Threshold Gradient Descent Regularization in the Additive Risk Model with High-Dimensional Covariates

Shuangge Ma[1] and Jian Huang[2]

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

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

* *e-mail:* shuangge@u.washington.edu

June 2005

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 346

Summary An additive risk model is a useful alternative to the Cox model (Cox, 1972) and may be adopted when the absolute effects, instead of the relative effects, of multiple predictors on the hazard function are of interest. In this article, we propose using the threshold gradient descent regularization (TGDR) method for covariate selection, estimation and prediction in the additive risk model for right censored survival data when the dimension of the covariate vector can be greater than the sample size. Such "small n, large p" problems may arise in the investigation of association between survival times and gene expression profiles. We propose using the V-fold cross validation and a modified Akaike's Information Criterion (AIC) for tuning parameter selections. The proposed method is demonstrated with two real data examples. The results show that the TGDR is effective in model reduction and can identify more relevant covariates than the least absolute shrinkage and selection operator approach.

KEY WORDS: Censored survival data; Cross validation; Gene expression; High-dimensional covariates; Regularization; Semiparametric model.

1. Introduction

The need for analyzing right censored survival data when the number of covariates is greater than the sample size arises in investigating association between microarray gene expression profiles and a censored survival outcome. For example, in the diffuse large B-cell lymphoma (DLBCL) study reported in Rosenwald et al. (2002), the goal is to identify genes from a total of 13413 genes printed on cDNA microarrays that are associated with time to relapse based on a data set of sample size 40. For such "small n, large p" problems, standard methods for censored survival data developed for the case where $n >> p$ are no longer directly applicable, since the parameters in the model are not estimable without any constraints or regularization. When the dimension of the covariate vector is smaller than, but comparable to the sample size, standard approaches usually yield unstable estimates, since in such cases, the observed information matrix tends to be ill-conditioned.

A fruitful approach for coping with "small n, large p" problems is via regularization. Commonly used regularization methods can be divided into two types based on the characteristics of the parameter estimates. The first type includes the ridge regression (RR) and the partial least squares (PLS) which discourage dispersion among the absolute estimated parameter values; the second type includes the least absolute shrinkage and selection operator (LASSO) and the least angle regression (LARS) that encourage dispersion among the absolute estimated parameter values and sparsity in the estimated model. Both types of regularization methods have been applied to censored survival data with high dimensional covariates. Recent studies include the application of the standard PLS method for the Cox model by Nguyen and Rocke (2002), the LARS-LASSO procedure for the Cox model by Gui and Li (2004), and the LASSO method for the additive model by Ma and Huang (2005). Although the LASSO based methods can produce sparse solutions, they may miss important covariate factors by shrinking their coefficients to zero. This is particularly problematic for genomic data, where there may exist

many covariates effects with small to moderate coefficients. In addition, if there is a group of variables among which the pair-wise correlations are high, the LASSO approaches tend to select only one variable from that group. Extensive discussions on various regularization methods and their characteristics can be found in Friedman and Popescu (2004).

In a linear regression model with high-dimensional covariates, Friedman and Popescu (2004) proposed a threshold gradient descent regularization (TGDR) method for model reduction, parameter estimation and prediction. The basic idea of this method is to first define a set of candidate models as a path in the space of joint parameter values. Then a point on this path is chosen as the final model by minimizing an appropriate objective function. The high dimensional parameter path can be traversed by varying tuning parameters. The TGDR can provide a path connecting the solutions roughly corresponding to PLS/RR and the solutions roughly corresponding to LASSO/LARS by varying the threshold values. Moderate to large threshold values create paths that involve more diverse absolute coefficient values than the PLS/RR solutions but less than the LASSO/LARS solutions. Empirical studies show that for models whose true parameters are between dense and sparse, the TGDR tends to yield more accurate predictive models.

In this article, we apply the TGDR to right censored survival data with high dimensional covariates, under the semiparametric additive risk model assumption. The additive risk model (Aalen, 1980) is a useful alternative to the Cox model (Cox, 1972) and may be adopted when the absolute effects, instead of the relative effects, of multiple predictors on the hazard function are of interest. This model has been studied extensively by many authors. In particular, Lin and Ying (1994) proposed an elegant and simple estimating equation approach for parameter estimation in the additive risk model. This estimating equation can be cast in the framework of a least squares problem. The availability of such a least squares structure makes the adaptation of the TGDR method for censored survival data computationally convenient. We

note that in the Cox model, no simple least squares structure is available, and minimization of the negative partial likelihood involves iterative reweighted least squares problems.

