



Partial least squares dimension reduction for microarray gene expression data with a censored response

Danh V. Nguyen *

*Division of Biostatistics, Public Health Sciences, School of Medicine, University of California,
One Shields Avenue, Davis, CA 956168638, USA*

Received 24 November 2003; received in revised form 8 September 2004; accepted 21 October 2004

Abstract

An important application of DNA microarray technologies involves monitoring the global state of transcriptional program in tumor cells. One goal in cancer microarray studies is to compare the clinical outcome, such as relapse-free or overall survival, for subgroups of patients defined by global gene expression patterns. A method of comparing patient survival, as a function of gene expression, was recently proposed in [Bioinformatics 18 (2002) 1625] by Nguyen and Rocke. Due to the (a) high-dimensionality of microarray gene expression data and (b) censored survival times, a two-stage procedure was proposed to relate survival times to gene expression profiles. The first stage involves dimensionality reduction of the gene expression data by partial least squares (PLS) and the second stage involves prediction of survival probability using proportional hazard regression. In this paper, we provide a systematic assessment of the performance of this two-stage procedure. PLS dimension reduction involves complex non-linear functions of both the predictors and the response data, rendering exact analytical study intractable. Thus, we assess the methodology under a simulation model for gene expression data with a censored response variable. In particular, we compare the performance of PLS dimension reduction relative to dimension reduction via principal components analysis (PCA) and to a modified PLS (MPLS) approach. PLS performed substantially better relative to dimension reduction via PCA when the total predictor variance explained is low to moderate (e.g. 40%–60%). It performed similar to MPLS and slightly better in some cases.

* Tel.: +1 530 754 6510; fax: +1 530 752 3239.

E-mail address: ucdnguyen@ucdavis.edu

Additionally, we examine the effect of censoring on dimension reduction stage. The performance of all methods deteriorates for a high censoring rate, although PLS–PH performed relatively best overall.

© 2005 Published by Elsevier Inc.

Keywords: DNA Microarray; Dimension reduction; Gene expression; Partial least squares; Principal components; Proportional hazard regression

1. Introduction and motivating applications

DNA microarray technologies have found broad applications, especially in biomedical research. For an overview of the technological and biological aspects of DNA microarrays, including applications, see [1] by Nguyen et al. and references therein. The application of DNA microarray technologies in cancer research is one specific area of interest. In this paper we examine a method for analyzing censored patient survival times (the censored response variable) with their corresponding gene expression profiles as covariates (predictors). To be more concrete, consider the following two motivating applications:

1. *Example: Diffused large B-Cell lymphoma.* In a study of diffused large B-cell lymphoma (DLBCL), mRNA expression for over 5622 gene probes were measured from microarray experiments [2]. In addition to the gene expression data collected, patient survival times were ascertained for $N = 40$ DLBCL patients. However, not all survival times could be observed by the end of the study. Of the 40 observed survival times, 22 were times of death; thus the percentage of censored observations is 55%.
2. *Example: Breast carcinomas.* Similarly, in a prospective breast carcinomas study, thousands of mRNA gene expression measurements were obtained simultaneously from microarray experiments [3]. Tissue samples for the microarray experiments were obtained from patients in a prospective study on locally advanced breast cancer with no distant metastases. Survival data was available for $N = 49$ patients with approximately 61% censoring.

Similar cancer microarray studies with survival data can be found in [4] for central nervous system embryonal tumors, [5] for DLBCL, [6] and [7] for prostate cancer, and [8] for lung carcinomas among others.

The data structure in the examples presented above can be more formally described as follows. Suppose that Y_i is the true survival time of the i th patient. The variate of interest at the end of a study, the survival time, cannot be observed completely. Instead, we are only able to observe $T_i = \min(Y_i, Z_i)$, where Z_i is a censored value. Also, recorded for the i th patient are p gene expression values, denoted by $\mathbf{x}'_i = (x_{i1}, \dots, x_{ip})$ ($i = 1, \dots, N$), obtained from microarray experiments. This vector of gene expression values is called the (patient-specific) gene expression profile or pattern. We refer to the expression profile generally as covariates or predictors. Thus, the typical data set, illustrated by the examples above, consists of N samples. In addition, each sample contains the triple $\{T_i, \delta_i, \mathbf{x}_i\}$, where \mathbf{x}_i is the gene expression pattern, T_i is the survival time if $\delta_i = 1$, and it is the right-censored time if $\delta_i = 0$. Note that the number of observed survival times that

are censored is $\#\{i: \delta_i = 0\}$. It is assumed that the censoring mechanism or the censoring time distribution is independent of the survival time distribution.

For microarray data, the sample size is small relative to the number of covariates: $N \ll p$. The problem of small sample size is further compounded in this context. This is because the effective sample size, the number of uncensored samples, is actually smaller: $\#\{i: \delta_i = 1\} < N$. In contrast to high-dimensional microarray data, there are many more samples than there are predictors ($N \gg p$) in the traditional data setting. For the traditional setting, the proportional hazard (PH) regression model, introduced by Cox in [9], is a powerful tool for studying the relationship between a censored response, such as survival time, and a set of predictors. The PH regression model is

$$h(t; \mathbf{x}_i; \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta}), \quad (1)$$

where h is the hazard function associated with covariate \mathbf{x}_i and h_0 is an unspecified baseline hazard function. The hazard function $h(t)$ measures the instantaneous risk of ‘dying’ at time t (details in Section 2). The proportional hazard regression model (1) can be expressed in terms of the survival function, $S(t) = \Pr(T > t)$, as

$$S(t; \mathbf{x}_i; \boldsymbol{\beta}) = S_0(t)^{\exp(\mathbf{x}_i' \boldsymbol{\beta})},$$

where $S_0(t)$ is the corresponding baseline survival function. (We will elaborate on this in Section 2.)

However, when the number of predictors (p) is greater than the number of available samples (N), as in microarray data, traditional regression methodologies do not work. This includes the PH regression model (1), where $p > N$ leads to non-unique estimates of model parameters. To cope with the high-dimensionality in microarray data, one can employ dimension reduction techniques. Thus, we previously proposed a two-stage procedure to predict survival based on gene expression data. More precisely, we reduce the dimensionality of gene expression data by partial least squares (PLS) in the first stage and then predict survival probabilities by applying PH regression model to the reduced data [11,12]. In stage 1, we used PLS dimension reduction to extract K dimensions ($K \ll N$) with both optimal predictor variance and correlation between predictors and survival time (details in Section 2). We refer to this two-stage procedure as partial least squares proportional hazard (PLS–PH) regression.

For example, Fig. 1(a) displays the estimated survival curves (\hat{S}) via PLS–PH regression for two DLBCL subgroups (example 1): (1) germinal centre (GC) B-like and activated B-like. The estimated survival curves reveal a strong divergence in clinical behavior of GC B-like and activated B-like DLBCLs. The clinical outcomes for the two groups, defined molecularly by gene expression variation, are significantly different. Similarly, Fig. 1(b) gives the PLS–PH regression estimates of the survival curves for five novel subgroups of breast carcinomas (example 2): basil-like, *ERBB2*+, normal breast-like, luminal subtype A, and luminal subtype B + C [3]. In both examples, tumor subgroups were determined by variation in the global expression patterns obtained from DNA microarrays. PLS–PH regression is a method to predict the clinical outcomes for these molecularly distinct subgroups. For details on these applications of PLS–PH regression, see [11,12].

