Journal of Data Science **5**(2007), 183-198

Principal Component Analysis in Linear Regression Survival Model with Microarray Data

Steven Ma
University of Washington

Abstract: As a useful alternative to the Cox proportional hazards model, the linear regression survival model assumes a linear relationship between the covariates and a known monotone transformation, for example logarithm, of an event time of interest. In this article, we study the linear regression survival model with right censored survival data, when high-dimensional microarray measurements are present. Such data may arise in studies investigating the statistical influence of molecular features on survival risk. We propose using the principal component regression (PCR) technique for model reduction based on the weight least squared Stute estimate. Compared with other model reduction techniques, the PCR approach is relatively insensitive to the number of covariates and hence suitable for high dimensional microarray data. Component selection based on the nonparametric bootstrap, and model evaluation using the time-dependent ROC (receiver operating characteristic) technique are investigated. We demonstrate the proposed approach with datasets from two microarray gene expression profiling studies of lymphoma cancers.

Key words: Linear regression model, microarray, principal component regression, survival analysis.

1. Introduction

Microarray technologies that are capable of monitoring tens of thousands of gene expression profiles simultaneously have been extensively used in medical and biological studies. Our research is partly motivated by biomedical experiments like the well known gene expression profiling study for diffuse large B-cell lymphoma (DLBCL) reported by Alizadeh *et al.* (2000), where gene expression profiles of 4026 clones and survival information for 40 patients are recorded. One main goal of the DLBCL study is to identify the statistical influence of tumor molecular features on survival risk. A better understanding of the molecular biology that underlies variations of phenotype among subjects may provide a more accurate and rational method of risk stratification to guide treatment and may

suggest new therapeutic approaches as well. Classification and prediction of occurrence of cancer using microarray data have been shown to be successful. See Alon *et al.* (1999), Golub *et al.* (1999) and Garber *et al.* (2001) among many others for reference. Due to presence of censoring and usage of more complicated semiparametric or nonparametric models, survival analysis using microarray data has been less investigated. It is thus of special interest to develop sound statistical methodologies that can effectively use high dimensional microarray measurements in survival analysis.

Recent studies of right censored survival data with high dimensional microarray measurements include, but are not limited to, the following. An ad hoc approach suggested by Alizadeh et al. (2000) with microarray data is to cluster genes first, and then use the within cluster averages of the gene expression levels in the Cox model. Another well developed clustering based algorithm is the gene harvesting procedure of Hastie *et al.* (2001). Nguyen and Rocke (2002) apply the standard partial least squares (PLS) method to survival data and use the resulting PLS components in the Cox model. Li and Luan (2003) develop a penalized estimation procedure for the Cox model using kernels, under the assumption that the covariate effects are smooth functions of gene expression levels. General penalization methods have also been developed for the Cox model (Fan and Li, 2002). Empirical studies show that performances of different approaches are data dependent, with no approach dominating the others.

Despite the extensive study of the Cox model and the additive risk model, research on the linear regression survival model remains rare for right censored survival data with high dimensional microarray measurements. The linear regression model assumes a linear relationship between the covariates and a known monotone transformation, for example logarithm, of a failure time of interest (Buckley and James, 1979; Ying, 1993). Since the event time, instead of conditional risk function, is modeled directly, the linear regression survival model can be more interpretable under certain circumstances and more suitable for prediction of survival time. See Wei (1992) for an illuminating discussion. In this article, we propose using the PCR (principal component regression) technique for dimension reduction with the linear regression model. Properties of PCR estimators have been extensively investigated for simple linear regressions (Jolliffe, 1986; Kollo and Neudecker, 1993). Compared with other dimension reduction techniques, the PCR estimators are computationally less sensitive to the number of covariates, easier to compute using existing software even for high dimensional data and their theoretical properties are more transparent. Hence the PCR method can be more suitable for model reduction with high dimensional microarray data.

The goal of this paper is to develop theoretically well-behaved and compu-

tationally stable estimators for the linear regression model with right censored survival data and microarray measurements. The article is organized as follows. In section 2, we define the linear regression model and corresponding PCR estimators. Inference and model evaluation are investigated in the same section. In section 3, we present analysis of the Mantle cell lymphoma data and the Follicular lymphoma data. Concluding remarks are in section 4.