This article is organized as follows. In section 2, we first give a brief description of the Lin and Ying estimator for the additive risk model. We then describe the TGDR method by reformulating the Lin and Ying estimator as a solution to a least squares problem. Tuning parameter selections, which involve selecting the number of iterations needed in the gradient descent search and the value of the threshold parameter for the gradients, are based on the V-fold cross validation and a modified Akaika's Information Criterion (AIC), respectively. In section 3, we illustrate the proposed method using two data sets, one is the primary biliary cirrhosis (PBC) data set (Fleming and Harrington, 1991) and the other is the DLBCL data set (Rosenwald et al., 2002). Concluding remarks are given in section 4.

2. TGDR estimate in additive risk model

2.1 Additive risk model

Consider the additive risk model as described in Lin and Ying (1994), where the conditional hazard at time t is

$$\lambda(t|Z(\cdot)) = \lambda_0(t) + \beta' Z(t), \tag{1}$$

given a p-dimensional vector of possibly time-varying covariates $Z(\cdot)$. Here β and $\lambda_0(\cdot)$ denote the unknown regression parameter and the unknown baseline hazard function, respectively. The components in β denote the absolute change in λ per unit increase in the corresponding covariates.

Consider a random sample of size n. For the i^{th} data unit, denote $\{N_i(t) = I(X_i \leq t, \delta_i = 1); t \geq 0\}$ and $\{Y_i(t) = I(X_i \geq t); t \geq 0\}$ as the observed event process and the at-risk process, respectively. In the additive risk model (1), the parameter β can be estimated by solving the

following estimating equation

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} Z_i(t)\{dN_i(t) - Y_i(t)d\hat{\Lambda}_0(\beta, t) - Y_i(t)\beta' Z_i(t)dt\} = 0, \tag{2}$$

where $\hat{\Lambda}_0(\beta, t)$, the estimate of $\Lambda_0(t) = \int_0^t \lambda_0(u)du$, satisfies

$$\hat{\Lambda}_0(\hat{\beta}, t) = \sum_i \int_0^t \frac{\{dN_i(u) - Y_i(u)\hat{\beta}' Z_i(u)du\}}{\sum_{i=1}^n Y_i(u)}. \tag{3}$$

The resulting estimator of β is obtained by solving the equation

$$\left[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t)\{Z_i(t) - \bar{Z}(t)\}^{\otimes 2}dt\right]\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^{\infty} \{Z_i(t) - \bar{Z}(t)\}dN_i(t)\right], \tag{4}$$

where $\bar{Z}(t) = \sum_{i=1}^n Y_i(t)Z_i(t)/\sum_{i=1}^n Y_i(t)$. Denote $L^i = \int_0^{\infty} Y_i(t)\{Z_i - \bar{Z}(t)\}^{\otimes 2}dt$ and $R^i = \int_0^{\infty}\{Z_i - \bar{Z}(t)\}dN_i(t)$.

Denote the (s, l) element of L^i as $L^i_{s,l}$ and the s^{th} components of R^i and β as R^i_s and β_s, respectively. We can see that equation (4) is equivalent to the following p equations:

$$\left(\sum_{i=1}^{n} L^i_{s,1}\right)\beta_1 + ... + \left(\sum_{i=1}^{n} L^i_{s,p}\right)\beta_p = \sum_{i=1}^{n} R^i_s, \quad s = 1, \ldots, p. \tag{5}$$

It is obvious the estimate defined by (5) is the same as

$$\hat{\beta} = argmin_{\beta}\left\{M(\beta) = \sum_{s=1}^{p}\left\{\left(\sum_{i=1}^{n} L^i_{s,1}\right)\beta_1 + ... + \left(\sum_{i=1}^{n} L^i_{s,p}\right)\beta_p - \sum_{i=1}^{n} R^i_s\right\}^2\right\}. \tag{6}$$

When p is comparable to n, serious collinearity in Z may exist, and thus the estimate obtained by solving (4) may be numerically unstable. When p is larger than n, unique solution to equation (4) does not exist.

2.2 The TGDR estimate

The TGDR parameter path finding algorithm proposed by Friedman and Popescu (2004) for a linear regression model can be adapted to the additive risk model (1) as follows. Denote Δ_ν as the fixed positive infinitesimal increment, and $\nu_k = k \times \Delta_\nu$ as the index for the point along

the parameter path after k steps. Let $\beta(\nu_k)$ denote the parameter estimate corresponding to the index ν_k. For any fixed threshold value $0 \leq \tau \leq 1$, the TGDR path finding algorithm consists of the following iterative steps:

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

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

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

4. Update $\beta(\nu + \Delta_\nu) = \beta(\nu) + \Delta_\nu \times g(\nu) \times f(\nu)$ and $\nu = \nu + \Delta_\nu$.

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

The product of f and g in step 4 is component-wise. A possible variation of the above algorithm is to use the standardized negative gradient $g(\nu) = g(\nu)/max_j|g_j(\nu)|$ 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 RR/PLS. Since each increment is made in a direction in an acute angle with the negative gradient, each successive point on the parameter path has nonincreasing $M(\beta)$.

