \theta_k + \varepsilon_k$ (as a convention assuming $\widehat{\theta}_k^T\theta_k > 0$) where $\| \varepsilon_k \|_2 = O_P(\sqrt{\frac{p_1}{N}})$,

$$u_k = \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_k+\sigma^2}}\frac{1}{\sqrt{N}}\eta_k(1+o_P(1)) + \frac{\sigma}{\sqrt{\lambda_k+\sigma^2}}\frac{1}{\sqrt{N}}W\theta_k(1+o_P(1)) + \delta_k \tag{40}$$

where $\| \delta_k \|_2 = O_P(\sqrt{\frac{p_1}{N}})$. To prove this last statement we only need to use *Theorem 2* together with the well-known fact that $\| \frac{1}{N}W^TW \|_2 = 1 + o_P(1)$ (for an easy proof see Paul (2004)), since

$$\begin{aligned} \| \frac{1}{\sqrt{N}}W\varepsilon_k \|_2^2 &\le \| \frac{1}{N}W^TW \|_2 \| \varepsilon_k \|_2^2 = O_P(\frac{p_1}{N}), \\ \text{and} \quad |\varepsilon_k^T\theta_l| &\le \| \varepsilon_k \|_2 \| \theta_l \|_2 = O_P(\sqrt{\frac{p_1}{N}}), \qquad \text{for } 1 \le l \ne k \le M, \end{aligned}$$

and finally, $\| \eta_l \|_2 = \sqrt{N}(1+o_P(1))$ for all $l = 1,\ldots,M$.

From this it follows that

$$\widetilde{u}_k = \frac{1}{\sqrt{N}}\eta_k(1+o(1)) + \frac{\sigma}{\sqrt{\lambda_k}}\frac{1}{\sqrt{N}}W\theta_k(1+o(1)) + \widetilde{\delta}_k \tag{41}$$

where $\| \widetilde{\delta}_k \|_2 = O_P(\sqrt{\frac{p_1}{N}})$. Note that the vectors $\{W\theta_k : k = 1,\ldots,M\}$ are independent $N_N(0,I)$ and independent of $\{\eta_k : k = 1,\ldots,M\}$ since θ_k's are mutually orthonormal.

To establish consistency of $\widetilde{\beta}_k$, $1 \le k \le K$ note that

$$\begin{aligned} \widetilde{\beta}_k &= \beta_0\frac{1}{\sqrt{N}}\widetilde{u}_k^T\mathbf{1} + \sum_{l=1}^K \beta_l\frac{1}{N}[(\eta_k + \frac{\sigma}{\sqrt{\lambda_k}}W\theta_k)(1+o_P(1)) + \sqrt{N}\widetilde{\delta}_k]^T\eta_l + \widetilde{u}_k^TZ \qquad \text{(by (41))} \\ &= \beta_0(O_P(\frac{1}{\sqrt{N}}) + o_P(1)) + \beta_k(1 + o_P(1) + \widetilde{\delta}_k^T\frac{1}{\sqrt{N}}\eta_k) \\ &\quad + \sum_{l\ne k}\beta_l(O_P(\frac{1}{\sqrt{N}}) + \widetilde{\delta}_k^T\frac{1}{\sqrt{N}}\eta_l) + O_P(\frac{1}{\sqrt{N}}) \\ &= \beta_k(1+o_P(1)) \end{aligned}$$

since $\frac{1}{N}\eta_k^T\eta_l = O_P(\frac{1}{\sqrt{N}})$ if $k \ne l$, and $\frac{1}{N}\eta_l^TW\theta_k = O_P(\frac{1}{\sqrt{N}})$ for all k,l (by independence), $\| \widetilde{\delta}_k \|_2 = o_P(1)$, and $\widetilde{u}_k^TZ = \| \widetilde{u}_k \|_2 \langle \frac{\widetilde{u}_k}{\|\widetilde{u}_k\|_2}, Z\rangle$. Note that the second term in the last product is a $N(0,\frac{\tau^2}{N})$ random variable and the first term is $\sqrt{\frac{\lambda_k+\sigma^2}{\lambda_k}}(1+o_P(1))$ by (41).