2. Principal Component Regression in Linear Regression Survival Model

2.1 Data steeing

Let T_i be the logarithm or a known monotone transformation of the failure time and X_i a d-dimensional covariate vector for the ith subject in a random sample of size n. Here T_i may denote the (transformed) time to death due to cancer or time to occurrence of cancer. Since transformation of the original time is used here, the "time" T_i may be negative. X_i denotes the gene expression profiles. Without loss of generality, we assume the logarithm transformation of the event time hereafter. The linear regression survival model assumes

$$T_i = \alpha + X_i'\beta + \epsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

where α is the intercept, $\beta \in \mathbb{R}^d$ is the regression coefficient and ϵ_i is the error term. When T_i is subject to right censoring, we can only observe (Y_i, δ_i, X_i) with $Y_i = \min\{T_i, C_i\}$, where C_i is the logarithm of the censoring time and $\delta_i = 1_{\{T_i \le C_i\}}$ is the censoring indicator. Suppose that a random sample $(Y_i, \delta_i, X_i), i = 1, \ldots, n$ with the same distribution as (Y, Δ, X) is available.

The model (2.1) assumed here shares the same format as the AFT (accelerated failure time) model in Buckley and James (1979). Two semiparametric methods have received special attention for analyzing such model. One is the Buckley-James estimator (Buckley and James, 1979) which adjusts censored observations using the Kaplan-Meier estimator. The other is the rank based estimator which is motivated by the score functions of the partial likelihood (Wei, Ying and Lin, 1990). However, the linear regression survival model has not been widely used in practice, mainly due to the difficulties in computing the semiparametric estimators of the aforementioned methods, even in situations when the number of covariates is relatively small (Jin, Lin, Wei and Ying, 2003).

The Stute estimator (1999) uses the Kaplan-Meier weights to account for censoring and the objective function has a simple least squares format. This simple objective function makes PCR natural with the Stute estimator, as can be seen in section 2.2. We note that this simplicity of the objective function is not shared by the Buckley-James estimator and the rank based estimator for

the linear regression survival model or the Cox model, which needs iterative maximization of a weighted objective function.

The Stute approach can be summarized as follows. It is first assumed that X_i are iid distributed. Let $\widehat{F}_n$ be the Kaplan-Meier estimator of the unconditional distribution function F of T. Following Stute (1999), $\widehat{F}_n$ can be written as $\widehat{F}_n(y) = \sum_{i=1}^n w_{ni} 1\{Y_{(i)} \le y\}$, where w_{ni}'s are the Kaplan-Meier weights defined as

$$w_{n1} = \frac{\delta_{(1)}}{n}, \text{ and } w_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \ i = 2, \ldots, n.$$

Here $Y_{(1)} \le \cdots \le Y_{(n)}$ are the order statistics of Y_i's and $\delta_{(1)}, \ldots, \delta_{(n)}$ are the associated censoring indicators. Similarly, let $X_{(1)}, \ldots, X_{(n)}$ be the associated covariates of the ordered Y_i's. Stute (1999) proposed the weighted least squares estimator $(\hat{\alpha}, \hat{\beta})$ that minimizes $M(\alpha, \beta) = \frac{1}{2} \sum_{i=1}^n w_{ni} (Y_{(i)} - \alpha - X'_{(i)}\beta)^2$. Under certain mild regularity conditions, Stute (1999) proved that $(\hat{\alpha}, \hat{\beta})$ is $\sqrt{n}$ consistent and asymptotically normal as $n \to \infty$ for a fixed d. Let

$$\bar{X}_{wi} = \frac{\sum_{i=1}^n w_{ni} X_{(i)}}{\sum_{i=1}^n w_{ni}}, \quad \bar{Y}_{wi} = \frac{\sum_{i=1}^n w_{ni} Y_{(i)}}{\sum_{i=1}^n w_{ni}}. \tag{2.2}$$

To obtain a simplified format of the objective function, we replace $X_{(i)}$ and $Y_{(i)}$ in $M(\alpha, \beta)$ with

$$w_{ni}^{1/2} (X_{(i)} - \bar{X}_{wi}) \text{ and } w_{ni}^{1/2} (Y_{(i)} - \bar{Y}_{wi}), \tag{2.3}$$

respectively. For simplicity, we still use $X_{(i)}$ and $Y_{(i)}$ in $M(\alpha, \beta)$ to denote the weighted centered values. Using the weighted centered values, the intercept estimate $\hat{\alpha}$ is zero. So the weighted least squared objective function can be written as

$$M(\beta) = \frac{1}{2} \sum_{i=1}^n (Y_{(i)} - X'_{(i)}\beta)^2. \tag{2.4}$$

The objective function $M(\beta)$ in (2.4) takes a least squared format, which is easier to compute compared with the rank based estimates as in Wei, Ying and Lin (1990). This simple form also motivates using the PCR for model reduction with high dimensional microarray measurements. Compared with other model reduction methods, the PCR approach only involves simple matrix calculations and is relatively insensitive to the number of covariates. The PCR estimate is also easy to obtain using existing software, even for the $d >> n$ microarray data.

