Regularized Estimation in the Accelerated Failure Time Model with High Dimensional Covariates

Jian Huang[1,2], Shuangge Ma[3], and Huiliang Xie[1]

[1]Department of Statistics and Actuarial Science, and [2]Program in Public Health Genetics,

University of Iowa

[3]Department of Biostatistics, University of Washington–Seattle

* *e-mail:* jian@stat.uiowa.edu

June 2005

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 349

Summary The need for analyzing failure time data with high-dimensional covariates arises in investigating the relationship between a censored survival outcome and microarray gene expression profiles. We consider two regularization approaches, the LASSO and the threshold gradient directed regularization, for variable selection and estimation in the accelerated failure time model with high-dimensional covariates based on Stute's weighted least squares method. The Stute estimator uses the Kaplan-Meier weights to account for censoring in the least squares criterion. The weighted least squares objective function makes the adaption of this approach to high dimensional covariate settings computationally feasible. We use the V-fold cross validation and a modified Akaike's Information Criterion for tuning parameter selection, and a bootstrap approach for variance estimation. The proposed method is evaluated using simulations and demonstrated with a real data example.

KEYWORDS: Cross validation; LASSO; Microarray; Threshold gradient directed regularization; Variable selection; Weighted least squares.

1. Introduction

The accelerated failure time (AFT) model is a linear regression model in which the response variable is the logarithm or a known monotone transformation of a failure time (Kalbfleisch and Prentice, 1980). As a useful alternative to the Cox model (Cox, 1972), this model has an intuitive linear regression interpretation, see Wei (1992) for a lucid discussion. Semiparametric estimation in the AFT model with an unspecified error distribution has been studied extensively in the literature for right censored data. In particular, two methods have received special attention. One method is the Buckley-James estimator which adjusts censored observations using the Kaplan-Meier estimator. The other is the rank based estimator which can be motivated from the score function of the partial likelihood, see for example, Prentice (1978); Buckley and James (1979); Ritov (1990); Tsiatis (1990); Wei, Lin and Ying (1990); and Ying (1993), among others. However, the AFT model has not been widely used in practice, mainly due to the difficulties in computing the semiparametric estimators of the afore mentioned methods, even in situations when the number of covariates is relatively small (Jin, Lin, Wei and Ying, 2003). For high-dimensional covariates it is even more difficult to apply these methods, or their regularized versions, especially when variable selection is needed along with estimation.

The need for analyzing failure time data with high-dimensional covariates arises in investigating the relationship between a censored survival outcome and microarray gene expression profiles, see for example, Alizadeh et al. (2000) and Rosenwald et al. (2003). These studies use large scale gene expression profiling in analysis of various types of lymphoma. The sample size in these studies is at most in the hundreds, while the number of genes is at least in the thousands. One important goal of these studies is to identify genes that are associated with and are predictive of survival times. This is a variable selection problem from a statistical standpoint. Such studies call for methods that can simultaneously accomplish variable selection and estimation.

In this paper, we study two regularized versions of Stute's weighted least squares (LS) estimator

(Stute 1993, 1996) in the AFT model with high dimensional covariates, the least absolute shrinkage and selection operator method (LASSO, Tibshirani 1996) and the threshold gradient directed regularization method (TGDR, Friedman and Popescu 2004) for variable selection and model fitting. The Stute estimator uses the Kaplan-Meier weights to account for censoring in the least squares criterion. It is computationally more amenable to high-dimensional covariates than the Buckley-James and rank based estimators. It also has rigorously theoretical justifications under reasonable assumptions. The LASSO and the TGDR methods have also been applied to the Cox model with high dimensional covariates for variable selection and estimation (Tibshirani, 1997 and Gui and Li 2004, 2005). Because of the least squares structure in the criterion function for the Stute estimator, it is computationally efficient to apply the LASSO and the TGDR methods in the AFT model.

In the following, we first define Stute's weighted LS estimator. In Section 3, we describe the LASSO and the TGDR methods for regularization of the weighted LS objective function. We use the V-fold cross validation for tuning parameter selection. We propose using a bootstrap method for variance estimation of the regularized Stute estimators. Section 4 contains the asymptotic properties of the Stute estimator under the L_1 constraint assuming fixed dimensional covariates and large sample size. In Section 5, we use simulations to evaluate the proposed methods and apply them to a study that investigates the relationship between censored survival times of mantel cell lymphoma patients and their gene expression profiles as an illustration. Concluding remarks are given in Section 6.

2. Weighted least squares estimation in AFT model

Let T_i be the logarithm of the failure time and X_i a d-dimensional covariate vector for the ith subject in a random sample of size n. The AFT model assumes

$$T_i = \beta_0 + X_i'\beta + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where β_0 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 logarithm of the censoring time and $\delta_i = 1_{\{T_i \le C_i\}}$ is the censoring indicator. Suppose that a random sample $(Y_i, \delta_i, X_i), i = 1, \dots, n$ with the same distribution as (Y, Δ, X) is available.

Let $\widehat{F}_n$ be the Kaplan-Meier estimator of the distribution function F of T. Following Stute and Wang (1993), $\widehat{F}_n$ 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 jumps in the Kaplan-Meier estimator and can be expressed as,

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

which are also called the Kaplan-Meier weights. Here $Y_{(1)} \le \cdots \le Y_{(n)}$ are the order statistics of Y_i's and $\delta_{(1)}, \dots, \delta_{(n)}$ are the associated censoring indicators. Similarly, let $X_{(1)}, \dots, X_{(n)}$ be the associated covariates of the ordered Y_i's. Let $\theta = (\beta_0, \beta)$. Stute (1993, 1996) proposed the weighted least squares estimator $\widehat{\theta} \equiv (\widehat{\beta}_0, \widehat{\beta})$ that minimizes

$$M(\theta) = \frac{1}{2} \sum_{i=1}^n w_{ni} (Y_{(i)} - \beta_0 - X'_{(i)} \beta)^2. \tag{2}$$

Under reasonable conditions, Stute (1993, 1996) proved that $\widehat{\theta}$ is consistent and asymptotically normal as $n \to \infty$ for a fixed d. We describe these conditions in Section 4 below.