Consider an extreme case where there exist two nearly identical covariates (with the same norm and correlation coefficient $\rightarrow 1$). An ideal estimating approach should yield nearly equal coefficients for those two covariates. With the LASSO approach, it is easy to construct an example where the difference between the two estimated coefficients is bounded away from

0. With the TGDR, since nearly identical covariates have nearly equal gradients, the TGDR estimates, which are determined by the gradients, have similar values.

2.3 Tuning parameter selection

We select the tuning parameters k and τ using the following two-step approach. First we choose the tuning parameter k for any fixed τ using the V-fold cross validation (Smyth, 2001) for a pre-defined integer V. Partition the data randomly into V non-overlapping subsets of equal sizes. Choose k to minimize the cross-validated objective function

$$CV\ score(k) = \sum_{v=1}^{V} \left[M(\beta^{(-v)}) - M^{(-v)}(\beta^{(-v)}) \right], \tag{7}$$

where $\beta^{(-v)}$ is the TGDR estimate of β based on the data without the v^{th} subset for a fixed k and $M^{(-v)}$ is the function M defined in (6) evaluated without the v^{th} subset.

A byproduct of the above V-fold cross validation is the minimal CV scores (minimized over k) as a function of τ. To emphasize this relationship, we denote the minimums as $min\ CV\ score(\tau)$. Careful inspection of (7) reveals that the CV score defined here is similar to the sum of squared errors in linear regression. Considering the fact that τ directly determines the number of covariates in the model, we propose choosing τ as the solution to $argmin_\tau[n \times \log\{min\ CV\ score(\tau)/n\} + 2K(\tau)]$, where $K(\tau)$ is the number of covariates in the model for each τ. The proposed objective function mimics the AIC defined for simple linear regression. By minimizing the above objective function, the selected threshold τ provides a balance between goodness-of-fit and parsimony of model.

2.4 Model evaluation

Let $\hat{\beta}$ be the TGDR estimate with k and τ selected with the approach proposed in section 2.3. After β is estimated, the baseline cumulative hazard Λ_0 can be estimated by replacing β with $\hat{\beta}$ in (3).

For $0 \leq \tau \leq 1$ and finite k, the regularized estimates are usually not the exact least squares solutions, and the martingale structure in the residuals of the least squares estimate of Lin and Ying (1994) no longer holds. We propose using the sum of pseudo martingale residuals defined by

$$\sum_{i=1}^{n} (\delta_i - \exp(-\hat{\Lambda}_0(T_i) - \hat{\beta}' Z_i T_i))^2, \tag{8}$$

which roughly corresponds to the sum of squared errors in linear regression, for model evaluation. Loosely speaking, a better model should have smaller sum of pseudo martingale residuals.

An alternative model evaluation approach is based on the linear risk scores $\beta' Z_i$. Assume the censoring time C has density function $\phi(c)$ and is independent of Z. Then $Pr(\delta = 1|Z) = \int (1 - \exp(-\Lambda_0(c) - \beta' Z c))\phi(c)dc$. So $Pr(\delta = 1|Z)$ is a monotone function of $\beta' Z$ under mild regularity conditions. A better model should provide more accurate classification based on the estimated linear risk scores $\hat{\beta}' Z_i$. Relevant discussions can be found in Pepe, Cai and Zhang (2004). The accuracy of classification can be measured by the standard ROC (receiver operating characteristic) curves, in which case the AUC (area under curve) can be used as a single comparison criteria.

2.5 Variance approximation

The TGDR algorithm of Friedman and Popescu (2004) itself does not directly lead to variance estimation of the estimated parameters. We rewrite this algorithm in a more explicit form and propose a sandwich variance estimator. The same notations as in section 2.2 are used here.

Denote the component-wise product of $f(\nu_k)$ and $g(\nu_k)$ as $h(\nu_k)$. To simplify the notations, denote $g_k = g(\nu_k), \quad f_k = f(\nu_k), \quad h_k = h(\nu_k),$ for $k \geq 0$. Let the initial values $\beta_0 = 0$ and

$g_0 = L'R = LR$. In the threshold gradient descent search,

$$\beta_k = \beta_{k-1} + \Delta_\nu h_{k-1},\ k = 1, 2, \ldots. \tag{9}$$

Denote $G = L'L$, then the gradient update is

$$g_k = g_{k-1} - \Delta_\nu G h_{k-1},\ k = 1, 2, \ldots$$

Let T_k be a diagonal matrix whose diagonal elements are those of the vector f_k. All the nonzero elements in the diagonal of T_k are 1. Since $h_k = f_k \cdot g_k$ (component-wise product), we have $h_k = T_k g_k$. Therefore,

$$g_k = g_{k-1} - \Delta_\nu G T_{k-1} g_{k-1} = (I_p - \Delta_\nu G T_{k-1}) g_{k-1},\ k = 1, 2, \ldots,$$

where I_p is the $p \times p$ identity matrix.