2.2 PCR estimate

Consider the following principal component regression approach. Denote $\mathbb{X}$ as the $n \times d$ matrix composed of $X_{(i)}s$ and $\mathbb{Y}$ as the length n vector composed of $Y_{(i)}s$. The estimate $\hat{\beta}$ defined as the minimizer of (2.4) satisfies $\{\mathbb{X}'\mathbb{X}\}\hat{\beta} = \mathbb{X}'\mathbb{Y}$. Since $\mathbb{X}'\mathbb{X}$ is a semi-positive-definite matrix, there exists a d dimensional square matrix P satisfying

$$P'\mathbb{X}'\mathbb{X}P = M = diag(m_1, m_2, \ldots, m_k, 0, \ldots, 0) \quad \text{and} \quad PP' = I_d. \tag{2.5}$$

Here I_d denotes the d-dimensional identity matrix. k is the rank of $\mathbb{X}'\mathbb{X}$. For microarray data with $n << d$, we have $k < d$. P is composed of eigenvectors of $\mathbb{X}'\mathbb{X}$, and the m_is correspond to eigenvalues of $\mathbb{X}'\mathbb{X}$.

So we have $P'\mathbb{X}'\mathbb{X}PP'\hat{\beta} = P'\mathbb{X}'\mathbb{Y}$ and $MP'\hat{\beta} = P'\mathbb{X}'\mathbb{Y}$. If we denote $P'\hat{\beta} = \hat{\gamma}$ and $M^G = diag(1/m_1, \ldots, 1/m_k, 0, \ldots, 0)$, it can be seen that one special solution to the Stute estimate when $n << d$ is $\hat{\gamma} = M^G P'\mathbb{X}'\mathbb{Y}$ and $\hat{\beta} = P\hat{\gamma}$.

Empirical studies show for small to medium sample size cases, when d is comparable to or larger than n, some components of $\hat{\gamma}$ can have estimated variances several orders larger than the other components, which indicates unstable estimates. This poses especially serious concerns for analysis of microarray data, which usually have $n < 100$ and $d \sim 10^3$ or more. This phenomenon motivates using the PCR to yield more reliable estimators by excluding certain principal components from the regression. This stability arises from the well known bias-variance tradeoff, as has been noticed for linear regression by Jolliffe (1986). Related discussions can also be found in Huang and Harrington (2004).

Denote S as the component-selection matrix with certain diagonal elements equal to 1 and all other elements equal to 0. For example, if only the principal components corresponding to the first p elements of $\hat{\gamma}$ are selected, then $S = diag(I_p, 0)$, where I_p denotes the p-dimensional identity matrix. For now, we assume the matrix S is known. Determination of S is postponed to section 2.3. The PCR estimator can then be defined as $\hat{\gamma}_{pc} = S\hat{\gamma}$ and $\hat{\beta}_{pc} = P\hat{\gamma}_{pc}$.

Under mild regularity conditions, we can establish the asymptotic bias of the PCR estimate $\hat{\beta}_{pc}$ assuming finite d and $n \to \infty$. It can also be shown that $\hat{\gamma}$ is asymptotically normal distributed. The proof is omitted here and is available upon request from the author. The asymptotic normality of $\hat{\gamma}$, combined with the nonparametric bootstrap proposed in section 2.3, can be used to determine the component selection matrix S via hypothesis testing for significant components.

Principal component regression has been used in a wide range of biomedical problems, including the analysis of microarray data in search of outliers genes as well as the analysis of other types of expression data (Raychaudhuri *et al.* 2000). When genes are used as variables, the PCR creates a set of principal gene components, also known as super genes (Lan *et al.*, 2003), that indicate the features of

genes that best explain the experimental responses they produce. Compared with penalization based methods where effects of single genes are identified, explanation of the PCR estimates may not be straightforward. However, if prediction and classification are of main interest, this limitation is not serious.