3. Regularized weighted LS regression for AFT model

If d is comparable to or greater than n, regularization is needed to obtain a stable estimator of θ with smaller prediction error. In addition, in the "small n and large d" settings, it is desirable to carry out variable selection and estimation simultaneously. We consider two methods for this purpose, the LASSO and the TGDR methods in (2).

We first center $X_{(i)}$ and $Y_{(i)}$ by their w_{ni}-weighted means, respectively. Let

$$\overline{X}_w = \frac{\sum_{i=1}^n w_{ni}\, X_{(i)}}{\sum_{i=1}^n w_{ni}}, \quad \overline{Y}_w = \frac{\sum_{i=1}^n w_{ni}\, Y_{(i)}}{\sum_{i=1}^n w_{ni}}.$$

We replace $X_{(i)}$ and $Y_{(i)}$ with $w_{ni}^{1/2}(X_{(i)} - \overline{X}_w)$ and $w_{ni}^{1/2}(Y_{(i)} - \overline{Y}_w)$, respectively. For simplicity, we still use $X_{(i)}$ and $Y_{(i)}$ to denote the weighted centered values. Using the weighted centered values, the intercept estimate is zero. So the weighted LS objective function can be written as

$$M(\beta) = \frac{1}{2}\sum_{i=1}^{n}(Y_{(i)} - X'_{(i)}\beta)^2. \tag{3}$$

3.1 The LASSO estimator

The LASSO estimator for linear regression is defined as the maximizer of an LS objective function under the L_1 constraint $\sum_{j=1}^d |\beta_j| \le u$, for a data-dependent tuning parameter u (Tibshirani, 1996). Here the constraint is only imposed on the regression coefficient β, not including the intercept. The tuning parameter u determines how many estimated coefficients are zero. The L_1 constraint is equivalent to adding an L_1 penalty to the objective function. Kim and Kim (2004) noticed that the L_1 boosting algorithm provided a computationally feasible and flexible solution to the LASSO type estimators, especially in high dimensional covariates cases. We apply it to the present LASSO estimation problem. For a fixed u, it can be implemented in the following steps:

1. Initialization $\beta = (0, \ldots, 0)'$.

2. With the current estimate of β, compute g(β), the negative derivative of $M(\beta)$ with respect to β. Denote the j^{th} component of g(β) as $g_j(\beta)$, $j = 1, \ldots, d$.

3. Find j^* that minimizes $\min(g_j(\beta), -g_j(\beta))$. If $g_{j^*}(\beta) = 0$, then stop the iteration.

4. Otherwise denote $\gamma = -\text{sign}(g_{j^*}(\beta))$. Find $\hat{\kappa} \in [0, 1]$ that minimizes $M((1-\kappa)\beta + \kappa \times$

$u \times \gamma\eta_{j^*})$, where η_{j^*} is a length d vector that has the p^*th element equals to 1 and the rest components equal to 0.

5. For the j^{th} component of β: $\beta_{(j)} = (1-\hat{\kappa})\beta_{(j)}$ for $j \neq j^*$ and $\beta_{(j^*)} = (1-\hat{\kappa})\beta_{(j^*)} + \gamma u \hat{\kappa}$. Let $m = m+1$.

6. Repeat steps 2–5 until convergence or a fixed number of iterations N has been reached.

The β at convergence is the LASSO estimate (Kim and Kim, 2004). We conclude convergence if the absolute value of $g_{(j^*)}(\beta)$ computed in step 3 is less than a pre-defined criteria, and/or if $M(\beta)$ is smaller than a pre-defined threshold.

An attractive feature of the L_1 boosting algorithm is that its convergence rate is independent of the dimension of input, which is particularly valuable for high dimensional genomic data (Kim and Kim, 2004). In addition, it has been known that for boosting methods, over-fitting usually does not pose a serious problem (Friedman, Hastie and Tibshirani, 2000). So the overall iteration number can be taken to be a large number to ensure convergence. We note that we can also compute the LASSO estimator using the LARS algorithm (Efron, Hastie, Johnstone and Tibshirani, 2004), where the number of iterations is larger than d.

3.2 The TGDR estimator

The TGDR algorithm proposed by Friedman and Popescu (2004) can be adapted to estimation of (3) as follows. Let Δ_ν be a fixed small positive number, and let ν be the index for the point along the parameter path. Let $\beta(\nu)$ denote the parameter estimate corresponding to the index ν. For any fixed threshold value $0 \leq \tau \leq 1$, the TGDR algorithm consists of the following iterative steps:

1. Initialize $\beta(0) = (0, \ldots, 0)'$ and $\nu_0 = 0$.

2. For the current estimate $\beta = \beta(\nu)$, compute the negative gradient $\mathrm{g}(\nu) = -\partial M(\beta)/\partial\beta$. Denote the j^{th} component of $\mathrm{g}(\nu)$ as $g_j(\nu)$. If $\max_j |g_j(\nu))| = 0$, stop the iteration.

3. Compute the vector $\mathbf{f}(\nu)$ of length d, where the j^{th} component of $\mathbf{f}(\nu)$: $f_j(\nu) = I\{|g_j(\nu)| \geq \tau \cdot \max_j |g_j(\nu)|\}$.

4. Update $\beta(\nu + \Delta_\nu) = \beta(\nu) + \Delta_\nu \mathbf{g}(\nu)\mathbf{f}(\nu)$ and $\nu = \nu + \Delta_\nu$.