It is easy to verify that $\widehat{\beta}_0 = \beta_0(1 + o_P(1))$. However, from the above analysis it is clear that the estimator $\widehat{\beta}_k = u_k^T y$, for $1 \leq k \leq K$, is not consistent in general. In fact $\widehat{\beta}_k = \sqrt{\frac{\lambda_k}{\lambda_k + \sigma^2}}\beta_k(1 + o_P(1))$ when λ_k's and σ^2 are fixed. However, as we indicated in the remark following equation (37), it is reasonable to assume that $\frac{\lambda_1}{\sigma^2} \to \infty$ as $p_1, N \to \infty$. This will ensure (via the first part of **Theorem 2**) that the factor $\sqrt{\frac{\lambda_k}{\widehat{\ell}_k}} \to 1$ in probability as $N \to \infty$ when **A1'** and **A2'** hold. And therefore we shall have $\widehat{\beta}_k = \beta_k(1 + o_P(1))$ for $1 \leq k \leq K$. This in a way validates the claim that having more genes (larger p_1) associated with a common latent factor gives better predictability.

8·5. Consistency of the coordinate selection scheme

Until now we have been working under the assumption that we are capable of selecting $\mathbf{X}_1$ exactly (or a superset with few spurious coordinates) and therefore treating these coordinates as fixed. However, when one employs a coordinate selection scheme, in order that the theoretical analysis to go through, we have to assume that at least asymptotically the selection scheme described in the paper is "consistent" (we elaborate on it later), and that the selection is made on an independent sample so that there is no selection bias involved. This last requirement is desirable because otherwise one will come across difficulties arising due to complex dependency structures, and even the notion of consistency will becomes somewhat more complicated since the set of selected coordinates itself is a random quantity. In practice, at least for N moderately large, this is not a big problem because of the massive averaging that goes on in the PCA step.

The issue of consistency of a coordinate selection scheme is rather tricky. Let us first consider the case when $p_1 < \infty$ is fixed (that is does not change with growing N). In this case by consistency we mean being able to select the set of p_1 coordinates that determine (35) and discarding all the coordinates that are not a part of this subset.

Now suppose p_1 is growing with N. In this case Σ_1 (and Σ_2) are matrices of growing dimension satisfying (35) and (36) such that model (39) also holds. Now the issue becomes complicated because, since each eigenvector is of unit norm, inevitably certain coordinates will tend to zero as N and p_1 increase. This will prevent a coordinate selection sheme from selecting *all* the "significant" coordinates. In this context "consistency" of a selection scheme should be treated as equivalent to ensuring consistency of the estimators of first K eigenvectors and eigenvalues. This boils down to

being able to recover all the "big" coordinates (i.e., the coordinates with large value for at least one among $\{\sqrt{\lambda_k}\theta_k : k = 1, \ldots, K\}$) among the first p_1 (under appropriate "mild" restrictions on the vectors $\{\theta_k : k = 1, \ldots, K\}$). A proof of this nontrivial fact is provided in (Paul 2004).

Let us see what the selection scheme in this paper attempts to recover. Since the rows of $\mathbf{X}_2$ are independent $N_{p_2}(0, \Sigma_2)$ r.v. independent of $\mathbf{X}_1$, invoking (37), we can express $s = \mathbf{X}^T y$ in the following form

$$s = \begin{bmatrix} (\sum_{k=1}^{M} \sqrt{\lambda_k}\theta_k \eta_k^T + \sigma W^T)y \\ \Sigma_2^{1/2}\mathbf{C}y \end{bmatrix} \tag{42}$$

where $\mathbf{C}$ is a $p_2 \times N$ matrix whose entries are i.i.d. $N(0,1)$ independent of $\mathbf{X}_1$ and Z (and hence y). Observe that W^T is independent of y.

This shows that if we consider the j-th element of s for $1 \le j \le p_1$, then

$$\begin{aligned} \frac{1}{\sqrt{N}} s_j &= \frac{1}{N}(\sum_{k=1}^{M} \sqrt{\lambda_k}\theta_{jk}\eta_k^T)(\beta_0 \mathbf{1} + \sum_{k'=1}^{K} \beta_{k'}\eta_{k'} + \sqrt{N}Z) + \frac{\sigma}{\sqrt{N}}(W^T y)_j \\ &= \beta_0 \sum_{k=1}^{M} \sqrt{\lambda_k}\theta_{jk} O_P(\frac{1}{\sqrt{N}}) + \sum_{k=1}^{K} \beta_k \sqrt{\lambda_k}\theta_{jk}(1 + O_P(\frac{1}{\sqrt{N}})) \\ &\qquad + \sum_{k=1}^{M} \sqrt{\lambda_k}\theta_{jk} \sum_{k' \ne k}^{K} \beta_{k'} O_P(\frac{1}{\sqrt{N}}) + \sigma(\sum_{k=0}^{K} \beta_k) O_P(\frac{1}{\sqrt{N}}) \\ &= \sum_{k=1}^{K} \beta_k \sqrt{\lambda_k}\theta_{jk} + O_P(\frac{1}{\sqrt{N}}) \end{aligned}$$

