Published in final edited form as:

Proc IEEE Int Symp Bioinformatics Bioeng. 2010 ; 2010(5521718): 26–31. doi:10.1109/BIBE.2010.14.

A supervised approach for predicting patient survival with gene expression data

Karthik Devarajan,

Division of Population Science, Fox Chase Cancer Center, Philadelphia, PA 19111,
karthik.devarajan@fccc.edu

Yan Zhou,

Division of Population Science, Fox Chase Cancer Center, Philadelphia, PA 19111,
yan.zhou@fccc.edu

Neeraj Chachra, and

A.C. Nielsen, Schaumburg, IL 60173

Nader Ebrahimi

Division of Statistics, Northern Illinois University, DeKalb, IL 60115, nader@math.niu.edu

Abstract

Rapid development in genomics in recent years has allowed the simultaneous measurement of the expression levels of thousands of genes using DNA microarrays. This has offered tremendous potential for growth in our understanding of the pathophysiology of many diseases. When microarray studies also contain information about an outcome variable such as time to an event or death, one of the goals of an investigator is to understand how the expression levels of genes (covariates) relate to the time-to-event (referred to as survival time) in the course of a disease.

In this article, we consider the case where the number of covariates, p , exceeds the number of observations, N , a setting typical of microarray gene expression data. For a given vector of responses representing survival times of N subjects and the corresponding $p \times N$ gene expression matrix, we examine the problem of predicting the survival probability when $N \ll p$. This is an ill-conditioned problem further compounded by the presence of possibly censored survival times. We propose a model that combines the partial least squares approach for dimensionality reduction with the accelerated failure time model, a widely used log-linear model for linking censored survival time to covariates. We develop parametric methods to account for censoring as well as for predicting patient survival probabilities. We illustrate the applicability of our methods using cancer microarray data and explore the biological relevance of our results using pathway analysis. Finally, we evaluate the performance of our methods using extensive simulation studies.

Keywords

gene expression; partial least squares; accelerated failure time; censored survival data; high-dimensional data

I. INTRODUCTION

The analysis of time-to-event data, generally called survival analysis, arises in many fields of study including medicine, biology, engineering, public health, epidemiology, and economics. The time-to-event is commonly referred to as “survival” time or “failure” time. Consider a clinical trial in which subjects are enrolled continuously in time, starting at some time origin. Let Y denote the survival time of a typical subject in the study, the length of time entry into the

study until a prescribed endpoint is attained. This endpoint may be the onset of a disease or event associated with it, or death itself. In addition, we let Z be the duration of observation of the subject, i.e. the time from entry into the study until removal. The random variable Z is referred to as a censoring variable. In general, we assume that both Y and Z are non-negative random variables of which only the first one to occur is observed. Thus, an observation consists of the pair (T, δ) , where $T = \min(Y, Z)$ and $\delta = I(T = Y)$. We also have data on p covariates from each subject. It is assumed that censoring is non-informative, i.e. the survival time Y and the censoring mechanism Z are independent, and that the covariates do not provide information about the censoring time Z . Survival data usually consists of N samples, each containing the triple $(T_i, \delta_i, \mathbf{x}_i)$ for $i = 1, \dots, N$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ is the covariate profile of the i -th subject, T_i is the survival time if $\delta_i = 1$ and it is the right censored time if $\delta_i = 0$. Typically one of the goals of an investigator is to estimate the survival probability for a subject with a certain set of values for the covariates.

Recent developments in DNA microarray technology has allowed the measurement of expression levels of thousands of genes simultaneously. The ability to measure gene expression en masse has resulted in data where the number of covariates (genes), p , far exceeds the number of observations (samples), N which is at most in the hundreds even in larger studies. Even with preprocessing of the data, which is designed to filter out genes that have incomplete information or those which are unlikely to be relevant, the “ $N \ll p$ ” problem still strongly persists. Direct application of statistical techniques results in ill-posed and often computationally infeasible problems due to the high-dimensionality of the covariate space. This problem is further complicated by the presence of censored observations. See Elston and Spence (2006) for more details.

Partial least squares (PLS) is a method for constructing linear predictive models in the high-dimensional setting with highly collinear covariates (Garthwaite (1994)). Since its introduction, PLS has been applied effectively as a tool for dimension reduction in many areas including chemometrics and bioinformatics. The interested reader is referred to Martens and Naes (1989) and Boulesteix and Strimmer (2007) for examples of applications. PLS sequentially finds linear combinations of covariates that are most strongly correlated with the response variable, while being orthogonal to the linear combinations that have been already found in the sequence. Due to the high collinearity in gene expression data from microarray studies, PLS finds a natural application for correlating such data with a response variable such as survival time. While dimensionality reduction can be achieved via PLS, it is also important to account for censored observations in this setting.