5. Steps 2–4 are repeated S times. S is taken to be a large number to yield a full parameter path.

The product of $\mathbf{f}$ and $\mathbf{g}$ in step 4 is component-wise: $\mathbf{f}(\nu)\mathbf{g}(\nu) = (f_1(\nu)g_1(\nu), \ldots, f_d(\nu)g_d(\nu))'$. A possible variation of the above algorithm is to use the standardized negative gradient $\mathbf{g}(\nu) = \mathbf{g}(\nu)/\max_j |g_j(\nu)|$ in step 4, so that each increment cannot be overly greedy and subtle structures are not missed. The threshold τ determines the relative degree of regularization: large τ yields estimates close to the LASSO/LARS, whereas estimates with small τ are close to those from the ridge regression. Since each increment is made in a direction in an acute angle with the negative gradient, each iteration decreases $M(\theta)$.

3.3 Tuning parameter selection

We use the V-fold cross validation (Wahba, 1990) to determine the tuning parameters: u for the LASSO and (k, τ) for the TGDR. For a pre-defined integer V, partition the data randomly into V non-overlapping subsets of equal sizes. We define the CV score and the Akaike's Information Criterion (AIC) type of score as

$$CV\ score = \sum_{v=1}^{V} \left[M(\hat{\theta}^{(-v)}) - M^{(-v)}(\hat{\theta}^{(-v)}) \right],\ AIC\ score = n \times \log(CV\ score) + 2K, \qquad (4)$$

respectively, for a fixed u (LASSO) or (k, τ) (TGDR). Here $\hat{\theta}^{(-v)}$ is the LASSO (TGDR) estimate of θ based on the data without the v^{th} subset, $M^{(-v)}$ is the function M defined in (3) evaluated without the v^{th} subset, and K is the corresponding number of non-zero coefficients.

For the LASSO, we choose u as the minimizer of the AIC score. For the TGDR, we use the following two-step procedure. For a fixed τ, k is chosen as the minimizer of the $CV\ score$. We then choose τ by minimizing the $AIC\ score$ with cross validated k from the first step.

Comparing with the original V-fold cross validation in Wahba (1990), the proposed approach puts a penalty on the number of non-zero coefficients, which favors smaller models with comparable prediction performance. This is desirable especially for survival data with high dimensional covariates. For the TGDR, k is chosen by minimizing the $CV\ score$ only. The rationale is as follows. For the TGDR with $\tau > 0$, k also determines the number of nonzero coefficients and the degree of shrinkage. So if we also choose k using the AIC type criteria, there will be ”double adjustment”, which is too severe and not necessary.

3.4 Variance estimation

For the LASSO and the TGDR, variances of the estimators can be estimated using the least squares type expressions as in Tibshirani (1997). However, this approach yields zero variance estimates for covariates with zero coefficients, which is not satisfactory. We estimate the variance using the nonparametric 0.632 bootstrap (Efron and Tibshirani, 1993). Sample $m \approx 0.632n$ from the n observations without replacement. Then the bootstrap sample is estimated with the same u (LASSO) or (τ, k) (TGDR). The bootstrap procedure is then repeated I times. After proper scale adjustment, the sample variance of the bootstrap estimates provides an estimate of the variance of β_k. We use $m = 0.632n$ since the expected number of distinct bootstrap observations is about $0.632n$. Computationally, it is more efficient to use a smaller bootstrap sample size.

3.5 Characteristics of LASSO and TGDR

The LASSO and the TGDR are both gradient directed iterative algorithms. One difference is that the LASSO estimate increases in the direction of one covariate, while the TGDR estimate may increase in the direction of multiple covariates at each iteration. For uncensored data, the LASSO has been studied in detail by Tibshirani (1996) and Efron et al. (2004). Friedman and Popescu (2004) showed that the TGDR can provide a path connecting the solutions roughly corresponding to the ridge regression and the solutions roughly corresponding to LASSO by varying the threshold values. Moderate to large threshold values create paths that involve more diverse absolute coefficient values than the ridge regression but less than the LASSO. When two covariates are strongly correlated, their corresponding gradients are close. So the TGDR yields similar estimates for strongly correlated covariates. This property is not shared by the LASSO. It is easy to construct an example where the difference between the two estimated coefficients is bounded away from 0, although their covariates are highly correlated. One drawback of the TGDR is that the TGDR may overestimate the number of non-zero coefficients. More studies are needed to better understand the relative merits of the LASSO and the TGDR for censored data.

4. Asymptotic distribution of the LASSO estimator for large n and fixed d

Stute (1993, 1996) proved consistency and asymptotic normality of the weighted least squares estimator with Kaplan-Meier weights under the assumptions $E(\varepsilon|X) = 0$ and other appropriate conditions. Because the LASSO penalty is not differentiable, Stute's proof is not applicable to the weighted LS estimator under the L_1 penalty. On the other hand, when there is no censoring, Fu and Knight (2000) derived the asymptotic distributions of the lasso-type estimators. Combining the methods of Stute and Fu and Knight, we can derive the asymptotic distribution of the Stute estimator under the L_1 penalty.

Let H denote the distribution function of Y. Under the assumption of independence between T and C, $1 - H(y) = (1 - F(y))\,(1 - G(y))$, where F and G are the distribution functions of T and C, respectively. Let τ_Y, τ_T and τ_C be the end points of the support of Y, T and C, respectively. Let $Z = (1, X')' = (Z_0, Z_1, \ldots, Z_d)'$ and F^0 be the joint distribution of (Z, T). Denote

$$\widetilde{F}^0(z,t) = \begin{cases} F^0(z,t), & t < \tau_Y \\ F^0(z,\tau_Y-) + F^0(z,\{\tau_Y\})1\{\tau_Y \in A\}, & t \geqslant \tau_Y. \end{cases},$$

with A denoting the set of atoms of H. Define two sub-distribution functions:

$$\widetilde{H}^{11}(z,y) = P(Z \leq z, Y \leq y, \delta = 1), \ \ \widetilde{H}^{0}(y) = P(Y \leq y, \delta = 0).$$