In this paper we provide a systematic assessment of the performance of the PLS–PH regression method. An analytical study of PLS–PH regression is not tractable because the dimension reduction method (PLS) involves complex non-linear functions of both the predictors and response variable [13]. Thus, we assess the dimension reduction performance of the methodology based on a

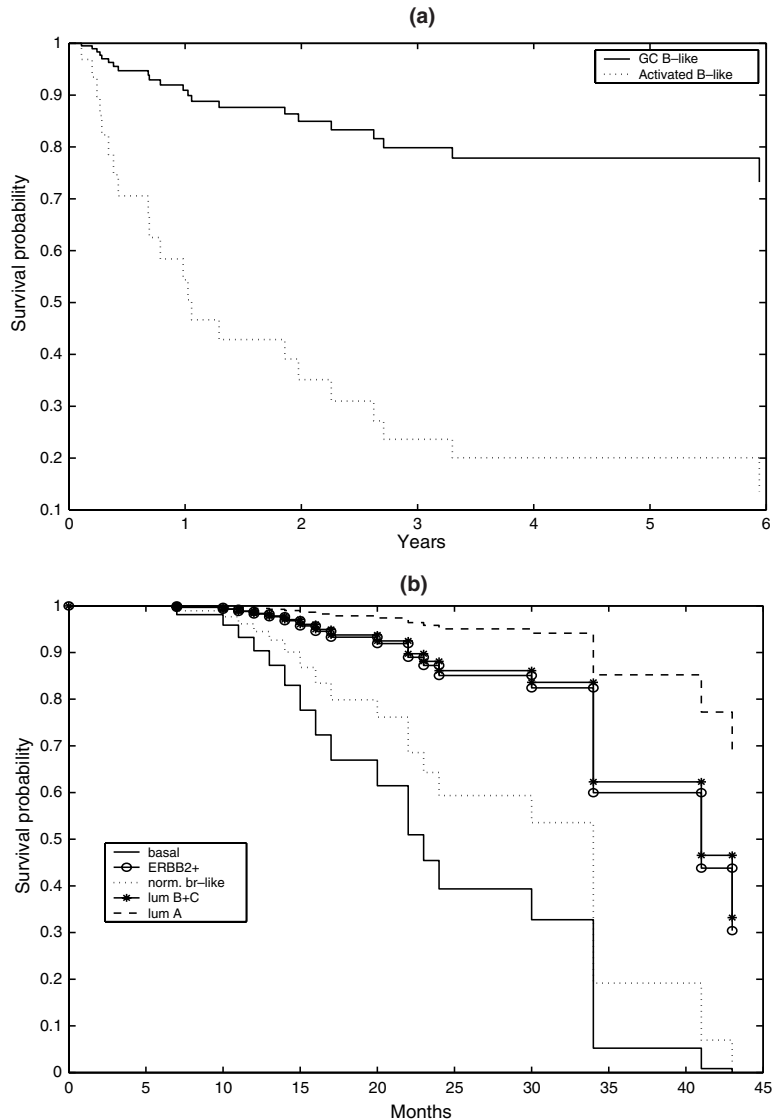


Fig. 1. (a, top) Given are estimated survival curves (germinal centre B-like—, and activated B-like ···) from the PLS–PH regression model fit to the diffuse large B-cell lymphoma microarray data set. (b, bottom) Similarly, given are the estimated survival curves for five subgroups of breast carcinomas: basal-like —, ERBB2+ –○–○–, normal breast-like ···, luminal subtype A – – – –, and luminal subtype B+C – * – * –. The curves were obtained for the group-average component profiles. Adopted from [12].

simulation model for gene expression data with a censored response variable. The aim of this study is two-fold. First, we compare the relative performance of PLS dimension reduction to dimension reduction via principal components analysis (PCA), a well-known technique widely used for microarray data. Additionally, we compare PLS to a modified PLS (MPLS) approach, which attempts to incorporate censoring information into the dimension reduction stage. Second, we examine the effect of the amount of censoring on the dimension reduction.

The paper is organized as follows. In Section 2, we describe the PLS–PH regression methodology introduced by Nguyen and Rocke [11,12]. This includes a description of PCA, PLS, and MPLS dimension reduction. The simulation model and procedure are described in Section 3. Results are summarized in Section 4 and we conclude in Section 5.

2. Dimension reduction and proportional hazard model

2.1. Proportional hazard regression model in high dimension

Let T^* be the survival time. The survival function, defined by $S(t) = \Pr(T^* > t)$, is the probability of still surviving at time t . The hazard function $h(t)$ measures the instantaneous risk in the next small time interval, given survival to time t . The hazard model under consideration is the proportional hazard (PH) regression model [9], given by

$$h(t; \mathbf{x}_i; \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{x}'_i \boldsymbol{\beta}), \quad (2)$$

where $h_0(t)$ is an unspecified baseline hazard function. Alternatively, the proportional hazard regression model (2) can be expressed in terms of the survival function, $S(t; \mathbf{x}_i; \boldsymbol{\beta})$, as

$$S(t; \mathbf{x}_i; \boldsymbol{\beta}) = S_0(t)^{\exp(\mathbf{x}'_i \boldsymbol{\beta})}, \quad (3)$$

where $S_0(t)$ is the corresponding baseline survival function.

An estimate of the model parameters, denoted $\hat{\boldsymbol{\beta}}$, can be obtained by maximizing the partial log-likelihood [9,10]. To estimate the survival function $S(t; \mathbf{x}_i; \boldsymbol{\beta})$, we use the standard product limit estimate [22] to obtain the baseline survival function estimate, denoted $\hat{S}_0(t)$. Thus, substituting $\hat{S}_0(t)$ and $\hat{\boldsymbol{\beta}}$ into (3) gives $\hat{S}(t; \mathbf{x}_i; \boldsymbol{\beta}) = \hat{S}_0(t)^{\exp(\mathbf{x}'_i \hat{\boldsymbol{\beta}})}$.

As mentioned in the Introduction Section, estimation of the model parameters, $\boldsymbol{\beta}$, is not possible when the dimension p is greater than the number of available samples N . Thus, one approach to this problem is to, first, reduce the dimensionality from p to $K \ll N$. We will describe dimension reduction strategies, in more details, in the next section. For the moment, suppose that $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_K]$, where \mathbf{t}_k ($k = 1, \dots, K$) is a $N \times 1$ vector of gene component values obtained after dimension reduction of gene expression matrix \mathbf{X} . After dimension reduction, we approximate \mathbf{X} with \mathbf{T} and apply the PH regression model (2) in the reduced subspace. Hence, a simple two-stage procedure introduced by Nguyen and Rocke [12] consists of dimension reduction via partial least squares (PLS) in stage one, followed by PH regression in stage two. As we will discuss in the next section, PLS is a dimension reduction method which uses information from both the predictors and response variable.

2.2. PLS and related dimension reduction techniques

PLS–PH regression involves dimension reduction via partial least squares. PLS is similar to principal components analysis (PCA) [14,15], a well-known dimension reduction technique. In PCA, linear combinations of the predictors are sequentially constructed to maximize a variance-objective criterion. More precisely, orthogonal linear combinations are constructed to maximize the variance of the linear combination of the predictors sequentially,

$$\mathbf{w}_k = \operatorname{argmax}_{\mathbf{w}'\mathbf{w}=1} \operatorname{var}(\mathbf{X}\mathbf{w}) = \operatorname{argmax}_{\mathbf{w}'\mathbf{w}=1} (N-1)^{-1} \mathbf{w}'\mathbf{S}\mathbf{w}, \quad (4)$$

subject to the orthogonality constraints $\mathbf{w}'_k \mathbf{S}\mathbf{w}_j = 0$, for all $1 \leq j < k$. We have denoted $\mathbf{S} = \mathbf{X}'\mathbf{X}$ in (4), where \mathbf{X} is the $N \times p$ matrix of predictor (gene expression) values. In practice, PCA is applied to the sample covariance or correlation matrix through the spectral decomposition. For example, the spectral decomposition of a correlation matrix, \mathbf{R} , is $\mathbf{R} = \mathbf{V}\mathbf{\Delta}\mathbf{V}'$, where $\{\lambda_k\}_{k=1}^{N-1}$ are the eigenvalues, $\mathbf{\Delta} = \operatorname{diag}\{\lambda_1 \geq \dots \geq \lambda_{N-1}\}$, and $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_{N-1})$ are the corresponding eigenvectors. The principal components (PCs) are constructed as linear combinations of the predictors with weights given by $\mathbf{w}_k = \mathbf{v}_k$; hence, the PCs are $\{\mathbf{t}_k = \mathbf{X}\mathbf{v}_k\}_{k=1}^{N-1}$. Furthermore, the proportion of variation explained by the k th PC is λ_k/p and the cumulative proportion for K PCs is $\sum_{k=1}^K \lambda_k/p$ [16]. Therefore, if there are only d ($< N < p$) underlying components that explain nearly all of the observed variation, then we expect that $\sum_{k=1}^d \lambda_k/p \approx 1$.