Several methods have been proposed in the literature to relate censored survival time to gene expression data. Rosenwald et al. (2002) described an approach for identifying genes related to survival time that fits Cox proportional hazards (PH) models to each gene and selects those that pass a threshold for significance. Hastie et al. (2001) described an approach that first clusters the genes, then fits a Cox PH model using the average expression level of each cluster as a covariate. Nguyen and Rocke (2002) considered the problem of relating survival time to gene expression by first reducing the dimensionality via PLS; the first few linear combinations of gene expressions obtained via PLS are then used in a Cox PH regression model for predicting the survival probabilities. While the Cox PH regression model is inherently designed to handle censored observations, Nguyen and Rocke (2002) treated the censored survival times as actual event times in the PLS model. Later, Nguyen (2005) proposed a PLS-based method for handling censored survival times. Other methods based on the Cox PH model for this problem include the penalized regression approach of Gui & Li (2005) and boosting using smoothing splines (Li & Luan, 2005).

Another widely used model that relates survival time to covariates is the accelerated failure time (AFT) model (Klein & Moeschberger, 2003). It is a censored linear regression model in which the covariates cause an acceleration (linear transformation) of the time scale. Unlike the Cox PH model, this model does not imply a constant relative risk and allows the hazard and survival curves corresponding to different values of a covariate to cross. This model can also incorporate a variety of survival time distributions such as the lognormal and Weibull in the parametric setting. Alternatively, one can also take a distribution-free approach to estimation in this model (Jin et al. (2003); Wei & Yuhyun (2003)). These attractive features make this a very flexible model and applicable to a variety of data structures. Recently, Datta et al. (2007) proposed different methods to account for censored observations in linking gene expression data with survival data using this model.

In this paper, we pursue PLS followed by AFT to correlate survival time to gene expression data in order to estimate the subject survival probabilities. Since both PLS and AFT fall within the linear models framework, it seems natural to consider a model that combines their strengths. Moreover, the ability of the AFT model to incorporate crossing hazard curves offers unparalleled flexibility for modeling large-scale biological data. Using this hybrid approach, we propose a parametric method for estimation that accounts for censored observations in the data. We demonstrate the applicability of our methods using microarray data on lung cancer and we evaluate their performance via extensive simulation experiments. We have implemented our methods using the open source R statistical language (www.r-project.org).

II. METHODS

A. Dimensionality Reduction via Partial Least Squares

Consider a $p \times N$ gene expression matrix where rows represent genes and columns represent subjects or samples. The goal of PLS is to sequentially find weight vectors w_k such that the linear combinations of the gene expressions Xw_k maximize the objective function

$$w_k = \arg \max_{w'w=1} \text{Cov}^2(Xw, \log t)$$

subject to xw_k being orthogonal to the already found linear combinations xw_j , $1 \leq j < k$. Here, t denotes the observed time-to-event. Let $u_{ik} = x_i w_k$, $i = 1, \dots, N$, $k = 1, \dots, K$, denote the linear combinations of gene expressions selected by PLS. These represent the latent variables or PLS components. The number of PLS components $K < p$ is typically chosen based on leave-one-out cross validation (LOOCV) to minimize the predicted residual sum of squares (PRESS) statistic (Nguyen & Rocke, 2005). More details on PLS can be found in Hastie et al. (2001) and Garthwaite (1994).

B. The Accelerated Failure Time Model

The AFT model relates the logarithm of survival time linearly to a set of covariates. It provides us with the flexibility of using a parametric as well as a non-parametric approach for model building. For a given set of observations, the parametric AFT model is given by

$$\log Y_i = \beta' u_i + \sigma^* W_i, i = 1, \dots, N, \quad (2.1)$$

where Y_i is the survival time for the i -th subject $u_i' = (u_{i0}, \dots, u_{iK})$ is a $(K + 1)$ vector for the i -th subject, $\beta = (\beta_0, \beta_1, \dots, \beta_K)$ is the $(K + 1)$ -vector of unknown regression parameters, W_i are independent error terms with a common distribution F_W and σ^* is the scale parameter. In our

