BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S120–S127

Linking gene expression data with patient survival times using partial least squares

Peter J. Park[1], Lu Tian[2] and Isaac S. Kohane[1]

[1]Children's Hospital Informatics Program and Harvard Medical School, 300 Longwood Ave, Boston, MA, 02115, USA and [2]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, MA, 02115, USA

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

There is an increasing need to link the large amount of genotypic data, gathered using microarrays for example, with various phenotypic data from patients. The classification problem in which gene expression data serve as predictors and a class label phenotype as the binary outcome variable has been examined extensively, but there has been less emphasis in dealing with other types of phenotypic data. In particular, patient survival times with censoring are often not used directly as a response variable due to the complications that arise from censoring.

We show that the issues involving censored data can be circumvented by reformulating the problem as a standard Poisson regression problem. The procedure for solving the transformed problem is a combination of two approaches: partial least squares, a regression technique that is especially effective when there is severe collinearity due to a large number of predictors, and generalized linear regression, which extends standard linear regression to deal with various types of response variables. The linear combinations of the original variables identified by the method are highly correlated with the patient survival times and at the same time account for the variability in the covariates. The algorithm is fast, as it does not involve any matrix decompositions in the iterations. We apply our method to data sets from lung carcinoma and diffuse large B-cell lymphoma studies to verify its effectiveness.

Contact: peter_park@harvard.edu

Keywords: microarrays; generalized linear models; survivial analysis; Poisson regression; principal components analysis.

INTRODUCTION

Simultaneous measurement of mRNA transcripts for thousands of genes using microarrays has made it possible to study gene expression on a genome-wide scale (for overview, see Collins (1999) and the articles that follow). The two most common types are oligonucleotide and cDNA arrays, but other platforms such as SAGE (Serial Analysis of Gene Expression) are also available. Expression profiling has been used in several contexts, notably in functional characterization of genes and classification of disease types.

One of the great challenges in medicine is to correlate genotypic data, such as gene expression measurements and presence of single nucleotide polymorphisms, and other covariates, such as age and gender, to a variety of phenotypic data from the patient. Capturing the relationship between the phenotype and the genotype would not only allow for a predictive model that can aid in diagnosis and treatment, but also bring about a better understanding of the basic biological processes.

The phenotypes considered in many studies so far have been limited to relatively simple cases. The most common is the binary type, typically comparing one disease against normal or another disease (Golub *et al.*, 1999; Alon *et al.*, 1999; Alizadeh *et al.*, 2000). Larger data sets containing several types of a disease have also become common, and multiclass classification has started to receive more attention recently (Ramaswamy *et al.*, 2001; Bhattacharjee *et al.*, 2001).

In general, however, phenotypic data can take several forms. It may be, for example, 'count' data, such as the number of recurrences of a disease, or continuous data, such as blood pressure. One particularly important case is that of patient survival time, such as the time from the beginning of a treatment to a 'failure', usually an occurrence of a particular condition or death. The difficulty in dealing with survival data is that failure times may not always be observed. That means for some patients, failure occurs past a certain time but the exact time is not known ('right-censoring'). This happens, for example, when a clinical trial is terminated before all the patients have failed, or a patient leaves the study early. Unfortunately, many of the current algorithms for linking gene expression data with phenotypic data cannot be easily extended to the more general cases.

A major source of difficulty in dealing with microarray data is that the number of variables (genes) is much

larger than the number of observations (samples). Recent oligonucleotide arrays contain more than 10 000 gene probe sets and this number is expected to increase further in the future. On the other hand, the number of samples is an order of magnitude smaller, with less than a few hundred in the largest studies. Even with preprocessing of the data to filter out those genes that are unlikely to be relevant, there is a high degree collinearity in the gene profiles. Direct applications of regression techniques often result in ill-posed or computationally infeasible problems.