2.3 Principal component selection

We propose selecting principal components based on marginal significance of $\hat{\gamma}$. At this point, it is not clear how to develop plug-in estimate for the variance of $\hat{\gamma}$. As an alternative, we consider the following nonparametric bootstrap, which was investigated in general by Efron and Tibshirani (1993).

First we sample $n' = 0.632n$ subjects from the n observations without replacement. Then the PCR estimates for the bootstrap samples are constructed in the same manner as proposed in section 2.2 with the component selection matrix $S = I_d$. Denote the bootstrap PCR estimate of γ as $\tilde{\gamma}$. The sampling and the estimation procedures are repeated many, for example 1000, times. Then after proper scale adjustment, the sample variance of $\tilde{\gamma}$ provides a fair estimate of the variance of $\hat{\gamma}$. Marginal $z-$scores and p-values can then be obtained based on the bootstrap variance estimates and the asymptotic normality of $\hat{\gamma}$. We use $n' = 0.632n$ since the expected number of distinct bootstrap observations is about $0.632n$. Computationally, it may be more efficient to use a smaller bootstrap sample size.

The cutoff for identifying important principal components can be based on the marginal p-values. Note that the dimension k of the principal components set is limited by $min(n, d)$. Empirical studies show that k can be much smaller than $min(n, d)$. In our study, we propose using the simple Bonferroni method (Johnson and Wichern, 1998) to account for multiple comparison adjustment. When the dimension of the principal components set is high, other techniques, for example the false discovery rate method (Benjamini and Hochberg, 1995), can be used.

When $n >> d$, it is expected that the validity of the nonparametric bootstrap can be proved following the general statements in Politis and Romano (1994). Simulation studies (not shown here) support the validity of the nonparametric bootstrap when $n >> d$. It is still unclear at this point whether similar arguments still hold under the current data setting with $n << d$. Limited empirical studies show that the nonparametric bootstrap variance estimates are well-behaved.

2.4 Model evaluation

In standard survival analysis, the focus is to assess the association between individual covariates with the censored survival outcome. However, when the

sample size is smaller than or comparable to the number of covariates, this standard approach of assessing significance may not be appropriate, since its validity typically relies on justifications assuming $n >> d$. Empirical studies show that when $n << d$, it is usually hard to correctly estimate individual covariate effects (Ghosh and Chinnaiyan, 2005). Our own simulation study for the linear regression survival model and the proposed approach supports Ghosh and Chinnaiyan's statement. In our study, the sample distribution of the PCR estimate $\hat{\beta}_{pc}$ is not clear. More importantly, the standard approach does not directly address the problem of prediction performance. Unlike in standard survival analysis where the association between survival outcome and covariates is of primary interest, the main goal of our study is to predict survival risk based on the PCR estimate. We consider the following approaches for assessing the performance of the proposed approach.

Consider the linear risk scores $X'\hat{\beta}_{pc}$. From model (2.1), we can see that smaller linear risk scores indicate on average smaller event times and hence higher survival risks. So a simple model evaluation procedure is as follows. First, we generate two hypothetical risk groups based on the PCR risk scores $X'\hat{\beta}_{pc}$ in a manner that there are equal number of subjects in the two risk groups. The empirical survival functions are then computed for the two risk groups. Better fitted models will yield more significantly different survival functions, and the difference of the survival functions can be measured by the simple logrank statistic and its corresponding p-value (Fleming and Harrington, 1991).

As an alternative, we also employ the time-dependent ROC (receiver operating characteristic) method for censored data approach. The time-dependent ROC technique was firstly proposed by Heagerty *et al.* (2000) in the context of the medical diagnosis and has been used as criteria for censored data regression with microarray gene expression data (Gui and Li, 2005). The essential idea is to treat the event indicator as binary outcome for each time point and evaluate the classification performance at each time using the standard ROC technique. In the ROC approach, the AUC (area under curve) can be used as the evaluation/comparison criteria and a larger AUC at time t indicates better predictability of the survival outcome at time t as measured by sensitivity and specificity evaluated at time t.

3. Examples

3.1 Mantle cell lymphoma data

Rosenwald *et al.* (2003) reported a study using microarray expression analysis of mantle cell lymphoma (MCL). One of the goals of this study is to discover gene expression signatures that correlate with survival in MCL patients. Among 101 untreated patients with no history of previous lymphoma included in this study, 92 were classified as having MCL, based on established morphologic and

Table 1: Mantle cell lymphoma data: the 30 genes with the largest absolute values of $\hat{\beta}_{pc}$.