Let $A_0 = I_p$, $U_j = I_p - \Delta_\nu G T_j$ and $A_k = \prod_{j=0}^{k-1} U_j,\ k = 1, 2, \ldots$. We have

$$g_k = U_{k-1} g_{k-1} = A_k g_0,\ k = 0, 1, \ldots, \tag{10}$$

It follows that

$$\begin{aligned} \beta_k &= \beta_{k-1} + \Delta_\nu T_{k-1} g_{k-1} \\ &= \beta_{k-1} + \Delta_\nu T_{k-1} A_{k-1} g_0 \\ &= \beta_0 + \Delta_\nu (T_0 A_0 + T_1 A_1 + \cdots + T_{k-1} A_{k-1}) g_0. \end{aligned} \tag{11}$$

Let $\Phi_k = T_0 A_0 + T_1 A_1 + \cdots + T_{k-1} A_{k-1}$. A more compact expression of β_k is $\beta_k = \Delta_\nu \Phi_k L' R$.

With the expression (11), we propose the following variance estimator of β_k. We first estimate $Var(R)$ by $C = \sum_{i=1}^n \int_0^n \{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dN_i(t)$. Similar to the sandwich variance estimator given in Lin and Ying (1994) and noting that L is symmetric, we propose to estimate the variance of β_k by

$$\Sigma_k = \frac{1}{n} \Delta_\nu^2 \Phi_k L C L \Phi_k'.$$

Computationally, we can use the following updating equations. For Φ_k, we use $\Phi_0 = T_0$, $\Phi_k = \Phi_{k-1} + T_{k-1}A_{k-1},\ k = 1, 2, \ldots$. For A_k, we use $A_0 = I_p,\ A_k = A_{k-1}U_{k-1},\ k = 1, 2, \ldots$. For computing T_kA_k, we note that T_k is a diagonal matrix whose non-zero elements are 1. So when it left-operates on A_k, it simply sets the all the rows of A_k corresponding to the zero rows of T_k to zero while keeping all the remaining rows intact. Similarly, for $U_k = I_p - \Delta_\nu G T_k$, T_k replaces all the columns of G corresponding to the zero columns of T_k by zeros.

In the special case of $\tau = 0$ (no truncation) and full rank L, $T_k = I_p$, and $A_k = (I_p - \Delta_\nu G)^k$,

$$\begin{aligned}
\beta_k &= \Delta_\nu[I_p + (I_p - \Delta_\nu G) + \cdots + (I_p - \Delta_\nu \Sigma)^{k-1}]L'R \\
&\to \nu[I_p - (I_p - \Delta_\nu G)]^{-1}L'R \\
&= \Delta_\nu \Delta_\nu^{-1}(L'L)^{-1}L'R \\
&= (L'L)^{-1}L'R \quad \text{as} \quad k \to \infty.
\end{aligned}$$

Therefore, when $\tau = 0$, the gradient search in the additive risk model is an iterative algorithm that computes the Lin and Ying estimator if k is allowed to converge to infinity. However, the speed of convergence can be quite slow if Δ_ν is very small. Furthermore, we note that for a finite k selected by cross validation, which may not be big enough to reach convergence, the resulting estimates can be different from the Lin and Ying estimator. At the other extreme when $\tau = 1$, the TGDR yields the most sparse solution. In addition, when $\tau = 0$ and L is of full rank, Σ_k converges as $k \to \infty$ to the sandwich variance estimator of Lin and Ying (1994).

3. Examples

3.1 PBC data

Between 1974 and 1984, the Mayo Clinic conducted a clinical trial of cirrhosis of the liver (PBC). We focus on the 276 patients with complete records. Descriptions of the covariates and a Cox model analysis can be found in Fleming and Harrington (1991). We employ the

additive risk model (1). Following the analysis of Fleming and Harrington (1991), we first make *log* transformations of the covariates *alkphos*, *bili*, *chol*, *copper*, *platelet*, *protime*, *sgot* and *trig*, so that the marginal distributions of the covariates are closer to normal. Another purpose of transformation is to make the covariates more comparable, so that the gradient descent procedure will not be dominated by a small number of covariates. Because of the relatively large sample size and moderate dimension of the covariate, this data set can be analyzed using standard methods. So it provides a test example for the proposed TGDR method as well as an opportunity to compare the TGDR method and the standard approach.

In applying the TGDR method, we select the number of iterations k in the gradient search using a 10-fold cross validation. For the threshold value τ, we only consider $\tau = 0, 0.1, \ldots, 0.9, 1.0$. With the modified AIC, the model with $\tau = 0.9$ is identified to be the best model (Table 1). Corresponding estimates are shown in Table 2. For comparison purposes, we also reproduce relevant results for the full additive model using the Lin and Ying estimator and the results of the LASSO approach from Ma and Huang (2005).