Note that in PCA, the weights are obtained independent of the response variable; thus, the selection of components with maximal variance may not necessarily be predictive of the response [15]. Thus, for prediction of survival times using gene expression data, we proposed PLS dimension reduction, which maximizes a covariance-objective criterion [12]. PLS optimizes the correlation between predictors and the response and the predictor variances simultaneously. More precisely, PLS components are linear combinations of the predictor variables, constructed to maximize an objective criterion based on the sample covariance between \mathbf{y} and $\mathbf{X}\mathbf{w}$. The k th PLS component is obtained by finding the weight vector, \mathbf{w} , satisfying

$$\mathbf{w}_k = \operatorname{argmax}_{\mathbf{w}'\mathbf{w}=1} \operatorname{cov}(\mathbf{X}\mathbf{w}, \mathbf{y}) = \operatorname{argmax}_{\mathbf{w}'\mathbf{w}=1} (N-1)^{-1} \mathbf{w}'\mathbf{X}'\mathbf{y}. \quad (5)$$

Unlike PCA, the PLS weights are non-linear functions of both the predictors and response variable [13]. The PLS components are obtained as $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k$, and, as in PCA, are orthogonal: $\mathbf{t}'_k \mathbf{t}_j = \mathbf{w}'_k \mathbf{S}\mathbf{w}_j = 0$ for all $1 \leq j < k$. Numerical algorithms to obtain \mathbf{w}_k can be found in [17,18]. Applications and implementations of PLS for tumor classification based on high-dimensional microarray gene expression data can be found in [19–21].

The PLS covariance-objective criterion (5) incorporates the response information in the dimension reduction stage. In addition, the PLS criterion (5) is designed for a continuous response variate. However, response outcomes such as overall survival or relapse-free survival are right-censored. Thus, it is of interest to also consider dimension reduction strategies which utilize censoring information together with the response information. One approach is based on modification of the PLS weights \mathbf{w}_k to include censoring information. We consider a modification based on the following expression for the PLS weights (5),

$$\mathbf{w}_k = \sum_{i=1}^N \theta_{ik} \mathbf{v}_i, \quad (6)$$

where \mathbf{v}_i is the i th eigenvector of $\mathbf{S} = \mathbf{X}'\mathbf{X}$ [13]. The scalars, θ_{ik} , depend on the response values, $\{y_i\}_{i=1}^N$, only through the dot product $a_i = \mathbf{u}'_i \mathbf{y}$, where \mathbf{u}_i is the eigenvector of $\mathbf{X}\mathbf{X}'$ [13]. This dot product is also the slope coefficient in the simple linear regression of Y on U_i . Because the response is censored in this case, we replace the dot product by the slope coefficient from the PH regression of Y on U_i . We explore this modification of PLS (MPLS) for dimension reduction. Details of the computation of MPLS components are given in the [Appendix A](#) section.

Table 1
Three dimension reduction methods

Method	Stage 1: dimension reduction		Stage 2: PH regression	
	Incorporates information from:		Incorporates information from:	
	Censoring data	Response data	Censoring data	Response data
PCA–PH			•	•
PLS–PH		•	•	•
MPLS–PH	•	•	•	•

Stage 1 is the dimension reduction step and stage 2 is the PH regression step.

In summary, the three dimension reduction methods, namely PCA, PLS and MPLS, deal with the response and censoring information in different ways. Dimension reduction via PCA completely ignores the response data and censoring information. PLS incorporates the response variable in the dimension reduction process, but treats it as uncensored. The modified PLS, namely MPLS, incorporates the response data and censoring information. However, after the dimension reduction step one, we note that censoring and response information are actually used in stage two, through the PH regression model (2) for all three methods. Table 1 summarizes the response and censoring information used in the two-stage procedures (PCA–PH, PLS–PH, and MPLS–PH). We examine the performance of these dimension reduction methods using a simulation model for gene expression data, which will be described in the next section.

3. Simulation experiments

We assess the performance of the dimension reduction methods under a simulation model. The simulation procedure consists of two main parts:

- *Generating gene expression values: the data matrix.* We consider a model for the data matrix of predictor values, \mathbf{X} , with $d (\ll N \ll p)$ underlying components, perturbed by a noise factor [13]. In addition, to flexibly cover a spectrum of real microarray data sets, the data matrix is generated such that first K principal components explain a specified proportion of predictor variability. This proportion can then be systematically varied.
- *Generating observed survival times: the censored response variable.* The survival times, which are functions of \mathbf{X} , are generated to satisfy the proportional hazard model (2). We vary the amount of censoring to assess the impact on the dimension reduction methods.

3.1. Generating gene expression values: the data matrix

The $N \times p$ data matrix is generated from a basic model with d underlying components and an error component [13]. The i th sample is obtained from

$$\mathbf{x}_i^* = r_{1i}\boldsymbol{\tau}_1 + \dots + r_{di}\boldsymbol{\tau}_d + \boldsymbol{\epsilon}_i \quad i = 1, \dots, N. \tag{7}$$

Table 2

Summary of simulation experiments: generating (A) the gene expression values and (B) the survival times

	Model	Parameters
<i>(A) Expression values</i>		
Components: $\{\tau_k\}_{k=1}^d$	$N(\mu_\tau, \sigma_\tau^2)$	μ_τ, σ_τ^2
Error: $\{\epsilon_i\}_{i=1}^N$	$N(\mu_\epsilon, \sigma_\epsilon^2)$	$\mu_\epsilon, \sigma_\epsilon^2$
Data matrix: $\{x_{ij}\}_{i,j=1}^{N,p}$	$LN(a_i, b_i^2)$	$a_i = \mu_\tau \sum_{k=1}^d r_{ki}$ $b_i^2 = \sigma_\tau^2 \sum_{k=1}^d r_{ki}^2 + \sigma_\epsilon^2$
<i>(B) Survival times</i>		
Survival times: $\{Y_i\}_{i=1}^N$	$Y_i = Y_{0i} \exp(-\mathbf{x}'_i \boldsymbol{\beta})$	$Y_{0i} \sim \text{Exp}(\lambda_y)$ and $\text{Weib}(\alpha_y, \lambda_y)$
Censored times: $\{Z_i\}_{i=1}^N$	$Z_i = Z_{0i} \exp(-\mathbf{x}'_i \boldsymbol{\beta})$	$Z_{0i} \sim \text{Exp}(\lambda_z)$ and $\text{Weib}(\alpha_z, \lambda_z)$
Observed survival: $T_i = \min(Y_i, Z_i)$		
Censoring indicator: $\delta_i = I(Y_i < Z_i)$		