UNIQID	Gene name	$\hat{\beta}_{pc} \times 10^2$
15981	Hs.524214, Myeloid leukemia factor 2	-2.568
16089	Hs.227817, BCL2-related protein A1	2.556
16541	Hs.30054, Coagulation factor V	2.765
16847	Hs.517717, Special AT-rich sequence binding protein 1	2.073
17322	Hs.428027, Pre-B-cell leukemia transcription factor 3	-1.900
17901	Hs.88218, Prepronociceptin	-2.023
23972	Hs.431009, Zinc finger protein, multitype 2	1.734
24084	Hs.467769, Family with sequence similarity 49, member A	-1.760
24379	Hs.120260, Immunoglobulin superfamily receptor	1.923
24880	Hs.459909, Transcribed locus	2.044
26192	Hs.530274, Aldolase B, fructose-bisphosphate	1.888
27067	–	1.742
27108	Hs.375108, CD24 antigen (small cell lung carcinoma cluster 4)	2.305
27678	–	2.340
27831	Hs.87205, Lymphocyte antigen 64 homolog, radioprotective 105kDa	-2.217
27838	Hs.87205, Lymphocyte antigen 64 homolog, radioprotective 105kDa	-2.626
27839	Hs.87205, Lymphocyte antigen 64 homolog, radioprotective 105kDa	-2.451
28216	Hs.484703, CD83 antigen (activated B lymphocytes)	1.858
28494	Hs.24529, CHK1 checkpoint homolog (S. pombe)	-2.100
28638	Hs.227817, BCL2-related protein A1	2.197
29286	Hs.118651, Hematopoietically expressed homeobox	-2.999
29791	–	-1.829
30130	Hs.368433, Tumor protein D52	-2.608
30596	Hs.120260, Immunoglobulin superfamily receptor	1.881
31219	Hs.118351, Ubiquitin protein ligase E3C	-2.415
31298	Hs.79347, Zinc finger protein 592	2.527
31837	Hs.436093, HLA-B associated transcript 2	-2.266
32249	Hs.494997, Complement component 5	-1.870
32874	Hs.363744, Transcribed locus	1.764
33831	Hs.370603, Tetratricopeptide repeat domain 7A	-3.178

immunophenotypic criteria. Survival times of 64 patients were available and other 28 patients were censored. The median survival time was 2.8 years (range 0.02 to 14.05 years). Lymphochip DNA microarrays (Alizadeh *et al.*, 2000) were used to quantify mRNA expression in the lymphoma samples from the 92 patients. The

gene expression data set that contains expression values of 8810 cDNA elements is publicly available and can be downloaded from *http://llmpp.nih.gov/MCL.* In Rosenwald *et al.* (2003), clustering based method is used for data analysis, assuming the Cox model. The underlying assumption is that all genes within the same cluster contribute additively and equally to the risk of survival, which is not realistic. As an alternative, we assume the linear regression survival model and apply the PCR approach to this dataset.

The PCR approach has no computational or methodological limitation on the number of genes that can be used in the prediction of patients' failure times. To gain further stability, we pre-process the genes as follows: (1). Fill in missing expression values with sample means; (2). Compute correlation coefficients of the uncensored survival times with gene expressions; (3). For each gene, compute the maximum and minimum of expression values across all the sample. Compute the differences between the maximum and minimum values; (4). Select the genes whose correlation with survival time is greater than 0.3 and the difference between the maximum and minimum is greater than 2.5. 364 genes pass the above selection criterion. We make the log transformation to the observed time and standardize the 364 selected genes to have mean 0 and variance 1. Similar gene pro-precessing has been proposed and discussed in detail in Dudoit *et al.* (2002). Since the number of the covariates (364) is larger than the sample size (92), Stute weighted least squared estimate is not unique.

We consider applying the proposed PCR approach to the MCL data. Using the nonparametric bootstrap and the Bonferroni adjustment, five principal components are significant at the 0.05 level. The final PCR estimate is constructed using those significant components only. In Table 1, we list the 30 genes with the largest absolute values of $\hat{\beta}_{pc}$. Roughly speaking, since all genes have been normalized to unit variance, the estimates are directly comparable: a larger estimated coefficient indicates stronger influence on survival. We also note that a direct evaluation of the influence of individual genes is not available. This is the inherent drawback associated with the PCR approach.