There are several ways of dealing with this severely ill-conditioned problem in the unsupervised setting. One approach is to identify a few genes whose behaviour is representative of the group. A more robust approach may be to combine similar profiles, as is done, for example, in Hastie *et al.* (2001) and Hedenfalk *et al.* (2001). One popular method is principal component analysis, which is a technique for finding the linear combinations of the original variables that best account for the variability in the data. A large reduction in the dimensionality results because only few such components are often needed to explain most of the variability. However, in relating the genotypic data to the phenotype, principal components or other similar methods may be ineffective because there is no guarantee that such linear combinations will predict the response well. In finding the principal components, the response variable is not taken into account. On the other hand, other methods such as ordinary least squares simply try to minimize the error in the fitted response variable and do not handle the collinearity problem.

Partial least squares is a compromise between principal component analysis and ordinary least squares regression. It is a method for constructing predictive models when there are many highly collinear covariates. First introduced in chemometrics (Wold, 1966), partial least squares attempts to find orthogonal linear combinations that explain the variability in the predictor space while being highly correlated with the response variable. Previously, this method has been extended by Marx (1996) to model a response variable belonging to an exponential family of distributions in generalized linear regression. In the context of expression data, it has been applied to the binary classification problem using logistic discrimination and quadratic discriminant analysis (Nguyen and Rocke, 2002).

In this paper, we extend the partial least squares method to deal with censored survival time data. Survival data have been used in the context of gene expression studies before, but only to verify that subclasses of the samples derived by a classification scheme have significantly different survival curves in the Kaplan–Meier analysis (Alizadeh *et al.*, 2000). A direct use of survival time as a response variable is an important problem, but the complications due to the censoring in the data make the analysis difficult. In this work, we show that the censoring issue can be circumvented by recasting the problem in a generalized regression setting. This allows us to treat the highly collinear gene expression data as predictors and directly link them to survival time as the response variable. Below, we first briefly review the partial least squares method, generalized linear regression, and models for survival data. We then describe how they can be combined to make an efficient method.

METHODS

Partial least squares

In the linear regression setting, the random response variable, y, is predicted from p covariates, $x_1, x_2, \ldots, x_p$. Given n observations, the data consist of an $n \times 1$ response vector $\mathbf{y}$ and a $n \times p$ covariate matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. If gene expression data are used as the only covariates, each column $\mathbf{x}_i$ represents a gene.

Partial least squares is an algorithm to find new variables by constructing appropriate linear combinations of original covariates. The underlying motivation for partial least squares is that there are 'latent' variables, $\mathbf{t}_1, \ldots, \mathbf{t}_s$, that explain both the response and covariate space:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \cdots + \mathbf{t}_s\mathbf{p}_s' + \mathbf{E}_s,$$

and

$$\mathbf{y} = \mathbf{t}_1 q_1 + \cdots + \mathbf{t}_s q_s + \mathbf{y}_s,$$

where $\mathbf{p}_i$ and q_i are suitably chosen weights and both $\mathbf{E}_s$ and $\mathbf{y}_s$ are small relative to the systematic part explained by the latent variables. In contrast to the dimension p of the covariate space defined by $\mathbf{X}$, the dimension of latent space, s, will be much smaller. Therefore, once the latent variables are recovered, a regular linear regression model can be fit with latent variables as predictors. The main advantage of partial least squares is its ability to handle a very large number of variables, particularly when p is much larger than n. There are many versions of the partial least squares approach. One iteratively re-weighted scheme is adopted in our algorithm described later.

Unlike principal components analysis, partial least squares chooses the linear combinations that are highly correlated with the response while accounting for the variability in the predictor space. Principal component analysis is based on the spectral decomposition of $\mathbf{X}'\mathbf{X}$; partial least squares is based on the singular value decomposition of $\mathbf{X}'\mathbf{y}$, hence reflecting the covariance structure between the predictors and the response. The relationship between partial least squares and principal component analysis or other linear regression techniques can be made precise mathematically and is described in Frank and Friedman (1993).

Generalized linear models

The linear regression of the form $\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \epsilon$ models a relationship between a set of covariates and a *continuous* response. For a different type of response variable, a generalized linear model with a transformed response is used (McCullagh and Nelder, 1989). Suppose the random response variable y comes from an exponential family, such as normal, binomial, or Poisson distributions, with $E(y_i) = \mu_i$. We can model the systematic components as before, using a linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. The connection between the random and systematic components is provided by $\eta_i = g(\mu_i)$, where $g(\cdot)$ is a given *link* function. Common generalized models include logistic regression for binary response and Poisson regression for count response.