More specifically, $x_{ij}^* = \sum_{k=1}^d r_{ki} \tau_{kj} + \epsilon_{ij}$, where $\{\tau_k = (\tau_{k1}, \dots, \tau_{kp})'\}_{k=1}^d$ are the components, $\{\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{ip})'\}_{i=1}^N$ are i.i.d. vectors of noise, and $\{r_{1i}, \dots, r_{di}\}$ is a set of fixed constants. We take the component values as $\tau_{kj} \sim N(\mu_\tau, \sigma_\tau^2)$ and the noise values as $\epsilon_{ij} \sim N(\mu_\epsilon, \sigma_\epsilon^2)$. (We denote the normal distribution with mean μ and variance σ^2 by $N(\mu, \sigma^2)$.) The matrix of predictor values are obtained as $x_{ij} = \exp(x_{ij}^*)$. Note that $x_{ij}^* \sim N(\mu_\tau \sum_{k=1}^d r_{ki}, \sigma_\tau^2 \sum_{k=1}^d r_{ki}^2 + \sigma_\epsilon^2) \equiv N(a_i, b_i^2)$. Thus, x_{ij} is distributed as log-normal with parameters a_i and b_i^2 , denoted as $x_{ij} \sim LN(a_i, b_i^2)$. A log-normal model has been used for microarray gene expression data [13]. Table 2A summarizes the simulation for the matrix of predictor values.

The expression matrix is generated with a mean noise of zero, $\mu_\epsilon = 0$, and with a component mean of $\mu_\tau = 5/d$. In order to make comparisons between dimension reduction methods that are not confounded by different model dimensions, we fix the number of components d . For the simulation we generated data from a model with $d = 6$ components.

In addition, the relative variance of the noise and component factors is controlled by the ratio of variance parameter $\delta = \sigma_\epsilon / \sigma_\tau$. As mentioned earlier, a predictor data matrix is generated so that the first K PCs explain a specified proportion of predictor variability. This is controlled by the simulation parameter δ . Data sets are generated so that the proportion of total predictor variability explained by the first K PCs, namely $\text{ave}(\lambda, K) \equiv \sum_{k=1}^K \lambda_k / p$, is between 0.40 and 0.70. For each percentage of variability explained, $100 \times \text{ave}(\lambda, K) \in \{40\%, 50\%, 60\%, 70\%\}$, and for each $p = 100, 300, 500, 800, 1000, 1200, 1400, \text{ and } 1600$, one hundred data sets were generated. (Details of simulation parameter settings, were previously given. See Table 3 in [13].) Furthermore, note that the data matrix is of size $N \times p$, where $p \gg N$. Since this is the case of interest in practice, we consider the number of predictors in the range of 100 to 1600, with the sample size, N , fixed at a small value of 50.

3.2. Generating observed survival times: the censored response

After generation of the expression matrix, $\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_N)$, we generated survival times satisfying the PH model (2). More precisely, we obtained survival times, Y_i , independent of censoring times, Z_i ($i = 1, \dots, N$), from the model

$$Y_i = Y_{0i} \exp(-\mathbf{x}'_i \boldsymbol{\beta}), \quad Z_i = Z_{0i} \exp(-\mathbf{x}'_i \boldsymbol{\beta}), \tag{8}$$

where β is a fixed coefficient vector. The observed, censored, response for individual i is obtained as $T_i = \min(Y_i, Z_i)$. The corresponding censoring indicator is $\delta_i = I(Y_i < Z_i)$.

We consider survival times with constant and increasing baseline hazard functions. For exponential (Exp) distributed survival times with density $f(t, \lambda) = (1/\lambda)\exp(-t/\lambda)$, the baseline hazard is constant: $h_0(t) = 1/\lambda$. For example, taking $Y_{0i} \sim \text{Exp}(\lambda_y)$ and $Z_{0i} \sim \text{Exp}(\lambda_z)$ for (8), we obtain survival data satisfying the PH model (2), since $h(t; \mathbf{x}_i; \beta) = 1/\{\lambda_y \exp(-\mathbf{x}'_i \beta)\} = h_0(t) \exp(\mathbf{x}'_i \beta)$. Also, the rate of censoring can be controlled by varying the parameters λ_z and λ_y since $\Pr(Y > Z) = \lambda_y/(\lambda_z + \lambda_y)$. For instance, taking $\lambda_y = 0.5$ and $\lambda_z = 1$ gives a censoring rate of $\Pr(Y > Z) = 1/3$.

Alternatively, for weibull (Weib) distributed survival times with density $f(t, \alpha, \lambda) = \alpha \lambda t^{\alpha-1} \exp(-\lambda t^\alpha)$, the baseline hazard function is increasing: $h_0(t) = \alpha \lambda t^{\alpha-1}$. To obtain survival data satisfying the PH model (2) with increasing baseline hazard, we can similarly take $Y_{0i} \sim \text{Weib}(\alpha_y, \lambda_y)$ and $Z_{0i} \sim \text{Weib}(\alpha_z, \lambda_z)$ for (8). The PH model (2) holds since $h(t; \mathbf{x}_i; \theta) = \alpha \lambda_y t^{\alpha_y-1} \{1/\exp(-\mathbf{x}'_i \beta)^\alpha\} = h_0(t) \exp(\mathbf{x}'_i \theta)$, where $\theta = \alpha_y \beta$.

The simulation procedure and model for survival times are summarized in Table 2B.

3.3. Examples

To illustrate the methods, namely PCA-PH, PLS-PH, and MPLS-PH, and the simulation experiments, we generated samples from a $d = 6$ components model (7) with $p = 100$ dimensions. Next, observed survival times were generated using (8) with exponential distributions and a censoring rate of 1/3: $Y_{0i} \sim \text{Exp}(1/2)$ and $Z_{0i} \sim \text{Exp}(3/2)$. For example, Fig. 2 displays three estimated survival curves $\hat{S}(t)$ evaluated at the average covariate vector $\bar{\mathbf{x}}$. These curves were obtained using the PH regression model (2) after dimension reduction from $p = 100$ to $K = 3$ dimensions using (a) PCA, (b) PLS, and (c) MPLS. The true survival distribution, $S(t)$, is also displayed in Fig. 2. Note that the estimated curve using PCA-PH is far above the true survival curve. However, the data was generated so that the first $K = 3$ PCs explained only 40% of the total predictor variability. To summarize the estimates from the various methods, relative to the true survival, let $s_i = S(t_i)$ denote the (true) probability of still surviving at time t_i . Denote its corresponding estimate by $\hat{S}(t_i)$, from PCA-PH, PLS-PH, or MLS-PH by \hat{s}_i . Using these notations, the (squared) Euclidean distance between the estimated and true vector of survival probability is

$$d^2(\mathbf{s}, \hat{\mathbf{s}}) = \|\mathbf{s} - \hat{\mathbf{s}}\|^2 = \sum_{i=1}^D (s_i - \hat{s}_i)^2, \quad (9)$$

where D is the number of observed survival time points. For example, the distances between the true and estimated survival curves, via PCA-PH, PLS-PH, and MPLS-PH in Fig. 2, are 0.3541, 0.0281, and 0.0467 respectively.