In Figure 1, we show the survival functions for the two risk groups defined using the PCR estimate. We can see that the difference of the survival functions are obvious (p-value< 0.001), which suggests that the proposed approach is effective in predicting survival risks based on gene expression profiles. We also show in Figure 1 the AUC as a function of time. For comparison, we also consider a simple linear regression survival model with the ten genes that are marginally most significantly associated with the outcome as covariates. In this case, the sample size $n = 92$ is greater than the number of covariates $d = 10$. So a simple Stute estimate is available. We can see the PCR estimate has dominatingly larger AUC, which suggests better model fitting.

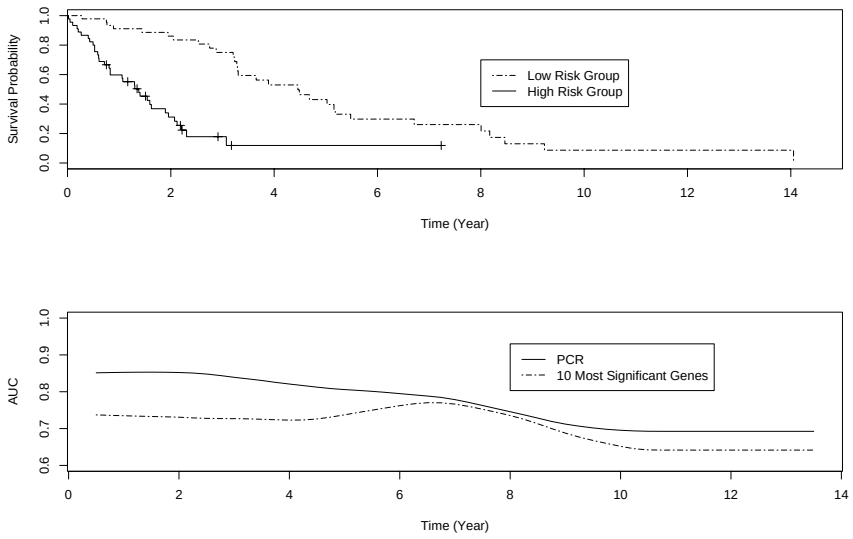

Figure 1: Mantle cell lymphoma data. Upper panel: survival function for the two risk groups defined by the PCR estimate. Lower panel: time-dependent ROC.

3.2 Follicular lymphoma data

Follicular lymphoma is the second most common form of non-Hodgkin's lymphoma, accounting for about 22 percent of all cases. An experiment was conducted to determine whether the length of survival among patients with follicular lymphoma can be predicted by the gene-expression profiles of the tumors at diagnosis. Fresh-frozen tumor-biopsy specimens and clinical data from 191 untreated patients who had received a diagnosis of follicular lymphoma between 1974 and 2001 were obtained. The median age at diagnosis was 51 years (range 23 to 81), and the median follow up time was 6.6 years (range less than 1.0 to 28.2). The median follow up time among patients alive at last follow up was 8.1 years. Eight records with missing survival information are excluded from the downstream analysis. Detailed experimental protocol can be found in Dave *et al.* (2004). The gene expression data and survival data can be downloaded from *http://content.nejm.org/cgi/content/abstract/351/21/2159*.

RNA was examined for gene expression with the use of Affymetrix U133A and U133B microarrays. A log 2 transformation was applied to the Affymetrix measurements. We first filter the 44928 gene measurements with the following criteria: (1). the max expression value of each gene across 191 samples must be

Table 2: Follicular lymphoma data: the 30 genes with the largest absolute values of $\hat{\beta}_{pc}$.