For $j = 0, \ldots, d$, denote

$$\begin{aligned}
\gamma_0(y) &= \exp\left\{ \int_0^{y-} \frac{\widetilde{H}^0(dw)}{1 - H(w)} \right\}, \\
\gamma_{1,j}(y;\theta) &= \frac{1}{1 - H(y)} \int 1_{\{w>y\}}(w - z'\theta) z_j \gamma_0(w) \widetilde{H}^{11}(dz, dw), \\
\gamma_{2,j}(y;\theta) &= \iint \frac{1_{\{v<y, v<w\}}(w - z'\theta) z_j \gamma_0(w)}{[1 - H(v)]^2} \widetilde{H}^0(dv) \widetilde{H}^{11}(dz, dw), \\
\gamma_l(y;\theta) &= \big(\gamma_{l,0}(y;\theta), \gamma_{l,1}(y;\theta), \ldots, \gamma_{l,d}(y;\theta)\big)', \quad l = 1, 2.
\end{aligned}$$

We assume that:

(A1) $E(\varepsilon|X) = 0$ and $E(T^2)$ is finite;

(A2) T and C are independent and $P(T \leq C|T, X) = P(T \leq C|T)$;

(A3) $E(ZZ')$ is finite and nonsingular;

(A4) $\tau_T < \tau_C$ or $\tau_T = \tau_C = \infty$;

(A5) (a) For the true parameter value $\theta^{*\prime} = (\beta_0^*, \beta^{*\prime}) = (\beta_0^*, \beta_1^*, \ldots, \beta_d^*)$, $E\left[(Y - Z'\theta^*)^2 ZZ'\delta\right] < \infty$; (b) $\int |(w - z'\theta^*) z_j| C^{1/2}(w) \widetilde{F}^0(dz, dw) < \infty$, for $j = 0, \ldots, d$ and $C(y) = \int_0^{y-} [(1 - H(w))(1 - G(w))]^{-1} G(dw)$.

In (A1), we only need that $E(\varepsilon|X) = 0$. The distribution of ε can depend on covariates. This allows heteroscedastic error terms. For example, the results below hold for $\varepsilon_i = \sigma(X_i)\varepsilon_{0i}$, where ε_{0i}'s are independent and identically distributed with mean 0. This is weaker than that in the Buckley-James method (Buckley and James, 1979) and the rank based method (Jin et al. 2003), where the error terms ε_i's are assumed to have a common distribution and to be independent of X_i's. (A2) assumes that δ is conditionally independent of the covariate X given the failure time

Y. It also assumes that Y and C are independent, which is the same as that for the Kaplan-Meier estimator. However, we note that (A2) does allow the censoring variable to be dependent on the covariates. In comparison, in the Buckley-James and rank based estimators, it is assumed that $T - \beta_0 - X'\beta$ and $C - \beta_0 - X'\beta$ are conditionally independent given X. (A3) is a standard assumption in linear regression models. (A4), together with (A1), renders the true value to be the minimizer of (2). (A5a) ensures that the weighted LS-LASSO estimator has finite variance. (A5b) guarantees that the bias of Kaplan-Meier integral is in the order of $o(n^{-1/2})$. It is related to the degree of censoring and the tail behavior of the Kaplan-Meier estimator. Note that (A5) is implied by the often used and simpler assumption that $\tau_T < \tau_C$. The practical interpretation is that the study can only be conducted for a finite length of time, see e.g. Andersen and Gill (1982). Therefore, the assumptions needed for theoretical justification of the Stute estimator are quite mild and comparable to those of the Buckley-James and rank based estimators.

The LASSO estimator $\widehat{\theta}$ can be defined as the minimizer of

$$M_n(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_{ni}(Y_{(i)} - \beta_0 - X'_{(i)}\beta)^2 + \frac{\lambda_n}{n}\sum_{j=1}^{d}|\beta_j|,$$

where λ_n is the penalty parameter.

Theorem 1. (Consistency) Suppose assumptions (A1) – (A4) hold and $\lambda_n/n \to 0$. Then $\widehat{\theta} \to_{a.s} \theta^*$ as $n \to \infty$.

Theorem 2. (Asymptotic Normality) Suppose that assumptions (A1) – (A5) hold and $n^{-1/2}\lambda_n \to \lambda_0 \geq 0$. Let $\Sigma_0 = E(ZZ')$. Then $\sqrt{n}(\widehat{\theta} - \theta^*) \to_D \arg\min(Q)$ as $n \to \infty$. Here

$$Q(\mathbf{b}) = -\mathbf{b}'\mathbf{W} + \mathbf{b}'\Sigma_0\mathbf{b} + \lambda_0\sum_{j=1}^{d}[b_j\text{sgn}(\beta_j^*)1\{\beta_j^* \neq 0\} + |b_j|1\{\beta_j^* = 0\}],$$

where $\mathbf{W} \sim N(0,\Sigma)$ with $\Sigma = \text{Var}\left\{\delta\gamma_0(Y)(Y - Z'\theta^*)Z + (1-\delta)\gamma_1(Y;\theta^*) - \gamma_2(Y;\theta^*)\right\}$.

If $\lambda_0 = 0$ (the penalty is asymptotically negligible), the asymptotic distribution simplifies to that of the Stute estimator. When $\lambda_0 > 0$, if some of the β_j's are zero, the limiting distributions put positive probabilities at 0. Thus LASSO achieves variable selection in the AFT model by using the L_1 penalty as in the uncensored case. Although the results here are for the case of "large n and small d," they provide insight into the LASSO estimator in the AFT model with high dimensional covariates. They also give partial justification for the use of LASSO method in the present problem under the conditions (A1)-(A5).

5. Simulation studies and data example

5.1 Simulation study I: finite sample comparison

We conduct a simulation study on the following six examples with $n = 200$ and $d = 30$ to evaluate the finite sample performance of the proposed methods. For all six models, the event times are generated from $T = \beta_0 + X'\beta + \epsilon$, where $\beta_0 = 0.5$ and $\epsilon \sim N(0, 0.5)$. The censoring variables are generated as uniformly distributed and independent of the events. The tuning parameters are chosen using five-fold cross validation. The censoring rates are about 30% for examples 1–3 and 70% for examples 4–6.