To find the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, we often use maximum likelihood estimation. For generalized linear models, this can be carried out via the Fisher scoring method:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}^{(t)})S(\hat{\boldsymbol{\beta}}^{(t)}).$$

Given a log-likelihood function $l(\cdot)$, $S(\hat{\boldsymbol{\beta}}^{(t)}) = (\partial l/\partial\beta_1, \ldots, \partial l/\partial\beta_p)'$ is the score vector and $\mathcal{I}$ is the expected information matrix, $\mathcal{I}_{jk} = -E\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right)$. This is very similar to the Newton–Raphson algorithm, except that the expected value is taken for the Hessian matrix in order to simplify the computation. For the Poisson regression case with link function $g(\cdot) = \log(\cdot)$, the two cases are identical.

For the likelihood function from the generalized linear model, the Fisher Scoring method can be formulated as a weighted least squares algorithm:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \left(\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\beta}}^{(t)})\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\beta}}^{(t)})\mathbf{z}(\hat{\boldsymbol{\beta}}^{(t)})$$

At each iteration, we obtain new $\hat{\boldsymbol{\beta}}$, update the linear predictor $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and recompute $\mathbf{z}(\hat{\boldsymbol{\beta}}^{(t)})$ where $z_i = \eta_i + (\partial\eta_i/\partial\mu_i)(y_i - \mu_i)$. The weight matrix $\mathbf{V} = \mathrm{diag}(v_{ii})$ reflects the covariance structure of the predictor variables with the components $v_{ii} = \left[\mathrm{var}(y_i)\,(\partial\eta_i/\partial\mu_i)^2\right]^{-1}$.

Cox model for survival data

In survival analysis, the hazard function $\lambda(t)$ is often modelled. It is the probability that a failure occurs at time t given that the individual has survived up to time t, and $\lambda(t) = f(t)/[1 - F(t)]$, where $f(t)$ and $F(t)$ are the probability density and cumulative distribution functions of the failure time, respectively. A parametric approach using an exponential or Weibull distribution can be used to model this function.

The most popular way to model survival time, however, is Cox's proportional hazards model (Cox, 1972). Specifically, given a vector of covariates $\mathbf{z}$, the hazard function satisfies

$$\lambda_z(t) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{z}},$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and e^{β_i} represents the rise in risk due to a unit increase in the covariate z_i.

The estimation and statistical inference of $\boldsymbol{\beta}$ can be handled elegantly based on the 'partial' likelihood function. The observations are $y_i = \min\{T_i, C_i\}$ with $\delta_i = 1(T_i < C_i)$, where T_i and C_i, respectively, are the failure time of interest and the censoring time; $1(\cdot)$ is an indicator function. For those with $\delta_i = 0$, we only know that they survived beyond time y_i. The complex semi-parametric likelihood can be reduced to the partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}'\mathbf{z}_j}}\right)^{\delta_i}, \qquad (1)$$

where R_i is the set of those patients still in the study including those censored at t_i.

Reformulation of survival data for a Poisson regression

There is a way to transform the failure time problem into a generalized linear regression problem. This transformation (Whitehead, 1980), which we describe below, results in a Poisson regression (with a modified $\mathbf{y}$ and $\mathbf{X}$; $\mathbf{X}$ should be scaled properly) that has the same likelihood function as in (1). Therefore, the estimates for the parameters $\boldsymbol{\beta}$ by the two methods are the same. We have verified that the estimates given by the Cox proportional hazards model are the same as those given by the Poisson regression formulation. With this reformulation, we are able to apply the partial least squares method directly to the problem, circumventing the censoring issue in using the patient survival time as an outcome variable.