Table 1 provides the summary of the proposed cross-validation calculations with the TGDR, including the values of the threshold parameter τ used in the calculation; the results of the cross validated steps k in the gradient descent search; the pseudo martingale residual values defined in (8); the values of the AUC; the number of covariates selected; the CV scores; and the AIC scores. We see that it takes more steps in the TGDR to find the minimum of the CV scores as the threshold value τ increases. The greater the threshold value τ is, the fewer the non-zero coefficients result in the final model. Interestingly, for the PBC data, the same model is selected for $\tau = 0.5, ..., 1.0$, and this model is identical to the model selected by LASSO (Ma and Huang 2005).

Table 2 presents the estimated regression coefficients, the estimated standard errors, and the corresponding z-scores based on the Lin and Ying method, the LASSO, and the TGDR.

Of the four covariates selected by the TGDR and the LASSO, two (*age* and log(*bili*)) have the biggest z-scores in the the full model. However, the other two covariates (*stage* and log(*copper*) selected by the TGDR and the LASSO only have modest z-scores. The LASSO and the TGDR give very similar estimated coefficients, but the LASSO gives bigger standard error estimates than the TGDR. As for the classification accuracy, the TGDR is similar to the LASSO in terms of the AUC values, and both approaches perform satisfactorily as can be seen from the large AUC values. Furthermore, the model selected by the TGDR has slightly higher AUCs.

The estimates from the TGDR differ significantly from their counterparts from the full model. The Lin and Ying estimates have zero bias under the additive risk model, but have a larger sum of pseudo-martingale residual squares. On the other hand, the TGDR estimates obtained based on the cross-validation selected model selected have smaller sum of pseudo-martingale residual squares. We also calculated the 10-fold *CV score* for the Lin and Ying estimate following a procedure similar to that in section 2.3. The Lin and Ying *CV score* is 61.23, which is considerably larger than the TGDR *CV scores* of about 45.

The characteristics of the estimates can be more easily seen in Figure 1. In the top panel, we show the plot of the estimated coefficients as a function of the threshold values of τ for cross validated k. It is clear that when $\tau = 0$ or near zero, the estimated coefficients are shrunk towards zero, exhibiting behavior similar to the ridge estimates. When τ increases to one, many coefficients become zero, exhibiting behaviors similar to the LASSO estimates. The bottom panel shows the estimated coefficients as a function of k for the final model with $\tau = 0.9$. All the coefficients start from zero, but as the gradient search progresses, the non-zero estimates selected by the TGDR become distinguished from the zero ones.

3.2 DLBCL data

Rosenwald et al. (2002) reported a gene expression profiling analysis for diffuse large B-

cell lymphoma (DLBCL), in which a total of 96 normal and malignant lymphocyte samples were profiled over 17856 cDNA clones. None of the patients included in this study had been treated before obtaining the biopsy samples. Among the 42 patients, 40 had follow-up survival information, including 22 deaths with survival times from 1.3 to 71.3 months and 18 censored observations with followup times from 51.2 to 129.9 months. The objective is to identify genes whose expression levels are associated with survival.

We assume the additive risk model (1) for the conditional hazards given the gene expression values and apply the proposed TGDR approach. The underlying model assumption is that multiple genes contribute to the hazard of event independently in an additive manner. We first apply the two-step approach in Rosenwald et al. (2002). For $s = 1 \ldots 13413$, we fit marginal additive models with the expression levels for the s^{th} gene as a one-dimensional covariate. All genes with marginal p-values less than 0.01 are included in the second step additive model fitting. 122 out of 13413 genes are identified to be marginally significant at the 0.01 level. For comparison purposes, we also consider the LASSO estimate (reproduced from Ma and Huang, 2005). Because the sample size $n = 40$, which is much smaller than the number of genes in the study and is also smaller than the number of genes selected in the first stage, standard methods do not apply if we want to consider all the 122 genes jointly in the model. We apply the TGDR method for the additive risk model to this data set consisting of the selected genes from the first stage.

Table 3 shows the models identified using 10-fold cross validation for different values of τ. Similar to the results from the PBC data, more iterations are needed in the gradient search with bigger values of τ, and there are more zero coefficients with bigger values of τ. However, unlike for the PBC data, the sums of martingale residuals generally show an increasing trend. Classification based on the model with $\tau = 1.0$ is the most accurate. The model with $\tau = 1.0$ is chosen as the best one using the modified AIC. This model includes 20 genes. The results

are summarized in Table 4. The results based on the LASSO method are also included in this table. The LASSO selects 9 genes in the model. Among these 9 genes, eight are also selected by the TGDR.