Example 1: the first ten components of β are equal to 1 and the rest components are 0. The pairwise correlation between the i^{th} and the j^{th} components of X is set to be $0.5^{|i-j|}$. In this example, we have a moderate number of large effects.

Example 2: the first 15 components of β are equal to 0.4 and the rest components are 0.2. X is the same as for example 1. In this model, we have a large number of moderate to small effects.

Example 3: The first 15 components of β are 1 and the rest are 0. The predictors are generated as follows:

$$X_i = Z_1 + \epsilon_i,\ \ Z_1 \sim N(0,1),\ i = 1,\ldots,5;\ \ X_i = Z_2 + \epsilon_i,\ \ Z_2 \sim N(0,1),\ i = 6,\ldots,10;$$
$$X_i = Z_3 + \epsilon_i,\ \ Z_3 \sim N(0,1),\ i = 11,\ldots,15;\ \ X_i \sim N(0,1),\ \ X_i\ \ i.i.d.\ \ i = 16,\ldots,30,$$

where ϵ_i are i.i.d $N(0, 0.01), i = 1, \ldots, 15$. In this model, we have 3 equally important groups and within each group there are 5 members. There are also 15 pure noises.

Examples 4 to 6 are the same as examples 1 to 3, respectively, except that the censoring rate is different. The simulation settings are similar to those in Zou and Hastie (2004). Since $n > d$, we are able to compare the LS estimates with the LASSO and the TGDR. The quantities of interest are the mean squared errors of the estimates and the number of nonzero coefficients. Summaries based on 100 replicates are shown in Table 1. The intercept is excluded from the constraint (LASSO) and the threshold step (TGDR) and hence is not counted in the number of nonzero coefficients.

From Table 1, we see that the LASSO can generate small models with reasonably small mean squared errors. When there exist grouping effects (examples 3 and 6), the LASSO estimates are more stable than the LS estimates. The LASSO may underestimate the number of non-zero coefficients when there exist a large number of small covariates effects (examples 2 and 5). The TGDR has smaller mean squared errors. Compared with the LASSO, the TGDR can better identify grouping effects as expected (examples 3 and 6). It is also worth noticing that the TGDR tends to overestimate the number of nonzero coefficients.

5.2 Simulation study II: variance estimation

We use simulations to evaluate the bootstrap approach for variance estimation. 100 datasets are generated from example 1. We consider the LASSO estimates with $u = 5.0$ and $u = 10.0$, and the TGDR estimates with $(k, \tau) = (200, 0.3)$ and $(k, \tau) = (200, 0.8)$. For each dataset/estimation approach, 50 bootstrap replicates are used for variance estimation. In Table 2, we show the standard deviations of the estimates and the means of the bootstrap estimated standard deviations for components 1–3 (with coefficients equal to 1) and 11–13 (with coefficients equal to 0) of β.

We can see from Table 2 that the bootstrap standard deviation estimates match the standard deviations of the estimates very well. Simulation studies with other data settings and tuning parameters yield similar results.

5.3 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 immunophenotypic criteria. Survival times of 64 patients were available and the remaining 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., 1999) were used to quantitate mRNA expression in the lymphoma samples from the 92 patients. The gene expression data set that contains expression values of 8810 cDNA elements is available at *http://llmpp.nih.gov/MCL.*

We apply the AFT model (1) with the LASSO and the TGDR methods to this dataset. These methods have no computational or methodological limitation on the number of genes that can be used in the prediction of patients' failure times. However, because many genes do not change across the patients or have very low corrleation with survival, 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.

Such a first stage filtering step is helpful to obtain more stable estimators since the sample size is relatively much smaller than the total number of genes. 364 genes pass the above selection criterion. We make the logarithm transformation to the observed times (measured in month) and standardize the 364 selected genes to have mean 0 and variance 1. Since the number of the covariates (364) is larger than the sample size (92), reguarization is needed in any estimation procedure. Table 3 shows the genes with nonzero coefficients from the Stute estimator with LASSO and TGDR, their unique identification numbers (UNIQID) in the Lymphochip, and GenBank accession numbers. Some information about these genes are available at the GenBank database (www.ncbi.nlm.nih.gov). Many of these genes are involved in cell proliferation and tumor growth. For example, the gene with UNIQID 15981 (GenBank accession # X65550, gene name MKI67) encodes the proliferation-related antigen Ki-67. This gene is associated with cell proliferation and is widely used in routine pathology as a ”proliferation marker” to measure the growth fraction of cells in human tumors. The gene with UNIQID 24612 (GenBank accession # AF343659, gene name IRTA1) is an immunoreceptor and is implicated in B cell development and lymphomagenesis. The gene with UNIQID 28027 (GenBank accession # NM_001880, gene name ATF2) is the activating transcription factor 2. Strong nuclear ATF2 expression was also associated melanoma and with poor survival in melanoma patients. The gene with UNIQID 24376 (GenBank accession # NM_175739, gene name GCET1) is a serine proteinase inhibitor and is highly restricted to and expressed in normal germinal center B cells. The gene with UNIQID 24488 (GenBank accession # NM_001402, gene name EEF1A1) is a translation elongation factor and interacts with the translationally controlled tumor protein. We note that although this set of genes exhibits strong statistical correlation with patients’ survival times, the analysis here does not provide information on whether they are just genomic markers correlated with the survival times or are actually in the pathways leading to MCL. Indeed, detailed discussion of the biological functions of these genes is beyond the scope of this paper.

For model evaluation, we define two hypothetical risk groups based on dichotomizing the estimated linear risk scores at the median risk score, which gives equal number of subjects in the two risk groups. Survival curves for the risk groups defined by the LASS0 and the TGDR estimates are shown in the top two panels in Figure 1, which suggests significant differences between the two risk groups. The p-values based on the log-rank test are both less then 0.001. Another advantage of the AFT model is that the event times can be actually estimated, which is not shared by the Cox model or the additive risk model. In the bottom two panels of Figure 1, we show the plots of the estimated event times versus the observed times. We notice that at the lower range of the survival times, the fitted times tend to overestimate the observed survival times. However, overall, for uncensored subjects, the fitted times provide reasonable estimates of the observed times.