To be more specific, let $G = \{i : \delta_i = 1\}$ denote the index set for all observed failure times. Let the observed times be ordered in a descending manner. At time y_i, $i \in G$, we create quasi-response variables $y_{i1} = 0, \ldots, y_{i,i-1} = 0, y_{ii} = 1$, with covariates $\mathbf{x}_1, \ldots, \mathbf{x}_i$, respectively. We let $y_{ij} \sim \mathrm{Poisson}(\mu_{ij}), i \in G, j = 1, \ldots, i$ and

$$\log(\mu_{ij}) = \phi_i + \boldsymbol{\beta}'\mathbf{x}_j \qquad (2)$$

Furthermore, we create a dummy variable

$$z_k = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i \end{cases}, \quad k \in G,$$

for observation y_{ij}, and let $\phi = (\phi_i, i \in G)$, $\mathbf{z}_{ij} = (z_k, k \in G)$. Then equation (2) becomes

$$\log(\mu_{ij}) = (\mathbf{z}'_{ij}, \mathbf{x}'_j) \begin{pmatrix} \phi \\ \beta \end{pmatrix}. \tag{3}$$

It can be shown that the maximum likelihood estimates of this problem are equivalent to those given by the partial likelihood in (1). Therefore, the Fisher scoring algorithm with partial least squares for finding the latent variables can be used directly to maximize this likelihood function and obtain the solution to the original problem. β are the parameters of interest, and ϕ serve as dummy variables which we discard at the end.

Example: Suppose we have the following survival times 17^+, 15, 12^+, 10, ('+' indicates censoring) with covariates $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, respectively. Then we have $G = \{2, 4\}$, $\{y_{21}, y_{22}, y_{41}, y_{42}, y_{43}, y_{44}\} = (0, 1, 0, 0, 0, 1)$, and

$$\begin{pmatrix} \log(\mu_{21}) \\ \log(\mu_{22}) \\ \log(\mu_{41}) \\ \log(\mu_{42}) \\ \log(\mu_{43}) \\ \log(\mu_{44}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \mathbf{x}'_1 \\ 1 & 0 & \mathbf{x}'_2 \\ 0 & 1 & \mathbf{x}'_1 \\ 0 & 1 & \mathbf{x}'_2 \\ 0 & 1 & \mathbf{x}'_3 \\ 0 & 1 & \mathbf{x}'_4 \end{pmatrix} \begin{pmatrix} \phi_2 \\ \phi_4 \\ \beta \end{pmatrix}$$

A proposed algorithm

Motivated by the partial least squares method and the standard Fisher Scoring method, we propose the following general algorithm. With n observations, $\mathbf{y} = (y_1, \ldots, y_n)'$ is the response variable and $\mathbf{X}$ is the $n \times p$ covariance matrix whose ith row contains the covariates of the ith subject (We use n and p here to denote the dimensions of the new Poisson problem). This algorithm is similar to the one proposed by Marx (1996). The four outer steps (1)–(4) resemble the weighted least squares algorithm described earlier and the four inner steps (a)–(d) incorporate the partial least squares algorithm.

(1) Initialization:

$$\mathbf{y}_0 = \psi(\mathbf{y}) - \frac{1}{n}\mathbf{1}'\hat{\mathbf{V}}\psi(\mathbf{y});$$

$$\mathbf{E}_0 = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\hat{\mathbf{V}}\mathbf{X};$$

$$\hat{\mathbf{V}} = \text{diag}\left[\frac{h'(g(\psi(y_i)))^2}{\text{var}(y_i)}\right]_{i=1,\ldots,n},$$

where $g(\cdot)$ is the link function, $h(\cdot) = g^{-1}(\cdot)$, $\mathbf{1}$ is a column vector containing 1s, and $\psi(y)$ is a suitably transformed version of y. (By $\psi(\mathbf{y})$, we mean $[\psi(y_1), \psi(y_2), \ldots, \psi(y_n)]'$.)

(2) Iterate (a)–(d) to obtain R latent variables: (A weighted version of partial least squares)