For the 8 genes selected by both the TGDR and the LASSO, their estimated coefficients are different, but the differences are small, and their effects have the same direction in that the signs of these coefficients are the same. One important difference between the TGDR and the LASSO is that the former is able to pick up genes with high correlation in their expression values. For example, for genes with rank 1, 3, and 22 (ID 14837, 4899, and 15171), the correlation coefficients are: 0.78 between 1 and 3, 0.79 between 1 and 22, and 0.89 between 3 and 22. Detailed correlation structures are available upon request from the authors. All these 3 genes were selected by the TGDR. However, the LASSO only selected the gene 1. It is interesting to note that all the z-scores based on the TGDR are not significant from the conventional point of view. But this does not mean that they are not correlated with the survival time. One plausible explanation is that the sample size ($n = 40$) is quite small compared to the number of covariates in the model (122), and so the standard errors are relatively large, which reflects the large uncertainty of the estimation results in such a "small n, large p" situation.

Similar to Figure 1, Figure 2 shows the plot of the estimated coefficients as a function of the threshold values for cross validated k (top panel) and the plot of the estimated coefficients as a function of k for the final model with $\tau = 1.0$ (bottom panel). However, for the DLBCL data, the differences between the estimated coefficients with τ near or at zero and those with τ near or at one are greater. In particular, the coefficients are shrunk more towards zero for small τ and are more diverse for τ near 1. The bottom panel shows that the some coefficients become stable early in the gradient search when they reach certain non-zero values, while the others emerge later in the search.

4. Discussion

Analysis of censored survival data with a large number of covariates is an important practical problem, especially now microarrays that can assay tens of thousands of genes are becoming a routine tool in the studies of various types of cancers and many other diseases. How to associate gene expression data with clinical outcomes such as patients' survival and identify important genes or clusters of functionally related genes presents a class of interesting and challenging problems in survival analysis. In this article, we proposed using the TGDR method of Friedman and Popsecu (2004) for analyzing right censored data with high-dimensional covariates under the additive risk model assumption, based on the estimating equation developed by Lin and Ying (1994). The two real data examples illustrate that the proposed approach can effectively reduce the dimension of the covariates, while providing satisfactory classification results. TGDR is capable of adapting to the degree of sparsity of the problem via cross-validation selection of the threshold value τ. As we see from Figures 1 and 2, smaller values of τ yield dense estimates, while bigger values of τ yield more sparse estimates. This is difference from the existing regularization methods through penalization such as the RR and the LASSO which are not as adaptive to the degree of sparseness, since the functional form of the penalty plays a major role in determining the sparsity of the solution. Another useful feature of the TGDR is that it is capable of selecting a set of covariates that have similar values or are highly correlated. This is in contrast to the LASSO method, which may only pick one from the set of correlated covariates.

Several important issues remain unsolved for the TGDR estimator in the additive risk model. In particular, the sampling distribution of the estimator is unknown. Because the TGDR algorithm is highly nonlinear and martingale structure no longer exists in the estimator, the standard methods for deriving asymptotic distributions do not apply to the TGDR estimator. The distributional property of the TGDR estimator is an interesting problem and

will be pursued in the future. We have suggested a sandwich variance estimator for the estimated regression coefficients for fixed values of threshold τ and the number of iterations k, further studies are needed to evaluate the impact of cross validation in this variance estimator. Initially, we have also considered using subsampling techniques, such as the bootstrap or the jackknife, to estimate the sampling distribution of the TGDR estimator. We did not pursue this strategy since it is not clear to us whether the bootstrap or the jackknife would be theoretically valid in the present problem. Our limited numerical studies suggest that the bootstrap/jackknife variance estimates do not seem plausible. Despite the difficulty in evaluating the sampling distributional properties of the TGDR method, we have demonstrated empirically that this method provides an effective and practical way for variable selection, model estimation, and prediction in "small n, large p" problems with censored data, to which the existing methods in survival analysis are not directly applicable.

Acknowledgment

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

References

Aalen, O.O. (1980) A model for regression analysis of counting processes. *Lecture Notes in Statistics, 2.* New York: Springer-Verlag.

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

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

Gui, J. and Li, K. (2004) Penalized Cox Regression Analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Center for Bioinformatics and Molecular Biostatistics*

Lin, D.Y. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Ma, S. and Huang, J. (2005) LASSO method for additive risk models with high dimensional covariates. *Manuscript.*

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

Pepe, M.S., Cai, T. and Zhang, Z. (2004) Combining predictors for classification using the area under the ROC curve. *University of Washington Biostatistics Working Paper Series.*

Rosenwald, A., Wright, G., Chan, W., Connors, J.M., Campo, E., Fisher, R., Gascoyne, R.D., Muller-Hermelink, K., Smeland, E.B. and Staut, L.M. (2002) The use of molecular profiling to predict survival after themotheropy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine* **346**, 1937–1947.