Other interesting comparisons, involving the quantiles of the survival distribution, can be made between two or more patient groups. The q th quantile, t_q ($0 < q < 1$), can be obtained by inverting $S(t_q) = q$. For instance, it is of interest to examine the estimated median survival times ($q = 0.50$) for patients in the GC-B like group relative to those in the activated B-like group for the diffuse large B-cell lymphoma study (see Fig. 1). Since the survival distribution in the simulated data example is $\text{Exp}(\lambda^*)$, the q th quantile can be explicitly obtained as $t_q = \lambda^* \log(q)$, where

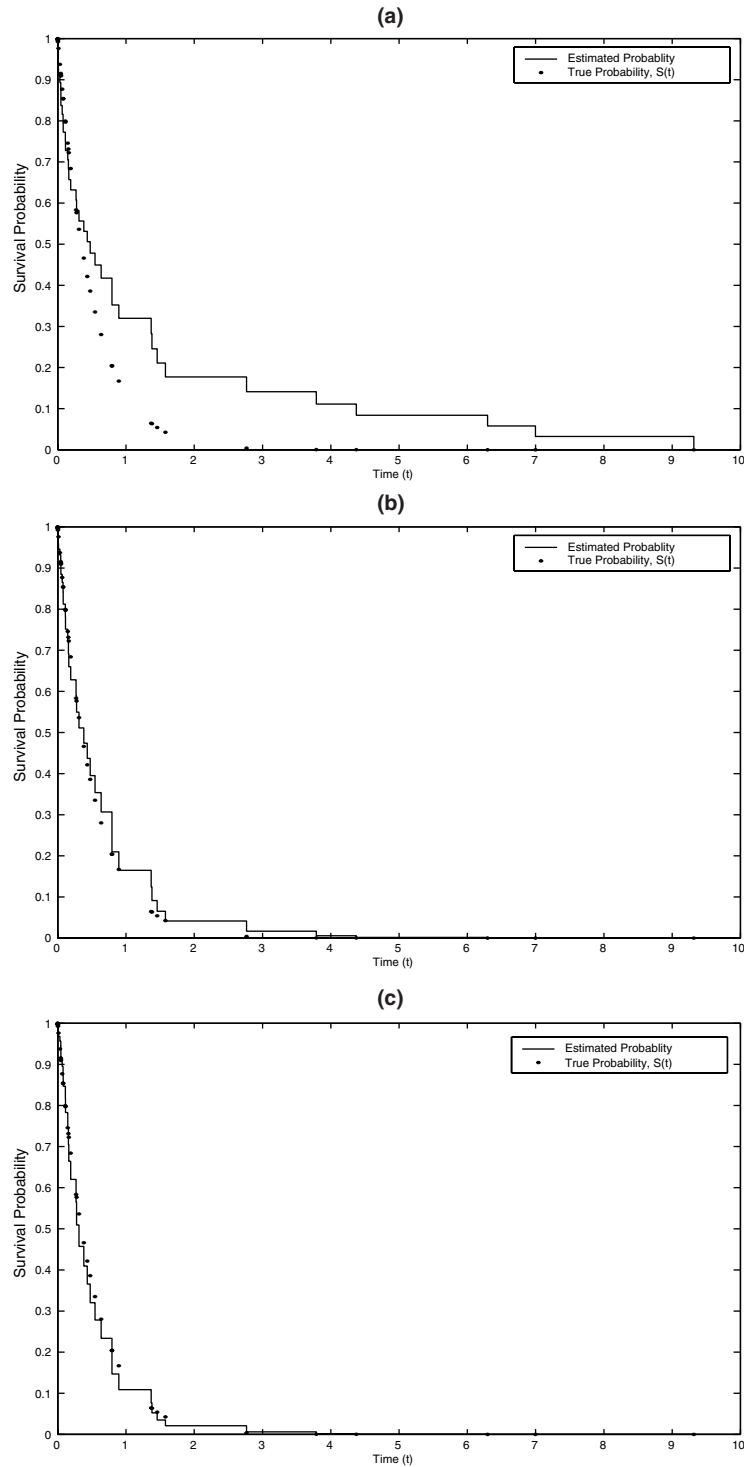


Fig. 2. Three estimated survival curves, $\hat{S}(t)$ (—), based on dimensionality reduction of $K=3$ from an original dimension of $p=100$. The dimension reduction methods are (a, top) PCA, (b, middle) PLS, and (c, bottom) MPLS. The true survival curve, $S(t)$ (***) is also given in each plot.

Table 3
Estimated and true quantile, t_q , for $q = 0.25, 0.50$, and 0.75

	Estimated q th quantile, t_q		
	25% ^a	50% ^a	75% ^a
PCA–PH	1.3794	0.4759	0.1134
PLS–PH	0.7952	0.3815	0.1467
MPLS–PH	0.6360	0.3116	0.1467
True quantile	0.6931	0.3466	0.1438

^a Dimension reduction.

$\lambda^* = \lambda \exp(\mathbf{x}'\boldsymbol{\beta})$. In practice, an estimate of t_q can be based on the estimated survival curve, \widehat{S} , and is given by

$$\hat{t}_q = \min\{t | \widehat{S}(t) \leq q\}. \quad (10)$$

For example, Table 3 gives the q th quantile, for $q = 0.25, 0.50$ and 0.75 , and their corresponding estimates from PCA–PH, PLS–PH, and MPLS–PH. From this one particular example (one simulation run), it appears that the PLS and modified PLS (MPLS) dimension reduction perform better than PCA. However, the general experimental results, which we will describe in the next section, indicate that this particular result does not hold unconditionally.

4. Experimental results

The first few principal components of real microarray data can explain a wide range of variability in the data. Thus, to flexibly cover a spectrum of real microarray data sets, the gene expression data matrix is generated such that first K principal components explain a specified proportion of predictor variability. We systematically varied this proportion to encompass a wide range of potential situations encountered in practice. In addition, we set the simulation size to reflect real microarray data, where $p \gg N$. More specifically, we simulated data with a relatively small sample size of $N = 50$ and increased the dimension from $p = 100$ to 1600. For a specified percentage of predictor variability explained (40%, 50%, 60%, and 70%) and dimension $p = 100, 300, 500, 800, 1000, 1200, 1400$, and 1600, one hundred data sets were generated. Thus, a total of 3,200 data sets were simulated. For each data set, we applied dimensionality reduction (in stage one) using PCA, PLS, and PLSM to extract $K = 3$ components. We obtained the estimated survival curve using the PH regression model (stage two) in the reduced subspace, as described in Section 2.1 and illustrated in Section 3.3.

We will describe the results in more detail for data sets where the percentage of total predictor variability accounted by a fixed number (K) of PCs is in the range (40%–70%). Table 4 gives the average observed proportion of total predictor variability explained by the first $K = 3$ PCs. Each value was obtained by averaging over the 100 simulated data sets (for each p and each targeted proportion of variation explained). The observed values are similar to the targeted percentage of total predictor variance explained (40%, 50%, 60%, and 70%) for each dimension p (see Table 4).

Table 4
Proportion of predictor variance explained by $K = 3$ principal components

p	Observed proportion			
100	0.43	0.53	0.60	0.73
300	0.41	0.51	0.58	0.72
500	0.43	0.51	0.58	0.73
800	0.40	0.50	0.59	0.70
1000	0.38	0.50	0.57	0.70
1200	0.42	0.48	0.58	0.78
1400	0.41	0.51	0.54	0.70
1600	0.40	0.48	0.58	0.73
	Target proportion			
	0.40	0.50	0.60	0.70

One hundred data sets were generated with $N = 50$ and p varying from 100 to 1,600. Each number given is the average (over 100 data sets) of the proportion of predictor variation accounted for by three PCs. The target total predictor variances explained by the K PCs are 40%, 50%, 60%, and 70%.