(a) Obtain the loadings for a new latent variable:

$$\mathbf{w}_k = \mathbf{E}'_{k-1}\hat{\mathbf{V}}\mathbf{y}_{k-1}.$$

(b) Create the latent variable:

$$\mathbf{t}_k = \mathbf{E}_{k-1}\mathbf{w}_k.$$

(c) Fit the linear model $\mathbf{y}_{k-1} = q_k\mathbf{t}_k + \epsilon_k$ and regress out the latent variable from the response variable:

$$q_k = \frac{\mathbf{y}'_{k-1}\hat{\mathbf{V}}\mathbf{t}_k}{\mathbf{t}'_k\hat{\mathbf{V}}\mathbf{t}_k},$$

$$\mathbf{y}_k = \mathbf{y}_{k-1} - q_k\mathbf{t}_k.$$

(d) Fit the linear model $\mathbf{E}_{k-1} = \mathbf{t}_k\mathbf{p}'_k + \epsilon_k$ and regress out the latent variable from the predictor variables:

$$\mathbf{p}' = \frac{\mathbf{t}'_k\hat{\mathbf{V}}\mathbf{E}_{k-1}}{\mathbf{t}'_k\hat{\mathbf{V}}\mathbf{t}_k},$$

$$\mathbf{E}_k = \mathbf{E}_{k-1} - \mathbf{t}_k\mathbf{p}'$$

(3) Update η:

$$\eta = \frac{1}{n}\mathbf{1}'\hat{\mathbf{V}}\psi(\mathbf{y}) + \sum_{k=1}^{R} q_k\mathbf{t}_k.$$

(4) Update the weight matrix:

$$\hat{\mathbf{V}} = \text{diag}\left[\frac{h'(\eta_i)^2}{\text{var}(y_i)}\right]_{i=1,\ldots,n}$$

and

$$\mathbf{y}_0 = \eta + \text{diag}\left[\frac{1}{h'(\eta_i)}\right]_{i=1,\ldots,n}(\mathbf{y} - h(\eta)).$$

(5) If $\Delta\eta$ is not sufficiently small, return to (2).

(6) Select s latent variables, $(\mathbf{t}_1, \ldots, \mathbf{t}_s)$, to fit the generalized linear model to estimate β.

The motivation behind steps (2a)–(2b) is to find the direction $\mathbf{t}_k = \mathbf{E}_{k-1}\mathbf{w}_k$ to maximize $\mathbf{t}'_k\hat{\mathbf{V}}\mathbf{y}_{k-1}/\sqrt{\mathbf{t}'_k\mathbf{t}_k}$, the weighted covariance of the latent variable with the response variable, subject to the orthogonality condition $\mathbf{t}'_k\hat{\mathbf{V}}\mathbf{t}_l = 0$, $l = 1, 2, \ldots, k-1$. In (2c)–(2d), we regress out the latent variable immediately from both the response variable and the matrix of predictor variables.

By working with the residuals, the next latent variable is constructed in a subspace orthogonal to the previous latent variables. Because of the complex nonlinear nature of the algorithm, it is difficult to make general statement about its convergence properties. However, in the two data sets we examined below, convergence was achieved in few steps with the $\Delta\eta$ criterion.

A major advantage of the partial least squares algorithm is that it does not involve any matrix decompositions. This is in contrast to most other regression methods, which require matrix inversions. This property makes it possible to deal with a large number of genes computationally, and it will be an increasingly important feature as the number of genes and patients in data sets grows larger.

RESULTS

In Bhattacharjee *et al.* (2001), 186 lung tumor samples were profiled using oligonucleotide arrays to examine subclasses of lung carcinomas. The specimens included histologically defined lung adenocarcinomas, squamous cell lung carcinomas, pulmonary carcinoids, and small-cell lung carcinomas. Among these, 125 adenocarcinoma samples were provided with clinical and histological data, such as tumor size, stage, type of operation performed, site of relapse, age, gender, and smoking history; the same samples were also associated with survival times and a censoring indicator. We note, however, that while all the samples belong to the adenocarcinoma subtype, there may be a significant heterogeneity due to the different stages of the disease when the specimens were taken. Deleting the subjects with any missing covariates (which included age, gender, and smoking history), we use 122 subjects, each with 366 gene expression level measurements. These genes were selected based on their high correlation between replicate pairs (Pearson correlation > 0.85). We can include more genes by lowering the threshold on the measurement reliability; the speed of algorithm we propose scales linearly with the number of genes.