The prediction properties of the proposed approaches are also investigated. We randomly divide the data into a training set of size 57 and a testing set of size 35. The censoring rates in the two sets are about the same. Estimation is carried out with the training set only. The linear risk scores for the testing set are computed using the estimates from the training set. We generate a risk group indicator based on the linear risk scores, so that there are 17 subjects in the low risk group and 18 in the high risk group. The p-values based on the log-rank test for the differences between two risk groups are 0.002 (LASSO) and 0.014 (TGDR), respectively. For reference, we carry out similar testing to the training set. The p-values for the training set are $<$0.001 (LASSO) and 0.036 (TGDR), respectively. Therefore, both the LASSO and TGDR methods can provide good prediction of low or high risk probabilities based on the expression values of the genes selected in the models.

6. Discussion

Analysis of censored failure time data with high dimensional covariates poses an important practical problem, especially now microarrays that can assay thousands of genes are becoming a routine tool in the studies of many diseases. How to estimate the relationship between gene expression data with patients' survival and identify important genes presents a class of interesting and challenging questions. In this article, we studied two regularized versions of the Stute's estimator for the AFT model with high-dimensional covariates, the LASSO and the TGDR, for simultaneous variable selection and estimation.

The simulation studies and the real data example illustrate that the proposed method can effectively reduce the dimension of the covariates, while providing satisfactory estimation and prediction results. The LASSO method achieves regularization by penalizing the L_1 norm of the regression coefficients. Although no explicit penalty is placed on the regression parameters, the TGDR method regularizes the estimator via cross-validation selection of the number of gradient search steps and the threshold value τ. For a given number of gradient search steps, smaller values of τ yield dense estimates similar to those of ridge regression, while bigger values of τ produce more sparse estimates. A useful feature of the TGDR is that it is capable of selecting a set of covariates that have similar values or are highly correlated. Theoretical properties of the TGDR will be pursued in the future.

In many studies, the covariates include both clinical as well as gene expression data, and investigators may know some covariates are important based on previous studies. So we may not want to subject such covariates to variable selection. Let Z denote the vector of covariates of known importance and α its associated coefficients, then the objective function is

$$M_w(\alpha, \beta) = \frac{1}{2} \sum_{i=1}^{n} w_{ni} (Y_{(i)} - Z'_{(i)}\alpha - X'_{(i)}\beta)^2. \tag{5}$$

We can put the L_1 penalty only on β to obtain a partial LASSO solution. We can also only threshold the derivatives with respect to β. Furthermore, there may be models in which different penalties are appropriate for different parameters. The algorithms described in Section 3 can be easily adapted to such situations.

We have derived the asymptotic distribution of the Stute estimator under the L_1 penalty when the number of covariates is fixed. For the TDGR method, if the gradient search is allowed to continue indefinitely, the TDGR estimator converges to the Stute estimator. Therefore, for the asymptotic distribution of the TGDR estimator is approximately the same for a fixed number of covariates and large search steps. However, several questions remain unanswered for the LASSO and TGDR methods in the AFT model based on the Stute estimator. In particular, it is a difficult and interesting problem to rigorously work out the approximate sampling distributions of the LASSO and TGDR Stute estimators for cross validated tuning parameters. Our simulation study suggests that the 0.632 bootstrap method for variance estimation works reasonably well. It is desirable to theoretically justify the bootstrap method for the estimators proposed here.

ACKNOWLEDGMENT

The work of Huang is supported in part by the NIH grant HL72288-01. The work of Ma is partly supported by N01-HC-95159 from the National Heart, Lung, and Blood Institute.

REFERENCES

Alizadeh, A.A., Eisen, M. B., Davis, R.E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000). Distinct types of diffuse large B-Cell lymphoma identified by gene expression profiling. Nature **403**, 503–511.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. Annals of Statistics, **10**, 1100-1120.

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

Cox, D. R. (1972). Regression models and life-tables (with discussion). Journal of Royal Statistical Society, Series B, **34**, 187-220.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. Annals of Statistics, **32**, 407–499.

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

Friedman, J. H. and Popescu, B. E. (2004). Gradient directed regularization for linear regression and classification. Technical report, Department of Statistics, Stanford University.

Fu, W. and Knight K. (2000). Asymptotics for lasso-type estimators. Annals of Statistics, **28**, 1356-1378.

Gui, J. and Li, H. (2004) Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data. Accepted by Bioinformatics.

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*.

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.

Kim, Y. and Kim, J. (2004). Gradient LASSO for feature selection. *Proceedings of the 21st International Conference on Machine Learning*.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

Prentice, R. L. (1978). Linear rank tests with right censored data. Biometrika, **65**, 167-179.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. Annals of Statistics, **18**, 303-328.

Rosenwald, A. Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., Gascoyne, R. D., Grogan, T. M., Muller-Hermelink, H. K., Smeland, E. B., Chiorazzi, M., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Henrickson, S., Yang, L., Powell, J., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Montserrat, E., Bosch, F., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Fisher, R. I., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Holte, H., Jan Delabie, J. and Staudt L. M. (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. (1993). Consistent estimation under random censorship when covariables are available. Journal of Multivariate Analysis, **45**, 89-103.

Stute, W. and Wang, J. L. (1993). The strong law under random censorship. Annals of Statistics, **14**, 1351-1365.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. Scandinavia Journal of Statistics, **23**, 461-471.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, B, **58**, 267–288.

Tibshirani, R. (1997) The LASSO method for variable selection in the Cox model. *Statistics in Medicine* **16** 385–295.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. Annals of Statistics, **18**, 303-328.

Van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Wahba, G. (1990). *Spline models for observational data.* SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics.