situation, u_{ik} , $k = 1, \dots, K$ represent the K PLS components for the i^{th} subject, $i = 1, \dots, N$. The parametric approach involves pre-specification of a distribution for the error terms W_i . Under this approach, given that the error term has some known distribution function F_W , all the unknown parameters are estimated using the method of maximum likelihood. From (2.1), given a gene expression vector $\mathbf{x}$ and PLS component $\mathbf{u}$, the survival function of Y is

$$\begin{aligned} S(y|\mathbf{x}) &= P(Y > y|\mathbf{x}) \approx P(\log Y > \log y|\mathbf{u}) \\ &= P(\beta'\mathbf{u} + \sigma^*W > \log y|\mathbf{u}) \\ &= P(W > (\log y - \beta'\mathbf{u})/\sigma^*) \\ &= 1 - F_W\left(\frac{\log y - \beta'\mathbf{u}}{\sigma^*}\right). \end{aligned} \quad (2.2)$$

Now, $S(y|\mathbf{x})$ is estimated by replacing the unknown parameters in (2.2) by their maximum likelihood estimates. In this paper, we consider the lognormal and Weibull distributions, two widely used distributions for modeling censored survival data.

C. PLS followed by AFT

In this section, we describe two procedures. In procedure 1, we use the observed data $t_1, \dots, t_N$ to obtain PLS components. We refer to this procedure as PLSAFT. In procedure 2, we adjust our censored observations and use uncensored observations as well as adjusted censored observations to obtain PLS components. We refer to this procedure as adjusted PLSAFT (APLSAFT).

1. *PLSAFT*: The steps of the PLSAFT method for regressing possibly censored survival time $(\log t_i, \delta_i)$ on gene expression data $\mathbf{x}_i$, $i = 1, \dots, N$ can be summarized as follows:
 - a. Pursue PLS to obtain weight vectors w_k . As specified in section IIA, we used the PRESS statistic based on LOOCV to choose K , the number of PLS components.
 - b. Now, use (2.1) with time to event observations (t_i, δ_i) and covariate $\mathbf{u}_i = (u_{i1}, \dots, u_{iK})$, $i = 1, \dots, N$ to model the effect of the covariates on Y .
 - c. Fit the AFT model (2.1) via maximization of the likelihood function based on the specific distribution of choice.
2. *APLSAFT*: In this section, we first adjust censored observations using a procedure that is similar in flavor to that of Buckley and James (1979). Below, we summarize the steps of APLSAFT.
 - a. Use uncensored data to obtain PLS components where the number of components is chosen as specified above. Choosing these components as covariates for model (2.1) and obtain the estimates of β and σ^* in model (2.1).
 - b. Let $v_i = \log t_i$. Replace v_i by

$$v_i^* = \delta_i v_i + (1 - \delta_i) \widehat{E}(\log Y_i | \log Y_i > v_i), i = 1, \dots, N.$$

Under the parametric model (2.1),

$$\widehat{E}(\log Y_i | \log Y_i > v_i) = \widehat{\beta}\mathbf{u}_i + \widehat{\sigma}^* E(W | W > \frac{v_i - \widehat{\beta}\mathbf{u}_i}{\widehat{\sigma}^*}),$$

where $\hat{\beta}$ and $\hat{\sigma}^*$ are obtained from step 1. The calculation of

$E(W|W > \frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*})$ is described below for some parametric cases.

For the log-normal model (where W is standard normal) one can easily show that

$$E\left(W|W > \frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right) = \frac{\varphi\left(\frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right)}{1 - \Phi\left(\frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right)}, \quad (2.3)$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ and $\Phi(x) = \int_{-\infty}^x \varphi(y) dy$.

For the Weibull model (where W has an extreme value distribution),

$$E\left(W|W > \frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right) = \frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*} + \frac{\Gamma(0, d)}{\bar{G}\left(\frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right)} \quad (2.4)$$

where $\Gamma(0, d) = \int_d^{\infty} \frac{1}{u} \exp(-u) du$, $d = \exp\left(\frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right)$ and $\bar{G}(x) = \exp\left(-\exp\left(\frac{v_i - \hat{\beta}u_i}{\hat{\sigma}^*}\right)\right)$.

- c. Use $v_1^*, \dots, v_N^*$ from step 2 to construct new PLS components. The number of components K for this adjusted survival data is determined using PRESS as described earlier. Use these components as covariates in model (2.1) and follow step 3 of Section (II-C1).