Smyth, P (2001). Model selection of probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **10**, 63–72.

Table 1. *PBC data: model features for different threshold τ. s: cross validated number of steps. residual: sum of martingale residuals. $K(\tau)$: number of variables with non-zero coefficients.*

τ	k	residual	AUC	$K(\tau)$	CV score	AIC score
TGDR						
0.0	608	75.27	0.840	17	43.96	-473.05
0.1	614	75.45	0.838	14	44.08	-478.29
0.2	645	68.76	0.827	15	44.86	-471.45
0.3	665	73.39	0.837	6	45.10	-487.98
0.4	702	72.44	0.833	5	45.37	-488.33
0.5	829	70.43	0.827	4	45.44	-489.91
0.6	930	70.04	0.828	4	45.44	-489.91
0.7	1022	68.87	0.826	4	45.33	-490.58
0.8	1025	68.15	0.823	4	45.16	-491.61
0.9	1101	68.03	0.821	4	45.14	-491.73
1.0	1148	67.70	0.820	4	45.40	-490.15
LASSO		67.58	0.817	4	–	–
Full model		78.36	0.821	17	–	–

Table 2. *PBC data: estimated coefficients* ($\times 10$)*, standard errors* ($\times 10$) *and* $z-scores$.

	Full Model			LASSO			TGDR		
Covariate	Estimate	SE	z-score	Estimate	SE	z-score	Estimate	SE	z-score
age	0.025	0.008	3.109	0.033	0.015	2.200	0.029	0.012	2.432
alb	-0.436	0.276	-1.582	0.000	0.000	0.000	0.000	0.000	0.000
$log(alkphos)$	-0.055	0.115	-0.477	0.000	0.183	0.000	0.000	0.000	0.000
$ascites$	2.736	1.260	2.172	0.000	0.000	0.000	0.000	0.000	0.000
$log(bili)$	0.597	0.153	3.884	0.769	0.206	3.733	0.788	0.185	4.253
$log(chol)$	-0.222	0.298	-0.744	0.000	0.000	0.000	0.000	0.000	0.000
$edtrt$	0.165	0.075	2.214	0.000	0.000	0.000	0.000	0.000	0.000
$hepmeg$	-0.047	0.186	-0.253	0.000	0.000	0.000	0.000	0.000	0.000
$log(platelet)$	0.128	0.233	0.549	0.000	0.106	0.000	0.000	0.000	0.000
$log(protime)$	1.569	1.039	1.510	0.000	0.000	0.000	0.000	0.000	0.000
sex	-0.067	0.309	-0.217	0.000	0.000	0.000	0.000	0.000	0.000
$log(sgot)$	0.302	0.224	1.347	0.000	0.000	0.000	0.000	0.000	0.000
$spiders$	0.154	0.258	0.698	0.000	0.000	0.000	0.000	0.000	0.000
$stage$	0.068	0.089	0.768	0.177	0.114	1.553	0.211	0.063	3.347
trt	0.035	0.150	0.233	0.000	0.000	0.000	0.000	0.000	0.000
$log(trig)$	0.034	0.200	0.170	0.000	0.000	0.000	0.000	0.000	0.000
$log(copper)$	0.183	0.107	1.717	0.062	0.148	0.419	0.121	0.040	3.034

Table 3. *DLBCL data: model features for different threshold τ. s: cross validated number of steps. residual: sum of martingale residuals. $K(\tau)$: number of variables with non-zero coefficients.*

τ	k	residual	AUC	$K(\tau)$	CV score	AIC score
TGDR						
0.0	1200	10.48	0.869	122	182.1	304.63
0.1	1229	10.50	0.866	119	182.4	298.69
0.2	1338	10.55	0.864	114	182.6	288.74
0.3	1461	10.46	0.851	96	185.4	253.35
0.4	1727	10.45	0.846	74	188.5	210.01
0.5	2398	10.52	0.843	51	191.9	164.72
0.6	4960	11.42	0.841	40	196.1	143.59
0.7	10665	12.54	0.833	32	195.4	127.45
0.8	70450	14.89	0.917	29	189.3	120.18
0.9	141102	15.51	0.929	26	181.9	112.58
1.0	163200	15.22	0.927	20	176.8	99.45
LASSO		5.152	0.861	9	–	–

Table 4. *DLBCL data: ID: gene ID. Rank: rank based on marginal p-values. Estimated coefficients* ($\times 10$)*, standard errors* ($\times 10$) *and* $z-$*scores.*