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.

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.

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

APPENDIX: PROOFS OF THEOREMS 1 AND 2

Proof of Theorem 1. Let $Z_i = (1, X_i')'$, $\mathbf{b} = (b_0, b_1, \ldots, b_d)'$. Recall

$$M_n(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_{ni}(Y_{(i)} - \beta_0 - X_{(i)}'\beta)^2 + \frac{\lambda_n}{n}\sum_{j=1}^{d} |\beta_j|.$$

When $\lambda_n/n \to 0$, by (A1) to (A4) and the result of Stute (1993), for every fixed θ, $M_n(\theta) \to_{a.s.} M(\theta)$ where

$$\begin{aligned} M(\theta) &= \frac{1}{2}\left\{\int_{\{T<\tau_Y\}}(T-\beta_0-X'\beta)^2 dP + 1\{\tau_Y \in A\}\int_{\{T=\tau_Y\}}(T-\beta_0-X'\beta)^2 dP\right\} \\ &= \frac{1}{2}E(T-\beta_0-X'\beta)^2 \\ &= \frac{1}{2}E[T-\beta_0^*-X'\beta^*+\beta_0^*-\beta_0+X'(\beta^*-\beta)]^2 \\ &= \frac{1}{2}\left\{\text{Var}(\varepsilon)+E[\beta_0^*-\beta_0+X'(\beta^*-\beta)]^2\right\}, \end{aligned}$$

where the last equality follows from (A1). By (A3), $M(\theta)$ is uniquely minimized at θ^*. Thus the consistency follows by the fact that M_n is a convex function of θ.

Proof of Theorem 2. By the definition of M_n, we have

$$M_n(\theta+n^{-1/2}\mathbf{b}) = \frac{1}{2}\sum_{i=1}^n w_{ni}[Y_{(i)}-(\beta_0^*+n^{-1/2}b_0)-X'_{(i)}(\beta^*+n^{-1/2}\mathbf{b}_{-0})]^2+\frac{\lambda_n}{n}\sum_{j=1}^d |\beta_j^*+n^{-1/2}b_j|,$$

where $\mathbf{b}_{-0} = (b_1, \dots, b_d)'$. Let $Q_n(\mathbf{b}) = n[M_n(\theta^*+n^{-1/2}\mathbf{b}) - M_n(\theta^*)]$. Then

$$Q_n(\mathbf{b}) = -\sqrt{n}\sum_{i=1}^n w_{ni}(Y_{(i)}-Z'_{(i)}\theta^*)Z'_{(i)}\mathbf{b}+\frac{1}{2}\sum_{i=1}^n w_{ni}\mathbf{b}'Z_{(i)}Z'_{(i)}\mathbf{b}+\lambda_n\sum_{j=1}^d \left[|\beta_j^*+n^{-1/2}b_j|-|\beta_j^*|\right].$$

By the results of Stute (1993, 1996), we have

$$\sum_{i=1}^n w_{ni}Z_{(i)}Z'_{(i)} \to_P E(ZZ'),$$

and when (A5a) and (A5b) hold,

$$\sqrt{n}\sum_{i=1}^n w_{ni}(Y_{(i)}-Z'_{(i)}\theta^*)Z_{(i)} \to_D W,$$

where $W \sim N(0, \Sigma)$, with Σ defined in the theorem. The last term in $Q_n(\mathbf{b})$

$$\lambda_n \sum_{j=1}^{d} \left[|\beta_j^* + n^{-1/2} b_j| - |\beta_j^*| \right] \to \lambda_0 \sum_{j=1}^{d} \left[b_j \text{sgn}(\beta_j^*) 1\{\beta_j^* \neq 0\} + |b_j| 1\{\beta_j^* = 0\} \right]$$

as $n \to \infty$. Thus $Q_n(\mathbf{b}) \to_D Q(\mathbf{b})$. Now the result follows from the argmax continuous mapping theorem, see e.g., Van der Vaart and Wellner (1996), page 286.

Table 1. Simulation study: comparisons of different estimation approaches. mse: mean squared error. count: number of nonzero coefficients.

Example (count)	LS		LASSO		TGDR	
	mse	count	mse	count	mse	count
1 (10)	0.351	29.9	0.654	10.0	0.153	16.2
2 (30)	0.352	30.0	3.004	7.4	0.144	29.9
3 (15)	318.9	30.0	31.90	4.7	0.079	19.5
4 (10)	1.363	29.9	2.875	8.3	0.578	21.5
5 (30)	1.450	30.0	3.174	5.9	0.531	29.6
6 (15)	1461.5	30.0	26.86	4.2	0.986	24.4

Table 2. Simulation study: bootstrap variance estimation. sd. est: standard deviations of estimates based on 100 replicates. mean. sd: mean of the estimated standard deviations.

		β_1	β_2	β_3	β_{11}	β_{12}	β_{13}
LASSO	sd. est	0.205	0.312	0.369	0.000	0.000	0.000
$u = 5.0$	mean sd.	0.213	0.333	0.358	0.003	0.003	0.000
LASSO	sd. est	0.163	0.229	0.212	0.000	0.000	0.000
$u = 10.0$	mean sd.	0.249	0.288	0.204	0.014	0.002	0.000
TGDR	sd. est	0.109	0.107	0.107	0.124	0.066	0.042
$\tau = 0.3$	mean sd.	0.101	0.095	0.093	0.112	0.070	0.058
TGDR	sd. est	0.189	0.232	0.199	0.008	0.000	0.000
$\tau = 0.8$	mean sd.	0.174	0.187	0.185	0.023	0.006	0.003

Table 3. Mantle cell lymphoma data: Genes with non-zero coefficients. Est.: estimates of coefficients. S.E.: bootstrap standard errors.