Gene ID	Gene Description	$\hat{\beta}_{pc} \times 10^2$
222450_at	transmembrane, prostate androgen induced RNA	-1.199
225981_at	chromosome 17 open reading frame 28	-1.606
227860_at	carboxypeptidase X (M14 family)	1.206
228209_at	CDNA FLJ38931 fis, clone NT2NE2013189	-1.868
228624_at	hypothetical protein FLJ11155	-1.208
228844_at	solute carrier family 13, member 5	1.242
231578_at	guanylate binding protein 1, interferon-inducible, 67kDa	-1.262
231822_at	hypothetical protein DKFZp547A023	-1.537
232018_at	leukocyte receptor cluster (LRC) member 1	-1.375
232303_at	zinc finger protein 608	-1.274
233834_at	Nuclear receptor coactivator 2	-1.161
234836_at	MRNA; cDNA DKFZp586G0822	1.330
235530_at	Sequestosome 1	-1.216
236916_at	Transcribed locus	1.637
237546_at	Interleukin 19	-1.369
237744_at	–	-1.916
238605_at	Nucleolar protein 4	-1.237
243430_at	seizure related 6 homolog (mouse)	-1.353
244407_at	cytochrome P450, family 39, subfamily A, polypeptide 1	-1.262
244657_at	Glucosidase, beta, acid 3 (cytosolic)	1.177
222545_s_a	chromosome 10 open reading frame 57	-1.775
229100_s_a	translocase of inner mitochondrial membrane 22 homolog	1.343
234419_x_a	–	-1.211
241748_x_a	DiGeorge syndrome critical region gene 14	-1.296
242938_s_a	forkhead box K2	-1.281
206641_at	tumor necrosis factor receptor superfamily, member 17	1.493
213771_at	interferon regulatory factor 2 binding protein 1	-1.457
215536_at	major histocompatibility complex, class II	1.300
209547_s_a	splicing factor 4	-1.254
209863_s_a	tumor protein p73-like	-1.740

greater than 9.186 (the median of the maximums of all probes). (2). the max-min should be greater than 3.874 (the median of the max-min of all probes). After steps (1) and (2), there are 6523 probes left. (3). Compute correlation coefficients of the uncensored survival times with gene expressions. Select the genes whose correlation with survival time is greater than 0.2. 729 genes pass this screening process. We normalize genes across samples to have mean 0 and variance 1.

We apply the proposed approach to the Follicular lymphoma data. Using the proposed nonparametric bootstrap, six principal components are significant at the 0.05 level with the Bonferroni adjustment. The 30 genes with the largest absolute value of the estimated PCR coefficients are shown in Table 2.

Model evaluation plots are shown in Figure 2. The survival functions for the two risk groups defined with the PCR estimate differ significantly with p-value 0.002. The AUC for the PCR is significantly larger than the AUC estimated with the ten marginally most significant genes (measured by the associations with the outcome) under the linear regression survival model.

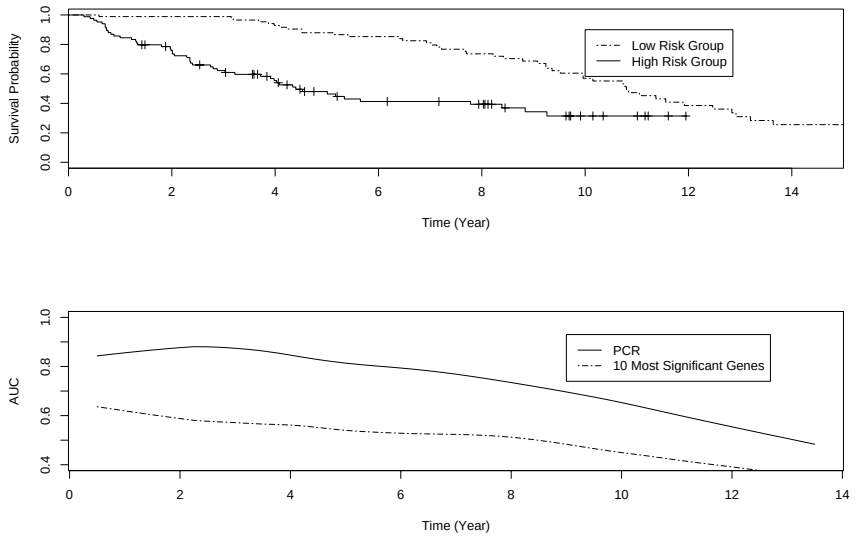

Figure 2: Follicular lymphoma data. Upper panel: survival function for the two risk groups defined by the PCR estimate. Lower panel: time-dependent ROC.

For the above two datasets, the principal components can be interpreted as composite genes (or so called "super genes"). This has been discussed in Lan *et al.* (2003) for uncensored data. It is also worth pointing out that interpretation of individual gene effects (with the PCR approach) may be obscured because one cannot identify individual effects. This limitation is the price to pay for a more parsimonious model.

4. Concluding Remarks

In this article, we investigate principal component regression with the linear regression survival model using high dimensional microarray measurements. Compared with the extensively used Cox model and the additive risk model, the linear regression model directly models the event times, is easy to interpret and hence preferable in some cases. The Stute least squared type estimating equation makes its adaption to microarray studies feasible. The PCR approach is one of the most natural with the least squared objective function. The computational cost involved is minimal compared with other estimation approaches. The most serious concern with analyzing microarray data is the extremely high dimensionality. In our study, this problem is partly solved by filtering genes first, i.e, removing genes with little variations before the analysis. The main solution is the computational simplicity inherent in the PCR method (Lan *et al.* 2003).