One way to evaluate the new variables $\mathbf{t}_i$ derived from the algorithm above is to fit them with the Cox model. In Table 1, we show the result of fitting the Cox model with the top 10 covariates obtained by partial least squares. We see that the first few have very low p-values, with four of them below 10^{-6}. We contrast this with using the first 10 directions from principal components analysis. At the $p = 0.01$ level, six latent variables and only one principal component are significant. The likelihood ratio test with 10 degrees of freedom gives p-values of $< 10^{-16}$ and 0.007 29 for partial least squares and principal components, respectively. In Table 2, we perform the same analysis but with only one covariate in each method. We see that using just a single component gives a highly significant result for the partial least squares (p-value of 3.6×10^{-7}) but not for the principal components (p-value of 0.02). The likelihood ratio test gives p-values of 8.6×10^{-8} for partial least squares and 0.0154 for principal components. Based on this single component, we divide the samples into two groups and plot the Kaplan–Meier survival curves for the two groups in Figure 1, similar to what is done in Hastie *et al.* (2001). The difference in survival between the groups is highly significant, with a p-value of 2×10^{-5}.

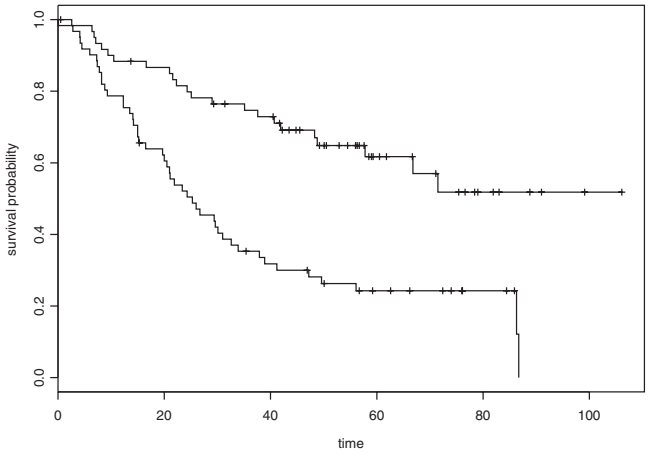

Fig. 1. Kaplan–Meier survival curves. Based on the first partial least squares component, we split the patients into two groups, one with higher than median risk and the other with lower than median risk. The x-axis is the time in months and the y-axis is the probability of overall survival. The p-value is $p = 0.000\,020$ for the null hypothesis of no difference between the two groups.

In Figure 2, the scatterplot shows the patients in terms of their first two latent variables. We see that even in two dimensions the patients with observed failure times appear to be separated from those with censored failure times. Because we do not know the actual failure times of censored patients, this description is of limited value, but it still shows the trend.

After applying partial least squares regression and obtaining a large number of latent components, we have a sequence of models with the number of covariates $1, 2, \ldots, R$. The optimal selection of the model size can be conducted through a K-fold cross-validation. First, the data set is split into K parts. For each $k = 1, 2, \ldots, K$, the partial least squares procedure is trained on all the data except the kth part, and then loss of predicting the kth part through the trained model is estimated. The results are averaged over $k = 1, \ldots, K$, and the optimal number of covariates is selected for the model minimizing the loss function. A negative log likelihood function and partial likelihood function can serve as the loss function in the generalized linear model and proportional hazards model, respectively. Our results show that a single latent variable model is selected based on this criterion.

Table 1. Significance of latent variables. To evaluate the latent variables, the Cox proportional hazards model was fit. Here we use 10 latent variables from partial least squares and 10 principal components for the lung carcinoma data. Many of the latent variables are highly significant. Six latent variables and one principal component are significant at $p = 0.01$

	Partial least squares				Principal components			
	coef	se(coef)	z	p-value	coef	se(coef)	z	p-value
1	0.007 64	0.001 56	4.91	9.3E−007	−0.0620	0.0209	−2.974	0.0029
2	0.017 39	0.003 33	5.22	1.8E−007	0.0573	0.0239	2.396	0.0170
3	0.011 33	0.002 80	4.05	5.2E−005	0.0204	0.0264	0.773	0.4400
4	0.022 07	0.004 34	5.09	3.6E−007	−0.0343	0.0302	−1.137	0.2600
5	0.027 94	0.004 99	5.60	2.1E−008	−0.0480	0.0286	−1.678	0.0930
6	0.010 51	0.004 01	2.62	8.7E−003	−0.0467	0.0343	−1.360	0.1700
7	0.005 23	0.003 46	1.51	1.3E−001	0.0192	0.0343	0.561	0.5700
8	0.010 59	0.004 33	2.45	1.4E−002	0.1182	0.0456	2.594	0.0095
9	0.008 50	0.006 29	1.35	1.8E−001	−0.0728	0.0455	−1.600	0.1100
10	0.006 56	0.004 58	1.43	1.5E−001	−0.0739	0.0434	−1.703	0.0890