		LASSO		TGDR	
UNIQID	Genbank Accession #	Est.	S.E.	Est.	S.E.
15981	X65550	-0.088	0.099	-0.091	0.081
16312	U19769	–	–	-0.093	0.061
16541	M14535	0.066	0.077	0.157	0.105
16561	AF008552	-0.032	0.017	-0.269	0.125
17149	X97795	–	–	0.127	0.084
17326	X67155	-0.046	0.023	–	–
17434	D14134	-0.039	0.086	–	–
20232	U75689	–	–	-0.038	0.108
23972	AI370174	–	–	0.002	0.045
24376	NM_175739	0.151	0.090	0.197	0.140
24379	AF343659	0.056	0.065	–	–
24488	NM_001402	0.082	0.105	0.164	0.117
24610	X62534	–	–	-0.021	0.027
24612	AF343659	–	–	0.052	0.114
24845	M26062	0.110	0.128	0.156	0.121
24897	X07203	0.057	0.078	–	–
26192	X02747	0.059	0.070	0.038	0.056
26475	M23452	-0.031	0.027	–	–
26944	M34065	-0.030	0.065	-0.112	0.106
27095	J04088	-0.077	0.066	–	–
27108	S75311	0.159	0.094	0.100	0.116
27678	X89986	0.024	0.035	0.016	0.031
27838	D83597	–	–	-0.070	0.064
28027	NM_001880	–	–	0.051	0.098
28638	U29680	–	–	0.034	0.056
29163	U43148	–	–	0.028	0.107
29875	NM_194319	–	–	0.009	0.011
30110	L04288	0.021	0.062	0.017	0.064
30898	NM_013242	-0.067	0.057	–	–
31049	AF052573	-0.093	0.076	–	–
31081	NM_018136	–	–	-0.082	0.089
31101	AA804900	-0.108	0.102	-0.006	0.077
31837	AI361774	–	–	-0.006	0.106
32249	M65134	-0.029	0.057	-0.030	0.084
32979	AA761214	0.019	0.058	0.009	0.054
33892	AB037883	–	–	0.033	0.032

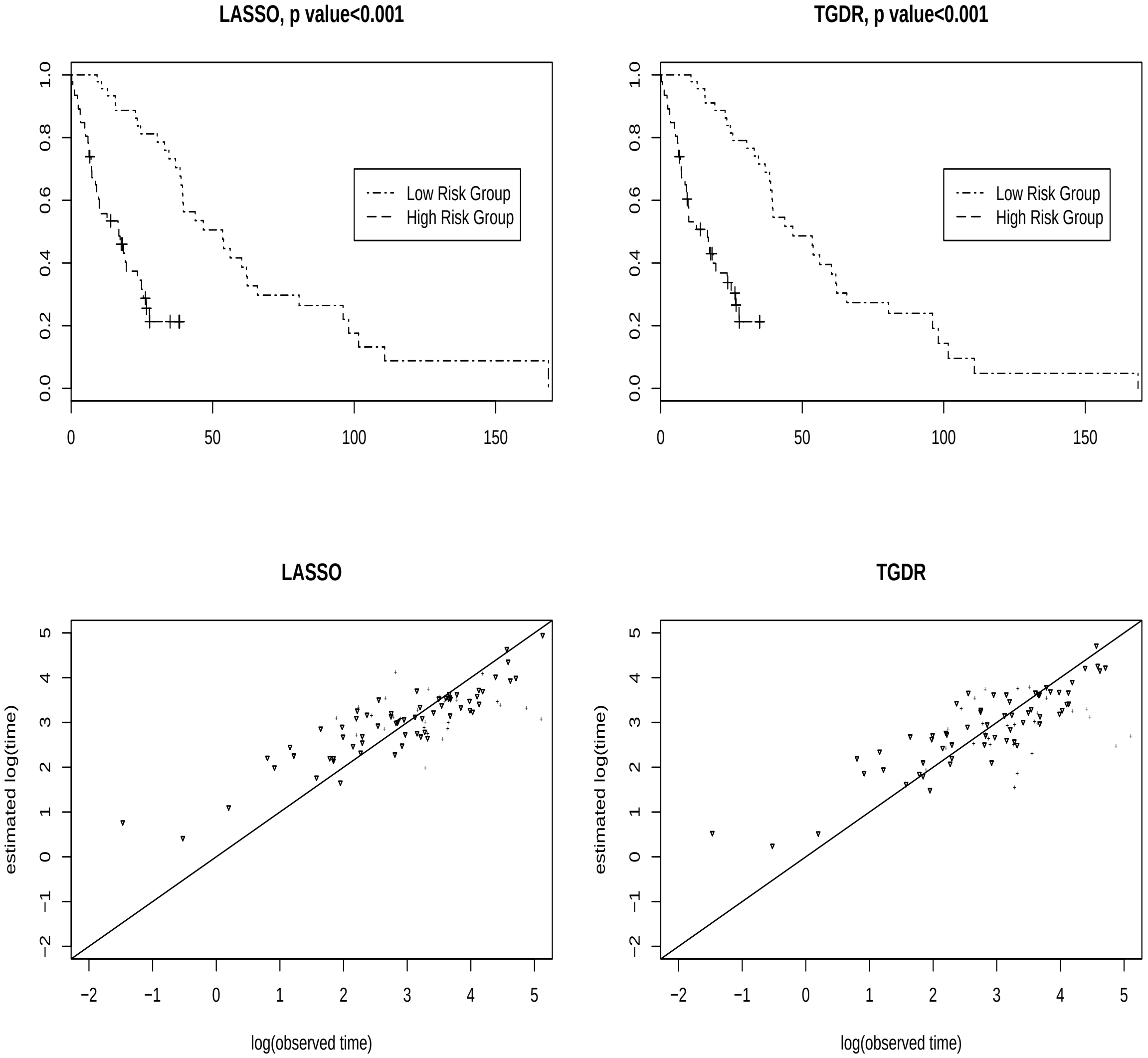

Figure 1: Mantle cell lymphoma data. Upper panels: survival curves for two risk groups. Lower panels: estimated event times versus observed times. Reversed triangles represent uncensored observations, while +'s represent censored observations.