The selection of the principal components is based on the marginal significance in this article. Other component selection techniques include selecting components with large eigenvalues, the cross validation techniques in Jolliffe (1986), and the mean squared error based selection in Hwang and Nettleton (2003). The performance of different selection techniques is data dependent and a detailed evaluation is beyond the scope of this article.

The linear regression survival model and the proposed PCR approach provide a useful alternative to existing dimension reduction techniques based on Cox's model for right censored survival data with microarray measurements. It is of interest to compare the relative efficacy of different models (for example the Cox model, the additive model, and the linear regression model) and different model selection techniques (for example, penalization methods like LASSO, PCR, and PLS). Based on previous work on simple linear models, it is expected that the relative performances of the different models/dimension reduction methods are data dependent, with no approach dominating the others. Comprehensive simulation studies and data analysis will be needed to draw more definitive conclusions.

Acknowledgment

The author is partly supported by N01-HC-95159 from the National Heart, Lung, and Blood Institute. The author gratefully acknowledges Dr. Jian Huang and the referee for insightful suggestions.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., *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., Gish, K., Mack, S. and Levine, J. (1999). Broad Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of National Academy of Science USA* **96**, 6745-6750.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289-300.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429-436.

Dave, S. S., Wright, G., Tan, B. *et al.* (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *The New England Journal of Medicine* **351** 2159-2169.

Dudoit, S., Fridyland, J. F. and Speed, T. P. (2002). Comparison of discrimination methods for tumor classification based on microarray data. *J. Amer. Statist. Asso.* **97**, 77-87.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74-99.

Fleming, T. R. and Harrington, D. P (1991). *Counting Processes and Survival Analysis*. Wiley.

Garber, M. E., Troyanskaya, O. C., Schluens, K., Petersen, S. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of National Academy of Science USA* **98**, 13784-13789.

Ghosh, D. and Chinnaiyan, A. M. (2005). Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology* **2:** 147-154.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* **286**: 531-537.

Gui, J. and Li, H. (2005). Threshold gradient descent method for censored data regression with applications in pharmacogenomics. *Proceedings of Pacific Symposium on Biocomputing 2005*.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A. A., Levy, R., Staudt, L., Chan, W. C., Bostein, D. and Brown, P. (2001). Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **2**, 1-21.

Huang, J. and Harrington, D. (2004). Dimension reduction in the linear model for right-censored data: predicting the change of HIV-I RNA levels using clinical and protease gene mutation data. *Lifetime Data Analysis* **10**, 425-443.

Hwang, G. and Nettleton, D. (2003). Principal components regression with data chosen components and related methods. *Technometrics* **45**, 70-79.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. L. (2003). Rank-based inference for the accelerated failure time model. Biometrika **90**, 341-353.

Johnson, R. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis.* Prentice-Hall.

Jolliffe I. T. (1986). *Principal Component Analysis*. Springer-Verlag.

Kollo, T. and Neudecker, H. (1993). Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices. *Journal of Multivariate Analysis* **47**, 283-300.

Lan, H., Stoehr, J. P., Nadler, S. T., Schueler, K. L., Yandell, B. S. and Attie, A. D. (2003). Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* **164**, 1607-1614.

Li, H. Z. and Luan, Y. H. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing* **8**, 65-76.

Nguyen, D. and Rocke, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray data. *Bioinformatics* **18**, 1625-1632.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions (in resampling). *Annals of Statistics* **22**, 2031-2050.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Proceedings of Pacific Symposium on Biocomputing*.

Rosenwald, A. Wright, G., Wiestner, A., Chan, W. C., *et al.* (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185-197.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica* **9**, 1089-1102.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* **11**, 1871-1879.

Wei, L. J., Ying, Z. L. and Lin D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. Biometrika **77**, 845-851.

Ying, Z. L. (1993). A large sample study of rank estimation for censored regression data. *Annals of Statistics* **21**, 76-99.

Received December 6, 2005; accepted February 9, 2006.

Steven Ma
Collaborative Health Studies Coordinating Center
Bldg. 29, Suite 310
University of Washington
6200 NE 74th Street
Seattle, WA 98115, USA
shuangge@u.washington.edu