Table 2. Significance of latent variables. The Cox proportional hazards model was again fit in order to evaluate the latent variables. This time, we use only one latent variable and one principal component in the Cox model. The latent variable from the partial least squares method is highly significant.

Partial least squares				Principal components			
coef	se(coef)	z	p-value	coef	se(coef)	z	p-value
0.007 43	0.001 46	5.09	3.6E−007	−0.0482	0.0208	−2.32	0.02

We also applied our method to the data set from patients with diffuse large B-cell lymphomas (DLBCLs), studied in Shipp *et al.* (2002). This data included the expression levels of 6817 genes in pre-treatment biopsies obtained from 58 DLBCL patients who have received chemotherapeutic regimens, as well as their long-term clinical outcomes. In the absence of duplicate arrays, we were unable to reduce the number of genes as we did in the previous example. Instead, we filtered out those genes that have expression levels lower than a threshold of 100 in more than 10 cases, and used a low-entropy filter to eliminate genes with extreme outliers. We then applied the partial least squares method with the remaining 2800 genes. The results were similar to the those obtained in the lung carcinoma data. For the model containing 10 latent variables, there are 6 variables with p-values $< 1 \times 10^{-5}$ and the likelihood ratio test give a p-value of $< 1 \times 10^{-16}$. Principal components analysis, on the other hand, involves computing eigenvectors of a 2800×2800 matrix and gives a less significant result.

DISCUSSION

The primary emphasis in partial least squares is on predicting the response. As a result, while it finds the new variables that best explain the response, it is often difficult to interpret those latent variables. When linear combinations are used to select new variables, interpretation is difficult in general, although the first one or two directions in principal components sometimes can be explained (Raychaudhuri *et al.*, 2000). If any gene is closely correlated with the survival times, we should be able to identify them by their large coefficients in the first few latent variables. In the data set we examined, we have not found any dominant genes, even though the latent variables we obtained were highly significant. In Figure 3, we show the coefficients of the genes for the first three latent variables.

The reformulation of the censored problem as a Poisson regression increases the dimension of the problem for the iteration step. If the original problem had p columns, it increases by the number of distinct failure times, through the addition of ϕ_i in equation (2). That number is bounded by the number of patients n, and the increase is usually

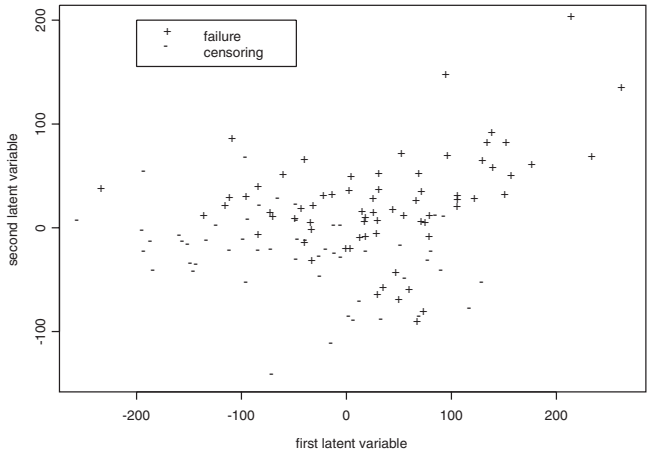

Fig. 2. Patients with observed and censored failures. Using the first two latent variables as the new coordinates, we see that the patients with observed failure times appear to be separated from those with censored failure times.