First, we examine the estimated survival curves based on (1) PCA, (2) PLS, and (3) MPLS dimension reduction. As illustrated in the example of Section 3.3, we compare each estimated survival curve to the true survival curve using the Euclidean distance measure (9). Since there are 100 generated data sets for each simulation configuration (i.e. for each p and targeted proportion of total predictor variance explained), we calculated the average distance between the estimated and true survival probabilities,

$$\text{ave}(d^2) \equiv 100^{-1} \sum_{b=1}^{100} d^2(\mathbf{s}_b, \hat{\mathbf{s}}_b) = 100^{-1} \sum_{b=1}^{100} \|\mathbf{s}_b - \hat{\mathbf{s}}_b\|^2. \tag{11}$$

For example, Fig. 3 compares $\text{ave}(d^2)$ for PCA, PLS, and MPLS dimension reduction. As expected, Fig. 3(a)–(d) shows that PCA improves ($\text{ave}(d^2)$ decreases) as the percentage of total predictor variance explained by K PCs increases. In addition, dimension reduction via PLS and MPLS performed substantially better than PCA when the total predictor variance explained is 40%–60%. However, when the total predictor variance is high (70%), all three methods performed similarly (see Fig. 3(d)). We also note that the ordinary PLS method, which treats the response variable as uncensored at the dimension reduction stage 1, performed slightly better than the modified PLS (MPLS) method when the predictor variance explained is 40%–50%.

Next, we compared the estimated quantile, \hat{t}_q , to the true quantile, t_q ($q = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, \text{ and } 0.9$). Similar to $\text{ave}(d^2)$, we computed the average estimated quantiles for each dimension reduction method. Fig. 4 compares these estimates with the corresponding true quantiles. For data sets where the percentage of total predictor variance accounted for is lower (40% and 50%), the estimated quantiles from PCA–PH is poor for smaller q ($q < 0.5$). However, as anticipated, PCA–PH improves as the total predictor variance explained increases to 60%. Furthermore, at 70%, PCA–PH estimates track the true quantiles well. This is true for PLS–PH as well. However, note that the performance of PLS–PH is good for all t_q ($q = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, \text{ and } 0.9$), even when the total predictor variance explained is lower. The performance

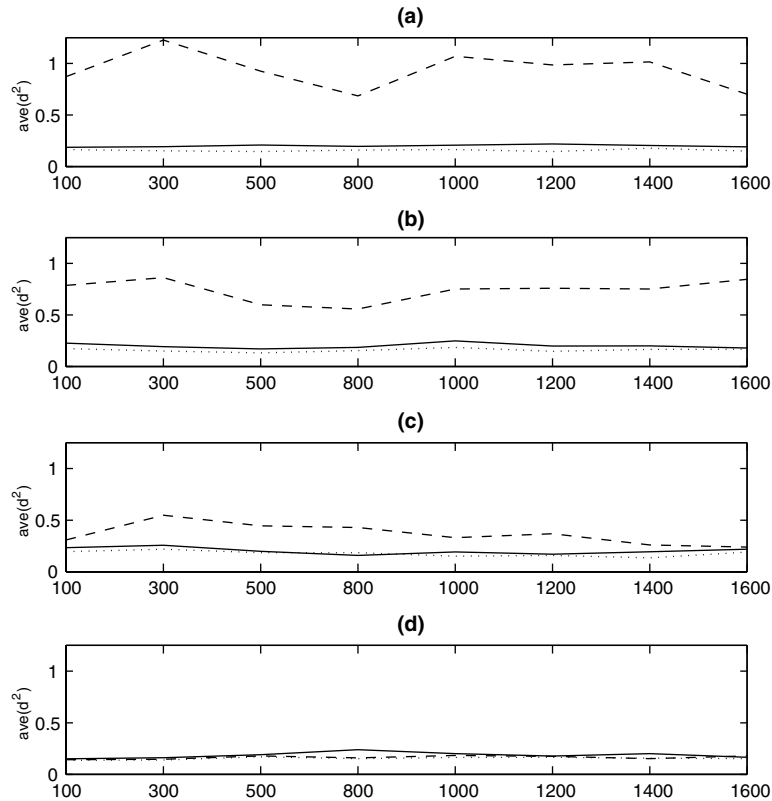


Fig. 3. 1/3 Censored. Averaged squared distance between the estimated and true survival probabilities, $\text{ave}(d^2)$, for data sets with approximately (a) 40%, (b) 50%, (c) 60%, and (d) 70% of total predictor variance accounted for by three PCs (PCA ---, PLS ···, and MPLS —).

of MPLS–PH is similar to PLS–PH, although it slightly over estimated the true quantiles when the total predictor variability accounted for is high (70%).

The relatively poor performance of PCA–PH, when the total predictor variance explained is lower, is not surprising. This is because the variance-objective criterion in PCA ignores response information completely. In addition, we did not select certain dimensions to be more predictive of survival time a priori. Thus, by design, the performance of PCA naturally improves as the total predictor variance explained increases. However, in practice, there is no guarantee that dimensions that explain a high percentage of predictor variance will be good predictors of a response [20,15]. Alternatively, the covariance-objective criterion in PLS aims to optimize both the predictor variance and the correlation between predictors and response. Thus, the performance does not solely depend on the amount of predictor variance explained. This is confirmed by the empirical performance of PLS, which remains similar as the predictor variance explained changes (e.g., see Fig. 4).

The results described thus far is based on exponential survival time and designed with a censoring rate of about 1/3. As stated in the Introduction Section, another aim of this study is to examine the affect of the amount of censoring on dimension reduction. Thus, we increased the