III. Implementation

We utilized the open source R statistical language for implementing the PLSAFT and APLSAFT algorithms and in our computations (R Development Core Team (2007)). Dimension reduction by PLS was based on the algorithm outlined in de Jong et al. (2001). The parametric AFT model was fitted using the function *survreg* in the *survival* package.

IV. Example

In this section, we describe the results of applying PLSAFT and APLSAFT to predict survival probabilities for a gene expression data set with patient survival information. We consider the lung cancer dataset described in Beer et al. (2002). We utilize the gene expression and survival data for the $N = 86$ patients. An interesting aspect of this dataset is that it includes a large proportion of censored observations (62 out of 86) and is well-suited for illustrating our methodology. We considered the 3761 most varying genes (based on coefficient of variation) in our analysis.

The patient survival probabilities were predicted using the PLS gene component profiles directly in the AFT model. Applying PLSAFT, we selected $K = 2$ PLS gene components (linear combinations of gene expressions) and then fitted the AFT model using these components as the covariates. As described in sections II.A and II.C the number of components was chosen by leave-one-out cross validation to minimize the PRESS statistic. For the choice of $K = 2$, the

total response variation explained was 91.92% for the unadjusted data. Using our parametric approach, we need some way to decide between various distribution options. By doing some diagnostics checks based on quantile-quantile (Q-Q) plots (see Klein and Moeschberger (2003) for more details), we found that the best suited distribution for the error term W was normal, i.e., the distribution of survival time is lognormal. The Q-Q plot based on PLSAFT is shown in Figure 1 and illustrates the goodness-of-fit of the lognormal distribution to our data. Based on this, the estimates of the regression parameters for our PLSAFT model are as given in Table 1. So the final model takes the form

$$\log Y = 1.87 + 26.64u_1 + 2.38u_2 + 1.68W. \quad (4.1)$$

Now, using the above equation

$$\widehat{P}(Y > y | x) \approx \widehat{P}(Y > y | u) = \widehat{P}(W > q) = 1 - \Phi(q), \quad (4.2)$$

where $q = \frac{\log y - [1.87 + 26.64u_1 + 2.38u_2]}{1.68}$ and Φ is the cumulative distribution of the standard normal distribution. Here, q is the residual for the PLSAFT model. Next, we applied our APLSAFT model to this data based on the lognormal distribution. As described in section II-C2, we adjusted the data for censored observations according to (2.3) and used this adjusted data to construct new PLS components based. Figure 2 shows the Q-Q plot for the adjusted data based on APLSAFT. Once again, we see that the lognormal distribution provides a good fit of the data. The number of PLS components was determined to be $K = 6$ using the PRESS statistic for the adjusted data. The proportion of response variation explained increased from 91.92% for the unadjusted data to 98.57% for the adjusted data. The parameter estimates are summarized in Table I. Based on these estimates, the survival probability is computed in a manner similar to that shown in (4.2). Figure 3 presents the survival curve using the residuals for the APLSAFT model. It is based on equivalent versions of (4.1) and (4.2) for the APLSAFT model (obtained using the parameter estimates in Table I) and the standard normal cumulative distribution.

These results clearly indicate an effect due to the adjustment for censored observations and this is reflected in the difference in estimated survival curves for both cases.

V. Biological Pathway Analysis

In order to elucidate relevant functional categories and biological pathways associated with patient survival, we mined the list of genes resulting from our analysis. To this end, we utilized the PLS coefficients as a means to assign a “relevance” score to each gene in the data set and rank the genes based on the absolute value of this score. The order of genes in this list signifies the correlation of each gene’s expression profile with patient survival. We considered gene rankings based on unadjusted as well as adjusted data in our analyses. The number of genes in common to the two methods is listed as a function of the number of top ranked genes considered in Table II.

Lists of the top 500 and 1000 genes resulting from this ranking were further interrogated using Ingenuity Pathways Suite (www.ingenuity.com) for the involvement of relevant functional categories and biological pathways. In this analysis, enrichment for a functional category is determined by comparing the number of genes in this list that belong to that category with the number of known genes in that category via a hypergeometric test. For each functional category, statistical significance of enrichment is given by the $-\log_{10}$ p-value associated with

this test. A minimum enrichment score (ES) of 1.3 (corresponding to $p\text{-value} < 0.05$) is required to achieve significance.