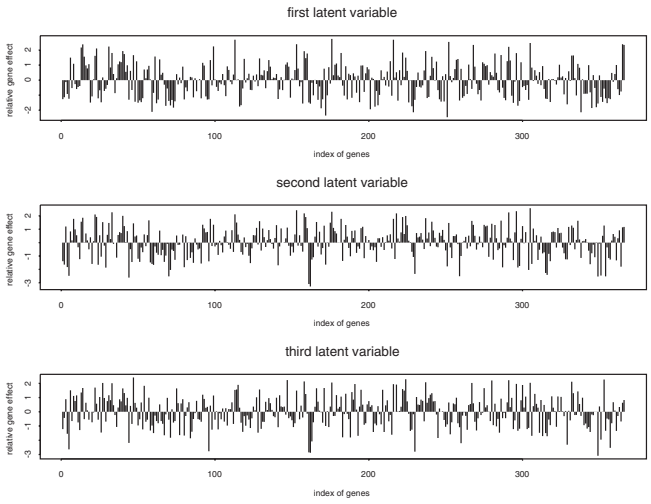

Fig. 3. Coefficients for the latent variables. Each latent variable is a linear combination of the original variables (genes). The coefficient for the 366 genes we used are shown here for the first three variables. No dominant genes seem to be present.

very small relative to p. On the other hand, the increase in the number of rows varies depending on the number of censored times and when they occur; in the worst case, the dimension goes from n to roughly $n^2/2$. Even for a moderate size n, this results in a substantial increase. Fortunately, the partial least squares iterations are fast. We have found that we achieve a reasonable speed even when there are more than 100 patients as in our data set.

The speed and effectiveness of the algorithm is in part due to the construction of an orthogonal sequence (in some weighted metric) in the iteration, which resembles the speedy conjugate gradient method. Indeed, a recursive formulation of the iteration can be shown to be equivalent to the conjugate gradient applied to the normal equations of generalized linear regression (Marx, 1996). The algorithm is highly nonlinear, as each latent variable is regressed out both from the response vector and the covariate matrix at each iteration. Because of this, understanding the convergence properties is not simple. With the lung carcinoma data, we have found that as the number of latent variables increases, the relative error for the weight matrix becomes somewhat inconsistent and hard to predict; however, we suspect that they still result in nearly identical predictors.

In this paper, we have shown a way to link censored data directly to gene expression data. With an iteratively reweighted partial least squares approach for selecting appropriate predictors, we are able to solve the transformed problem efficiently in the generalized linear model setting. Several latent variables we derive are highly significant, as the likelihood function values indicate. Eventually, a more comprehensive model relating various types of genotypic and phenotypic data will be necessary. Typical genotype data would include gene expression profiles and presence of mutations in genes; clinical data may include not just patient survival information but post-operative pathological staging, histopathological diagnosis, and site of disease recurrence and many others. Incorporating all of this information in a statistically coherent and computationally feasible framework remains an important challenge.

ACKNOWLEDGEMENT

We thank Meredith Goldwasser for a careful reading of this manuscript.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.G., Sabet,H., Tran,T., Yu,X. *et al.*, (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al.*, (2001) Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Collins,F.S. (1999) Microarrays and macroconsequences. *Nature Genetics*, **21(Supp)**, 2.

Cox,D.R. (1972) Regression models in life tables (with discussion). *J. Roy. Stat. Soc.* Ser. B, **34**, 187–220.

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

Golub,T.R., 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**, 531–537.

Hastie,T., Tibshirani,R., Botstein,D. and Brown,P. (2001) Supervised harvesting of expression trees. *Genome Biol.*, **2**, research0003.1–research0003.12.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M. *et al.*, (2001) Gene expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **244**, 539–548.

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

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*, second edition, Chapman and Hall.

Nguyen,D.V. and Rocke,D.M. (2002) Classification of acute leukemia based on dna microarray gene expressions using partial least squares. In Lin,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic, pp. 109–124.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.-H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J., Poggio,T., Gerald,W., Loda,M., Lander,E.S. and Golub,T.R. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Raychaudhuri,S., Stuart,J.M. and Altman,R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation data. *Pacific Symposium on Biocomputing*, 452–463.

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. *et al.*, (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Whitehead,J. (1980) Fitting Cox's regression model to survival data using GLIM. *Appl. Statist.*, **29**, 268–275.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiah,P. (ed.), *Multivariate Analysis*. Academic Press, New York, pp. 391–420.