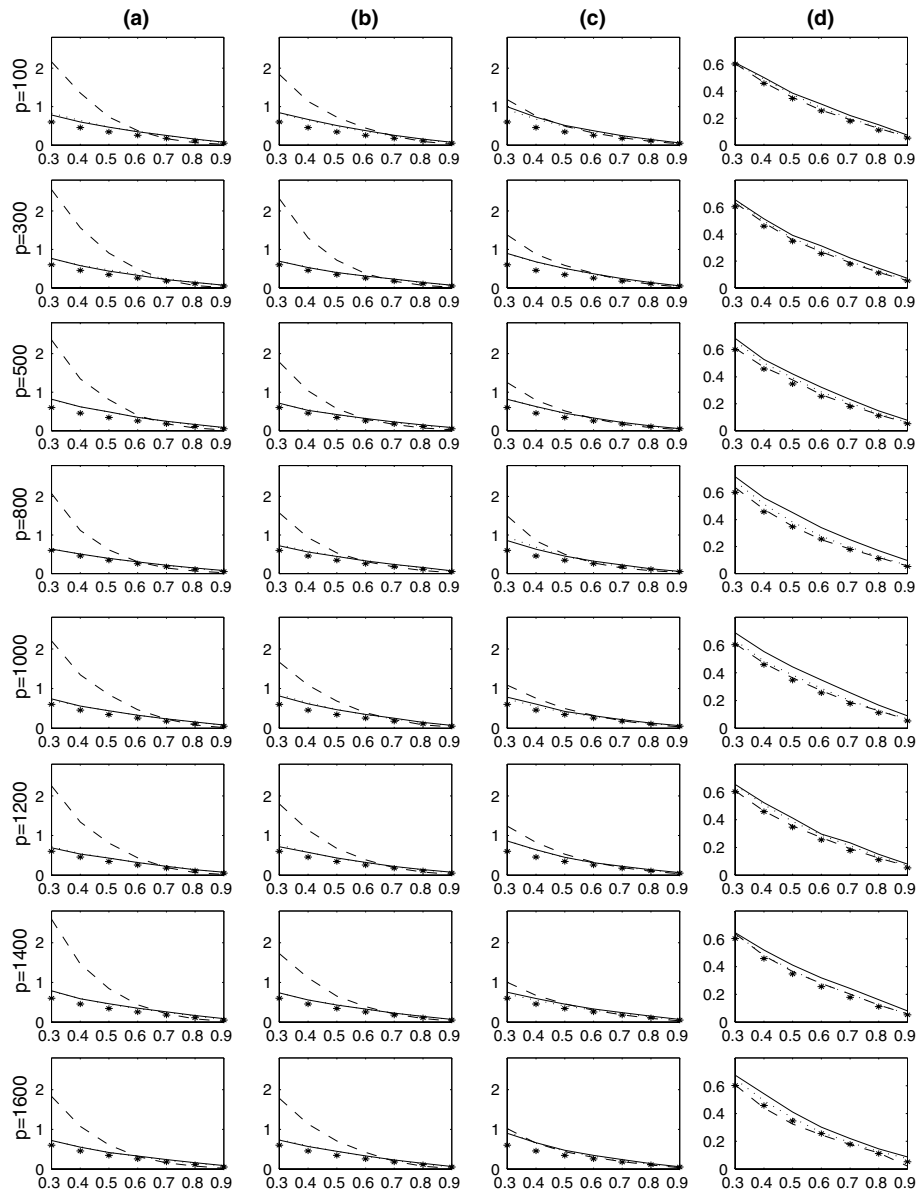


Fig. 4. $1/3$ censored. Given are the estimated quantiles, \hat{t}_q (y -axis), versus q (x -axis), based on PCA-PH (---), PLS-PH (···), and MPLS-PH (—). The true quantiles, t_q (***) are also given in each plot. The results are given for data sets with approximately (a) 40%, (b) 50%, (c) 60%, and (d) 70% of total predictor variability accounted for by three PCs. Each row of plots is for data sets with dimension $p = 100, 300, 500, 800, 1000, 1200, 1400, 1600$.

censoring rate to $1/2$; therefore, for a fixed sample size of $N = 50$, the effective sample size is now only 25 on average. This increase in censoring reflects the limited availability of samples in practice. The results for data with 50% censoring are given in Figs. 5 and 6. Generally, the pattern of the results is similar to the $1/3$ censoring case. However, due to the smaller effective sample size,

the performance (magnitude) deteriorates across all methods. Since PCA does not involve the censored response variable, censoring does not play a role in the dimension reduction step. Thus, although the overall magnitudes changed, the relative pattern of results did not. As before, PCA–PH improves as the total predictor variance increases. PLS–PH regression performs relatively better than PCA–PH and MPLS–PH. Also, estimates from PLS–PH track the true quantiles well (see Fig. 6). MPLS–PH appears to be most affected by the increased censoring. This is notable for the case where the total predictor variance explained is high (70%; Fig. 6(d)). In this case, MPLS–PH over estimates the true quantiles, while estimates from PCA–PH and PLS–PH are relatively closer to the true quantiles.

As mentioned earlier, the effective sample size is only 25 on average for a censoring rate of 1/2 and $N = 50$. Thus, estimating t_q for small values of q (less than 0.4 or 0.5) is not possible in some cases, since there may not be any long survival times observed in the data. For this reason only estimates of t_q for $q \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ are given in Fig. 6, omitting $q = 0.3$ and $q = 0.4$ in two cases. Furthermore, we reported results based only on exponential survival times, which implies a constant baseline hazard. Results for increasing baseline hazard (weibull survival times)

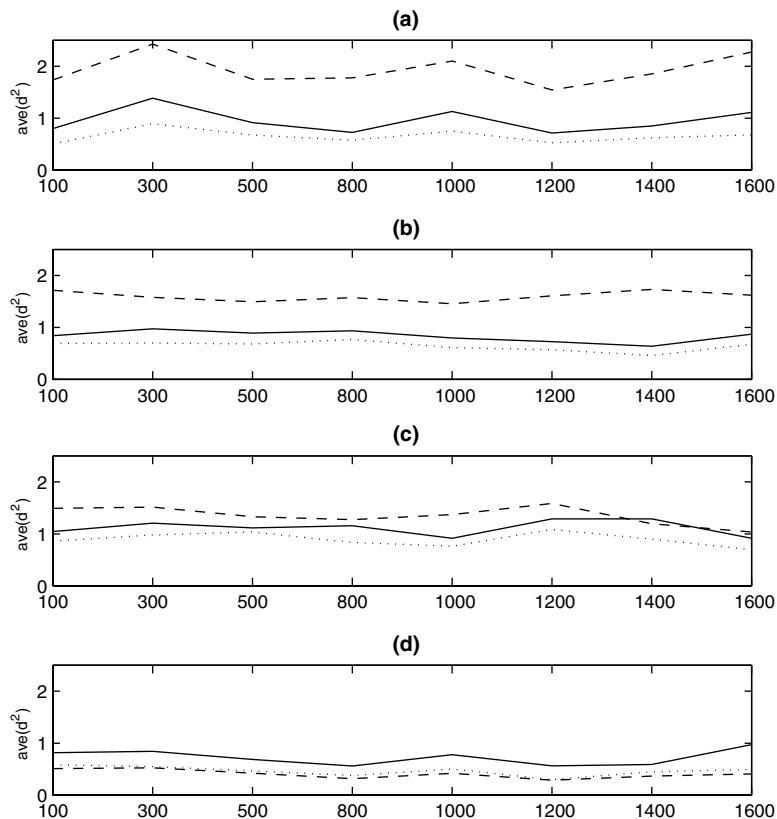


Fig. 5. 1/2 censored. Averaged squared distance between the estimated and true survival probabilities, $\text{ave}(d^2)$, for data sets with approximately (a) 40%, (b) 50%, (c) 60%, and (d) 70% of total predictor variance accounted for by three PCs (PCA ---, PLS ···, and MPLS —).