On the other hand, if $p_1 + 1 \le j \le p$ then assuming that $\| \Sigma_2 \|_2$ is bounded above

$$\frac{1}{\sqrt{N}} s_j = \frac{1}{\sqrt{N}}(\Sigma_2^{1/2}\mathbf{C}y)_j = (\Sigma_2^{1/2})_j^T \frac{1}{\sqrt{N}}\mathbf{C}y = O_P(\frac{1}{\sqrt{N}})$$

Thus, in order that the "signal" $\zeta_j^K := \sum_{k=1}^{K} \beta_k \sqrt{\lambda_k}\theta_{jk}$ is detectable, it must be $\gg \frac{1}{\sqrt{N}}$. Large deviation bounds suggest that we can recover with enough accuracy only those coordinates j for which $|\zeta_j^K| \ge c_0 \sqrt{\frac{\log N}{N}}$ for some constant $c_0 > 0$ (which depends on σ, λ_k's and β_k's and $\| \Sigma_2 \|_2$).

Potentially, a lot of ζ_j^K could be smaller than that and hence those coordinates will not be selected with a high probability. If we make the threshold too small we will include a lot of "fake" coordinates (i.e., the ones with $j > p_1$) and that can spell trouble in various ways that we discussed already.

If $K = 1$, j-th component of the signal vector ζ^K is proportional to $\sqrt{\lambda_1}\theta_{j1}$. So the scheme will select only those coordinates j for which $\sqrt{\lambda_1}|\theta_{j1}|$ is big. This may not exhaust the set $\{1, \ldots, p_1\}$,

but as far as consistent estimation of θ_1 and λ_1 is concerned, this is adequate. Thus, when $K = 1$ the coordinate selection scheme is consistent.

In the case $K > 1$, a simple argument shows that there is no guarantee that the selection strategy is consistent. In other words, even when at least one of the entries among $\sqrt{\lambda_1}\theta_{j1}, \ldots, \sqrt{\lambda_K}\theta_{jK}$ is quite big, we may miss that coordinate. This has to do with the fact that we are focussing on one specific linear combination of these numbers. A remedy, involving K different linear combinations, can be easily worked out as a generalization of our coordinate selection scheme. But since we have no occasion to use it in our paper, we omit the details.

9. Discussion

Supervised principal components represents a promising tool for prediction in regression and generalized regression problems. Here we have explored its application to gene expression studies.

Cancer diagnosis and prognosis is an especially important area for which gene expression studies hold promise. It is often difficult to choose the appropriate treatment for cancer patients. Cancer is a life-threatening disease, so it must be treated aggressively. However, the most aggressive possible treatment is not appropriate for all patients. Many forms of treatment for cancer have extremely toxic side effects. (Indeed, in some cases patients will die as a result of the treatment rather than the cancer itself.) If the disease can be cured using a less aggressive form of treatment, then such a treatment would be preferable. However, some tumors will not respond to a less aggressive therapy, meaning that a more intensive treatment must be used.

Unfortunately, it is not always obvious when a more aggressive treatment is needed. Some tumors immediately go into remission when a certain treatment is applied, but other seemingly identical tumors do not respond to the same treatment. This can occur because two tumors that appear to be identical may be entirely different diseases at the genetic level (Alizadeh et al. 2000, Sorlie et al. 2001, van 't Veer et al. 2002, van de Vijver et al. 2002, Lapointe et al. 2004, Bullinger et al. 2004).