Cancer, cellular growth and proliferation, cell death, cellular movement and development, and tissue development were the most enriched functional categories identified in this analysis. Each functional category consists of several sub-categories and an ES is computed for each of these sub-categories. These results are summarized in Table III. In this table, the functional categories are ranked based on the enrichment scores where we have listed the smallest ES among all sub-categories for a given functional category. It is evident from these results that APLSAFT has a consistently higher ES than PLSAFT in the top enriched categories.

Several cancer genes were identified in these categories. Some of the relevant genes include those from the ERBB, IGFBP, CXCL and DUSP families, among others, and are listed in Table IV. These genes were also part of various important biological pathways and networks identified by our analyses (data not shown). Furthermore, many relevant sub-categories were identified within these functional categories. These include tumorigenesis, neoplasia and lung cancer, among others.

VI. Simulation Studies

In this section, we study the performance of the adjustment method via extensive simulation experiments based on randomly generated censored survival data from the lognormal and Weibull distributions.

First, we created an artificial gene expression data set as follows. For each of 1000 mock genes, we generated expression measurements on $N = 40$ subjects based on a normal distribution with mean zero and variance 0.5. We also generated a PLS weight vector β uniformly distributed in $(-0.01, 0.01)$ for the 1000 mock genes.

A single run of our simulation procedure consisted of the following. For the parametric case, we generated time-to-event data based on model (2.1) where W_i , $i = 1, \dots, N$ is a standard normal (or a standard extreme value) random variable and the censoring proportion and scale parameter σ^* were fixed. This resulted in lognormally (or Weibull) distributed event times Y_i , $i = 1, \dots, N$ with a fixed shape parameter given by σ^* and varying scale parameter given by $\beta'X_i$ where β is as specified above and X_i is the gene expression vector for the i^{th} subject.

Next, we fit a PLS model to the Y_i s with the number of PLS components K determined by the PRESS statistic. Based on this, we applied the PLSAFT (unadjusted) and APLSAFT (adjusted) methods described in section II.C.1 and II.C.2 respectively. In each case, we obtained the maximum likelihood estimates of the model parameters β and σ^* and hence estimates of subject survival probabilities using (4.1) (or equivalently for the Weibull case). The true subject survival probabilities are determined by the distribution of the W_i . For each method, we computed the mean integrated squared error (MISE) as the mean squared error between the true survival probabilities and the estimated subject survival probabilities averaged across the $N = 40$ subjects.

We repeated the above steps for 1000 runs and computed the median MISE across the runs for each method. We considered 20% and 70% censored observations and several values of the scale parameter σ^* . Our results are shown in Tables V and VI for the lognormal and Weibull distributions respectively. Overall, there is a clear improvement (in terms of a decrease in median MISE) due to adjustment as the censoring proportion increases from 20% to 70%. Interestingly, there is also an improvement seen with increasing scale parameter σ^* for both lognormal and Weibull distributions.

VII. Conclusion

In this paper, we have explained the mechanics of combining PLS with the AFT model using parametric methods. We have demonstrated the applicability of our methods using microarray gene expression data with censored survival outcomes. Given gene expression measurements for an individual, it is possible to estimate the corresponding survival probability using PLS for dimension reduction and the AFT model for estimating the survival curve. Our simulations demonstrate an improvement due to adjustment in terms of a reduction in MISE, as the proportion of censored observations increases. There is also an improvement seen with an increase in the scale parameter σ^* . An advantage of our parametric approach is that the resulting survival curve can be smooth. The ability of our methods to handle censored observations broadens their applicability. Since our approach falls within the widely used linear models paradigm, they can also be directly used for correlating any continuous clinical outcome (such as blood pressure or prostate-specific antigen level) with high-dimensional data. The flexible nature of our hybrid approach makes it applicable for correlating censored patient survival data not only with microarray gene expression data but also with large-scale biological data stemming from other high-throughput technologies such as DNA copy number analysis and proteomics.