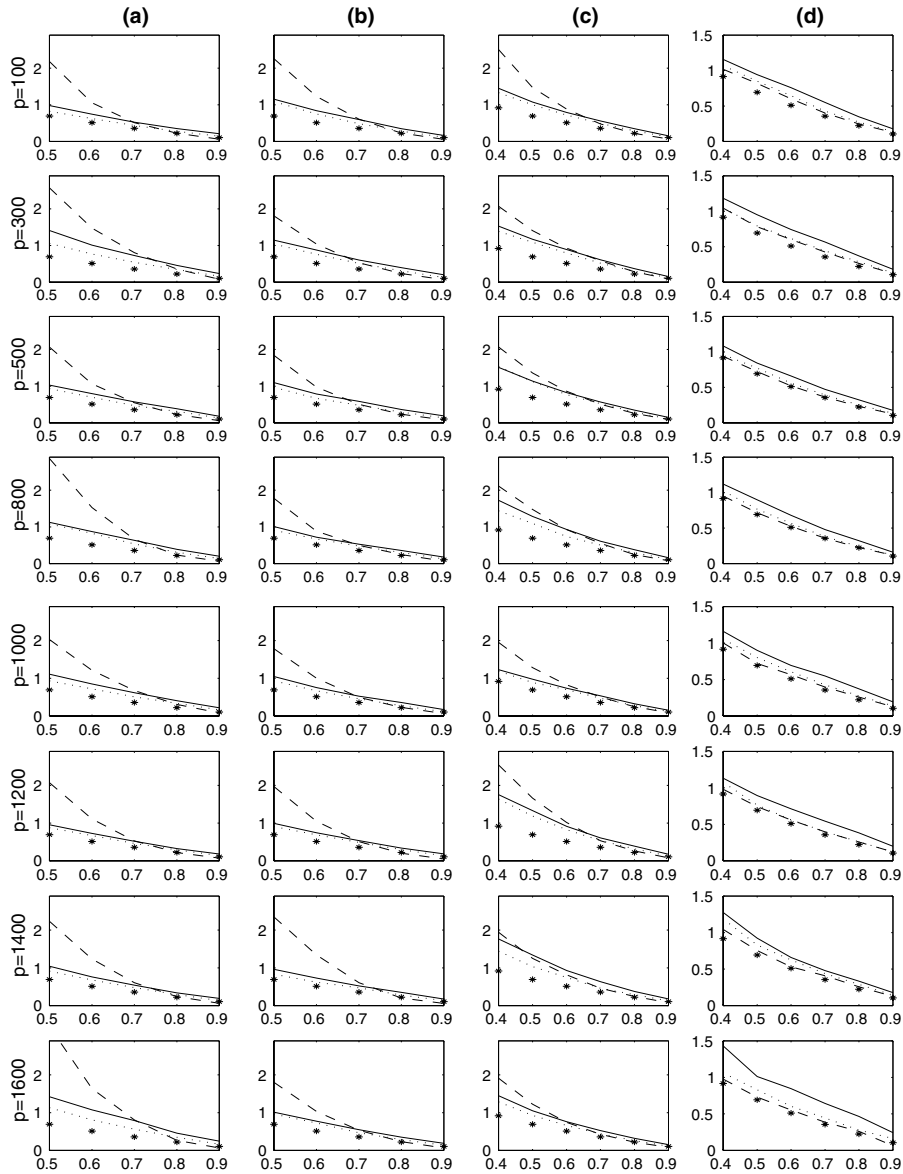


Fig. 6. 1/2 censored. Given are the estimated quantiles, \hat{t}_q (y -axis), versus q (x -axis), based on PCA-PH (---), PLS-PH (···), and MPLS-PH (—). The true quantiles, t_q (***), are also given in each plot. The results are given for data sets with approximately (a) 40%, (b) 50%, (c) 60%, and (d) 70% of total predictor variability accounted for by three PCs. Each row of plots is for data sets with dimension $p = 100, 300, 500, 800, 1000, 1200, 1400,$ and 1600 .

are similar and were omitted. However, this is expected since estimates from the PH regression model do not depend on the specific baseline hazard. It only requires that the data, $\{(T_i, \delta_i, \mathbf{x}_i)\}_{i=1}^N$, satisfies the PH model (2) and the simulation was designed to meet the PH model assumption.

5. Conclusion

In this paper we have provided a study of the performance of the PLS–PH regression method, previously proposed by Nguyen and Rocke [11,12]. More specifically, we have extended a simulation model for gene expression data with a censored response variable and examined the performance of the PH regression model under three dimension reduction techniques: PLS, PCA, and a modified PLS (MPLS) technique. Our conclusions, based on the simulation study, are as follow.

- PLS–PH and MPLS–PH outperform PCA–PH overall and perform substantially better when the total predictor variance explained (TPVE) is in the range 40%–60%.
- PLS–PH performs slightly better than MPLS–PH and is best overall.
- As the TPVE increases PCA–PH improves, as expected, and all methods perform similarly when the TPVE is high (e.g. at 70%).
- When the censoring rate is high (e.g. at 50%) the performance of the dimension reduction methods deteriorates, although the relative patterns of performance among the methods are similar.
- MPLS–PH appears to be most affected by the high censoring, overestimating survival time at high TPVE (70%).

The results suggests that the PLS–PH method works well, despite ignoring information on censoring at the dimension reduction stage 1.

Acknowledgement

We are gratefully to two reviewers for their detailed and constructive comments. The research reported here was partially supported by an American Cancer Society IRG program grant, through the UC Davis Cancer Center.

Appendix A. MPLS Calculation

The k th MPLS component is computed directly as a linear combination of the expression matrix \mathbf{X} . That is, $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k$. The weight vector can be computed as $\mathbf{w}_k = \sum_{i=1}^n \theta_{ik} \mathbf{v}_i$, where \mathbf{v}_i is the i th eigenvector of $\mathbf{X}'\mathbf{X}$. Thus, to compute the MPLS components, it remains to compute the constants $\{\theta_{ik}\}$. Close form expressions for these constants were previously published (Theorem 3.2 in Nguyen and Rocke [13]). We refer the reader there for the exact expression for $\{\theta_{ik}\}$. However, as mentioned earlier, the key is that these constants only involve the censored response values $\{y_i\}$ through the dot product $a_i = \mathbf{u}_i' \mathbf{y}$ (where \mathbf{u}_i is the i th eigenvector of $\mathbf{X}\mathbf{X}'$). Furthermore, a simple linear regression of the response Y on U_i will have estimated slope coefficient $\mathbf{y}'\mathbf{u}_i / (\mathbf{u}_i'\mathbf{u}_i)$. When the gene expression matrix is centered, $\mathbf{u}_i'\mathbf{u}_i = 1$. Hence, the dot product a_i is precisely the slope of the simple linear regression of Y on U_i . However, for a censored response variable, the use of a_i from simple linear regression has some drawbacks. For instance, it ignores the fact that a proportion of the response values are censored. Thus, we can incorporate information on censoring at the dimension reduction stage by taking a_i to be the estimated coefficient from the PH regression

model of survival time Y on U_i , instead of the simple linear regression slope coefficient. Additional details can also be found in [13].

References

- [1] D.V. Nguyen, A.B. Arpat, N. Wang, R.J. Carroll, DNA microarray experiments: biological and technological aspects, *Biometrics* 58 (2002) 701.
- [2] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Brolrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* 403 (2000) 503.
- [3] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P. Lonning, A.L. Borresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc. Natl. Acad. Sci. USA* 98 (2001) 10869.
- [4] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, T.R. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* 415 (2002) 436.
- [5] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nat. Med.* 8 (2002) 68.
- [6] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K.J. Pienta, M.A. Rubin, A.M. Chinnaiyan, Delineation of prognostic biomarkers in prostate cancer, *Nature* 412 (2001) 822.
- [7] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell* 1 (2002) 203.
- [8] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proc. Natl. Acad. Sci. USA* 98 (2001) 13690.
- [9] D.R. Cox, Regression models and life-tables (with discussion), *J. Roy. Statist. Soc. Series B* 34 (1972) 187.
- [10] D.W. Hosmer, S. Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley & Sons, New York, 1999.
- [11] D.V. Nguyen, D.M. Rocke, Assessing patient survival using microarray gene expression data via partial least squares proportional hazard regression, *Comput. Sci. Stat.* 33 (2001) 376.
- [12] D.V. Nguyen, D.M. Rocke, Partial least squares proportional hazard regression for application to DNA microarray survival data, *Bioinformatics* 18 (2002) 1625.
- [13] D.V. Nguyen, D.M. Rocke, On partial least squares dimension reduction for microarray-based classification: a simulation study, *Comput. Stat. Data Anal.* 46 (2004) 407.
- [14] W.F. Massey, Principal components regression in exploratory statistical research, *J. Am. Stat. Associat.* 60 (1965) 234.
- [15] I.T. Jolliffe, *Principal Component Analysis*, Springer, New York, 1986.
- [16] K.V. Mardia, J.T. Kent, J.M. Bibby, *Multivariate Analysis*, Academic Press, San Diego, 1979.
- [17] A. Höskuldsson, PLS regression methods, *J. Chemometr.* 2 (1988) 211.
- [18] S. De Jong, SIMPLS: An alternative approach to partial least squares regression, *Chemometr. Intell. Lab. System.* 18 (1993) 251.

- [19] D.V. Nguyen, D.M. Rocke, Classification of acute leukemia based on DNA microarray gene expressions using partial least squares, in: S.M. Lin, K.F. Johnson (Eds.), *Methods of Microarray Data Analysis*, Kluwer, Dordrecht, 2002, p. 109.
- [20] D.V. Nguyen, D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* 18 (2002) 39.
- [21] D.V. Nguyen, D.M. Rocke, Multi-class cancer classification via partial least squares using gene expression profiles, *Bioinformatics* 18 (2002) 1216.
- [22] J.D. Kalbfleisch, R.L. Prentice, Marginal likelihoods based on Cox's regression and like model, *Biometrika* 60 (1973) 267.