Until recently, there was no way to detect differences between two tumors at the molecular level. This is changing, however, with the advent of DNA microarrays. If subtypes of cancer are known to exist, various methods have been proposed that use microarrays to classify future tumors

into the appropriate subtype (Golub et al. 1999, Hedenfalk et al. 2001, Hastie, Tibshirani, Botstein and Brown 2001, Khan et al. 2001, Ramaswamy et al. 2001, Nguyen and Rocke 2002*a*, Nguyen and Rocke 2002*b*, Shipp et al. 2002, van 't Veer et al. 2002, van de Vijver et al. 2002, Tibshirani et al. 2001, Nutt et al. 2003). However, these methods are only useful if the subtypes have already been identified. For most types of cancer, however, no subtypes are known to exist, limiting the utility of these methods. Identifying such subtypes can be a difficult problem. Techniques such as hierarchical clustering have successfully identified cancer subtypes in several studies (Alizadeh et al. 2000, Sorlie et al. 2001, Bhattacharjee et al. 2001, Beer et al. 2002, Lapointe et al. 2004, Bullinger et al. 2004). Unfortunately, subtypes identified using hierarchical clustering are often uncorrelated with the prognosis of the patient as we demonstrated in an earlier study (Bair and Tibshirani 2004). Identification of cancer subtypes that are clinically useful remains a difficult challenge.

Supervised principal components is a technique that can be used to identify subtypes of cancer that are both biologically meaningful and clinically useful. In DNA microarray data, there are often groups of genes involved in biological processes that are not related to the survival of a cancer patient. Since the expression levels of such genes can vary from patient to patient, techniques such as hierarchical clustering may identify clusters of patients that are unrelated to the patients' survival (or other outcome of interest). Supervised principal components overcomes this problem by considering only those genes that are related to the outcome of interest. By pre-screening genes prior to performing principal component analysis, we greatly increase the likelihood that the resulting principal components are associated with the outcome of interest. Indeed, we have demonstrated that supervised principal components produces more accurate predictions than several competing methods on both simulated and real microarray data sets.

In order for a microarray predictor to be useful, it must provide information beyond what is observable using conventional diagnostic techniques. For example, clinicians routinely consider the grade and stage of a tumor when choosing the appropriate treatment for a cancer patient. If a diagnostic tool based on gene expression is not independent of the grade and stage of the tumor, the utility of such a diagnostic would be limited. As we discussed in Section 4, supervised principal components can incorporate other clinical parameters into the model to ensure that a predictor based on supervised principal components is independent of these parameters.

Identification of the most "significant" genes is an important problem in microarray analysis; several methods have been proposed for this (Ideker et al. 2000, Kerr et al. 2000, Newton et al. 2001, Tusher et al. 2001*b*). Supervised principal components allows us to calculate an "importance score" for each gene to help identify biologically significant genes. We observed in Section 3 that ranking genes based on their importance scores may produce fewer false positives than ranking genes on their raw Cox scores, a technique which is used in the "significance analysis of microarrays" procedure of Tusher et al. (2001*b*). Further research is needed to understand the relative merits of these approaches.

Finally, we note that the supervised principal components idea can be applied to other types of outcome measures, for example, classification outcomes. While this procedure seems promising, we have not yet found examples where it improves upon supervised methods such as the nearest shrunken centroid approach (Tibshirani et al. 2001). The explanation may lie in the soft-thresholding inherent in nearest shrunken centroids: it may have the same beneficial effect as the thresholding in supervised principal components.

We are currently developing an R language package `superpc` implementing supervised principal components for survival and regression data. It will be freely available on the last author's website.

Acknowledgments: Bair was partially supported by a National Science Foundation Graduate Research Fellowship. Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183. Hastie was partially supported by grant DMS-0204612 from the National Science Foundation, and grant 2R01 CA 72028-07 from the National Institutes of Health.

References

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., J., P., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R. Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. and Staudt, L. (2000), 'Identification of molecularly and clinically distinct substypes of diffuse large b cell lymphoma by gene expression profiling', *Nature* **403**, 503–511.

Alter, O., Brown, P. and Botstein, D. (2000), 'Singular value decomposition for genome-wide expression data processing and modeling', *Proceedings of the National Academy of Sciences* **97**(18), 10101–10106.

Bair, E. and Tibshirani, R. (2004), 'Semi-supervised methods to predict patient survival from gene expression data', *PLoS Biology* **2**, 511–522.

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B. and Hanash, S. (2002), 'Gene-expression profiles predict survival of patients with lung adenocarcinoma', *Nature Medicine* **8**, 816–824.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001), 'Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses', *Proceedings of the National Academy of Sciences* **98**, 13790–13795.

Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R., Tibshirani, R., Döhner, H. and Pollack, J. R. (2004), 'Gene expression profiling identifies new subclasses and improves outcome prediction in adult myeloid leukemia', *The New England Journal of Medicine* **350**, 1605–1616.

Frank, I. and Friedman, J. (1993), 'A statistical view of some chemometrics regression tools (with discussion)', *Technometrics* **35**(2), 109–148.

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999), 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–536.

Hastie, T. and Tibshirani, R. (2003), Efficient quadratic regularization for expression arrays, Technical report, Stanford University.

Hastie, T., Tibshirani, R., Botstein, D. and Brown, P. (2001), 'Supervised harvesting of expression trees', *Genome Biology* **2(1)**, 1–12.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Botstein, D. and Brown, P. (2000), 'Identifying distinct sets of genes with similar expression patterns via "gene shaving"', *Genome Biology* **1(2)**, 1–21.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer Verlag, New York.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A. and Trent, J. (2001), 'Gene-expression profiles in hereditary breast cancer', *The New England Journal of Medicine* **344**, 539–548.

Hoerl, A. E. and Kennard, R. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

Huffel, S. V. and Lemmerling, P., eds (2002), *Total Least Squares and Errors-in-Variables Modeling*, Kluwer, Dordrecht.

Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. E. (2000), 'Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data', *Journal of Computational Biology* **7**, 805–817.

Johnstone, I. M. and Lu, A. Y. (2004), 'Sparse principal component analysis', *Journal of American Statistical Association (to appear)* .

Kerr, M. K., Martin, M. and Churchill, G. A. (2000), 'Analysis of variance for gene expression data', *Journal of Computational Biology* **7**, 819–837.

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.

Lapointe, J., Li, C., van de Rijn, M., Huggins, J. P., Bair, E., Montgomery, K., Ferrari, M., Rayford, W., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D. and Pollack, J. R. (2004), 'Gene expression profiling identifies clinically relevant subtypes of prostate cancer', *Proceedings of the National Academy of Sciences* **101**, 811–816.

Lu, A. Y. (2002), Sparse principal component analysis for functional data, Technical report, Ph.D. thesis, Stanford University.

Mardia, K., Kent, J. and Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

Miller, R. G. (1986), *Beyond Anova: Basics of Applied Statistics*, John Wiley.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F. and Tsui, K. (2001), 'On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data', *Journal of Computational Biology* **8**, 37–52.

Nguyen, D. V. and Rocke, D. M. (2002*a*), 'Multi-class cancer classification via partial least squares with gene expression profiles', *Bioinformatics* **18**, 1216–1226.

Nguyen, D. V. and Rocke, D. M. (2002*b*), 'Tumor classification by partial least squares using microarray gene expression data', *Bioinformatics* **18**, 39–50.

Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R. and Louis, D. N. (2003), 'Gene expression-based classification of malignant gliomas correlates better with survival than histological classification', *Cancer Research* **63**, 1602–1607.

Paul, D. (2004), Ph.d. thesis, in preparation, Technical report, Stanford University.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Lodadagger, M., Lander, E. S. and Golub, T. R. (2001), 'Multiclass cancer diagnosis using tumor gene expression signatures', *Proceedings of the National Academy of Sciences* **98**, 15149–15154.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. and Staudt, L. M. (2002), 'The use of molecular profiling to predict survival after chemotherapy for diffuse large b-cell lymphoma', *The New England Journal of Medicine* **346**, 1937–1947.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C. and Golub, T. R. (2002), 'Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nature Medicine* **8**, 68–74.

Sorlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Lonning, P. and Borresen-Dale, A.-L. (2001), 'Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications', *Proceedings of the National Academy of Sciences* **98**, 10969–74.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2001), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci.* **99**, 6567–6572.

Tusher, V., Tibshirani, R. and Chu, G. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *Proc. Natl. Acad. Sci. USA*. **98**, 5116–5121.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R. (2002), 'A gene-expression signature as a predictor of survival in breast cancer', *The New England Journal of Medicine* **347**, 1999–2009.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415**, 530–536.

Wold, H. (1975), Soft modelling by latent variables: The nonlinear iterative partial least squares (NIPALS) approach, *in* 'Perspectives in Probability and Statistics, In Honor of M. S. Bartlett', pp. 117– 144.