References

1. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high dimensional genomic data. *Bioinformatics* 2007;8:32–44. [PubMed: 16772269]
2. Buckley J, James I. Linear regression with censored data. *Biometrika* 1976;66:429–436.
3. Datta S, Le-Rademacher J, Datta S. Predicting patient survival data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* 2007;63(1):259–271. [PubMed: 17447952]
4. Elston RC, Spence MA. Advances in statistical human genetics over the last 25 years. *Statistics in Medicine* 2006;25:3049–3080. [PubMed: 16900576]
5. Garthwaite PH. An interpretation of partial least squares. *Journal of the American Statistical Association* 1994;89:122–127.
6. Hastie T, Tibshirani R, Botstein D, Brown P. Supervised harvesting of expression trees. *Genome Biology* 2001;2:1–12.
7. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*. New York: Springer-Verlag; 2001.
8. Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for accelerated failure time model. *Biometrika* 2003;90:341–353.
9. de Jong S, Wise BM, Ricker NL. Canonical partial least squares and continuum power regression. *Journal of Chemometrics* 2001;15:85–100.
10. Klein, PJ.; Moeschberger, LM. *Survival Analysis: Techniques for censored and truncated data*. New York: Springer; 2003.
11. Martens, H.; Naes, T. *Multivariate Calibration*. Wiley; New York: 1989.
12. Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics* 2002;18:1625–1632. [PubMed: 12490447]
13. Nguyen DV. Partial least squares dimension reduction for microarray gene expression data with censored response. *Mathematical Biosciences* 2005;193:119–137. [PubMed: 15681279]
14. R Development Core Team. *R Foundation for Statistical Computing*. Vienna, Austria: 2004. R: A language and environment for statistical computing. URL <http://www.r-project.org>
15. Rosenwald A, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *New England Journal of Medicine* 2002;346:1937–1947. [PubMed: 12075054]
16. Wei LJ, Yuhyun P. Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* 2003;90:717–723.

17. Ingenuity Pathway Analysis Suite. <http://www.ingenuity.com>
18. Beer DG, Kardia SL, Huang C, et al. Gene expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 2002;8:816–825.
19. Li H, Luan Y. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics* 2005;21(10):2403–2409. [PubMed: 15713732]
20. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005;21(13):3001–3008. [PubMed: 15814556]

Fig. 1.
Normal Q-Q Plot: PLSAFT

Fig. 2.
Normal Q-Q Plot: APLSAFT

Fig. 3.
Survival Curve using Residuals for APLSAFT

TABLE I

Regression parameter estimates for Lung Cancer data using PLSAFT and APLSAFT

PLSAFT	$\hat{\beta}_0 = 1.87, \hat{\beta}_1 = 26.64, \hat{\beta}_2 = 2.38, \sigma^* = 1.68$
APLSAFT	$\hat{\beta}_0 = -1.11, \hat{\beta}_1 = 53.11, \hat{\beta}_2 = 12.06, \hat{\beta}_3 = 7.41, \hat{\beta}_4 = 5.70, \hat{\beta}_5 = 6.72, \hat{\beta}_6 = 4.74, \sigma^* = 0.75$

TABLE II

Comparison of Top Ranked Genes Between PLSAFT and APLSAFT

# of Top Ranked Genes	# of Common Genes
10	4
50	21
100	47
200	109
500	331
1000	741

TABLE III

Summary of Enrichment Analysis Based on Top 500 Genes

Functional Category	ES (PLSAFT)	ES (APLSAFT)
Cancer	4.50	4.80
Cell Growth	4.46	4.75
Cell Death	4.49	4.79
Cell Movement	4.46	4.77
Tissue Development	4.46	4.77

TABLE IV

Summary of Pathway Analysis

Functional Category	Gene Name
Cancer	ERBB2,3; IGFBP2,5; CXCL1,2,9,12; DUSP4,6; SERPINA1,3
Cell Growth	CXCL1,2,9,12; ERBB2,3; ID3; IL7R; BCAM
Cell Death	TFF3; ECM1; IGFBP2,3,5

TABLE V

Simulation Results: Lognormal

% Censored	Shape	Med. MISE	
		Unadj.	Adj.
20	2	0.154	0.151
70	2	0.177	0.175
20	5	0.155	0.153
70	5	0.183	0.162
20	10	0.155	0.154
70	10	0.186	0.159
20	20	0.156	0.156
70	20	0.194	0.159
20	50	0.160	0.159
70	50	0.204	0.161
20	100	0.161	0.160
70	100	0.207	0.162

TABLE VI

Simulation Results: Weibull

% Censored	Shape	Med. MISE	
		Unadj.	Adj.
20	2	0.182	0.177
70	2	0.178	0.148
20	5	0.182	0.177
70	5	0.177	0.144
20	10	0.182	0.177
70	10	0.177	0.144
20	20	0.182	0.177
70	20	0.177	0.139
20	30	0.182	0.177
70	30	0.177	0.132
20	50	0.182	0.177
70	50	0.177	0.125