		LASSO			TGDR		
Rank	ID	Estimate	SE	z-score	Estimate	SE	z-score
1	14837	0.058	0.105	0.553	0.022	0.041	0.543
3	4899	–	–	–	-0.022	0.094	-0.243
9	515	–	–	–	0.012	0.069	0.181
12	17879	0.056	0.086	0.647	0.441	0.776	0.569
16	12822	0.319	0.188	1.698	0.502	1.047	0.479
22	15171	–	–	–	-0.063	0.236	-0.266
29	12600	–	–	–	-0.756	3.046	-0.248
39	1591	–	–	–	0.081	0.423	0.192
62	4253	–	–	–	0.063	0.486	0.129
67	2689	–	–	–	0.478	1.217	0.393
68	2060	–	–	–	-0.079	0.257	-0.307
73	19274	-0.479	0.369	-1.297	-0.487	0.955	-0.510
88	2059	-0.740	0.389	-1.899	-0.842	1.551	-0.543
91	19307	0.417	0.227	1.833	–	–	–
95	14140	0.271	0.141	1.922	0.360	0.693	0.519
100	21333	–	–	–	0.208	0.600	0.348
101	10411	–	–	–	0.003	0.149	0.021
102	19282	-0.332	0.220	-1.510	-0.624	1.411	-0.442
103	14049	0.021	0.087	0.244	0.143	0.276	0.517
106	17499	–	–	–	0.127	0.343	0.370
110	15583	–	–	–	0.073	0.206	0.355

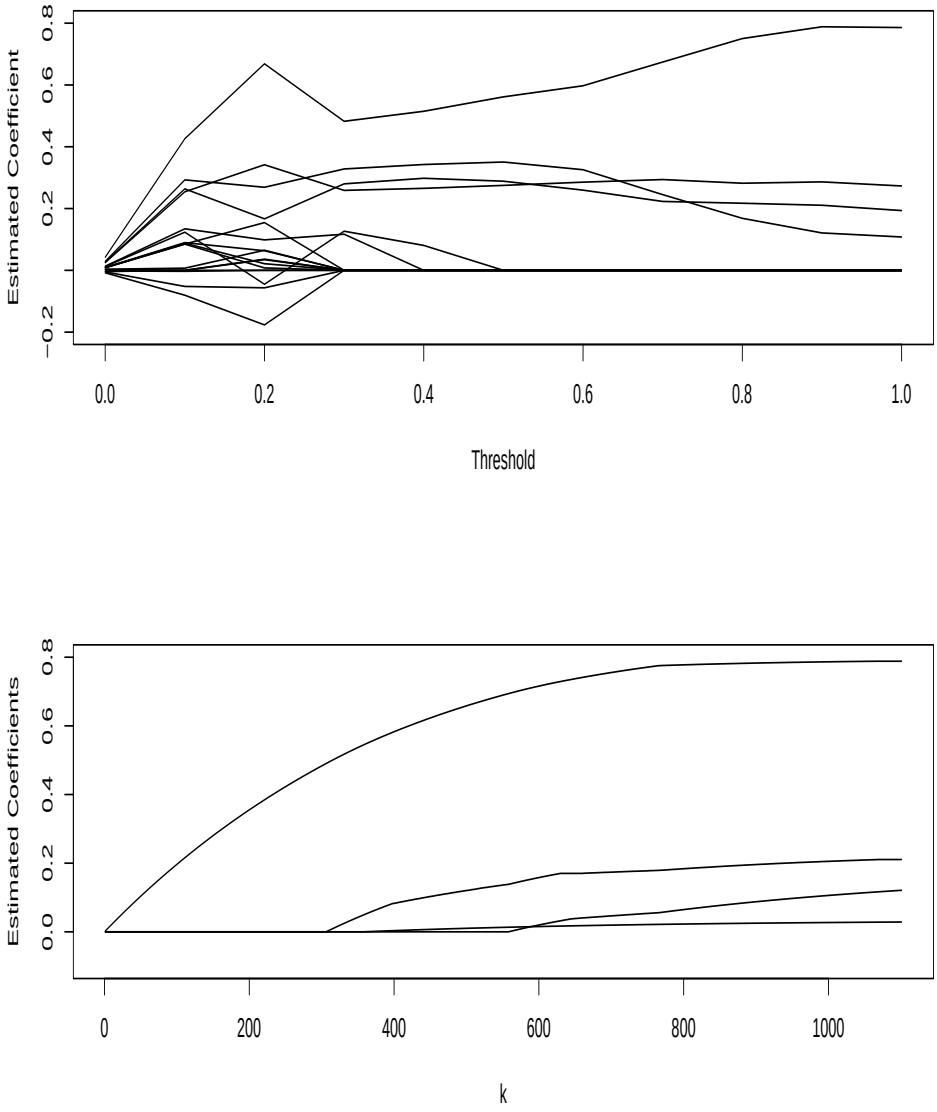

Figure 1: PBC data. Upper panel: estimated coefficients as a function of the threshold value. Lower panel: estimated coefficients as a function of step for $\tau = 0.9$.

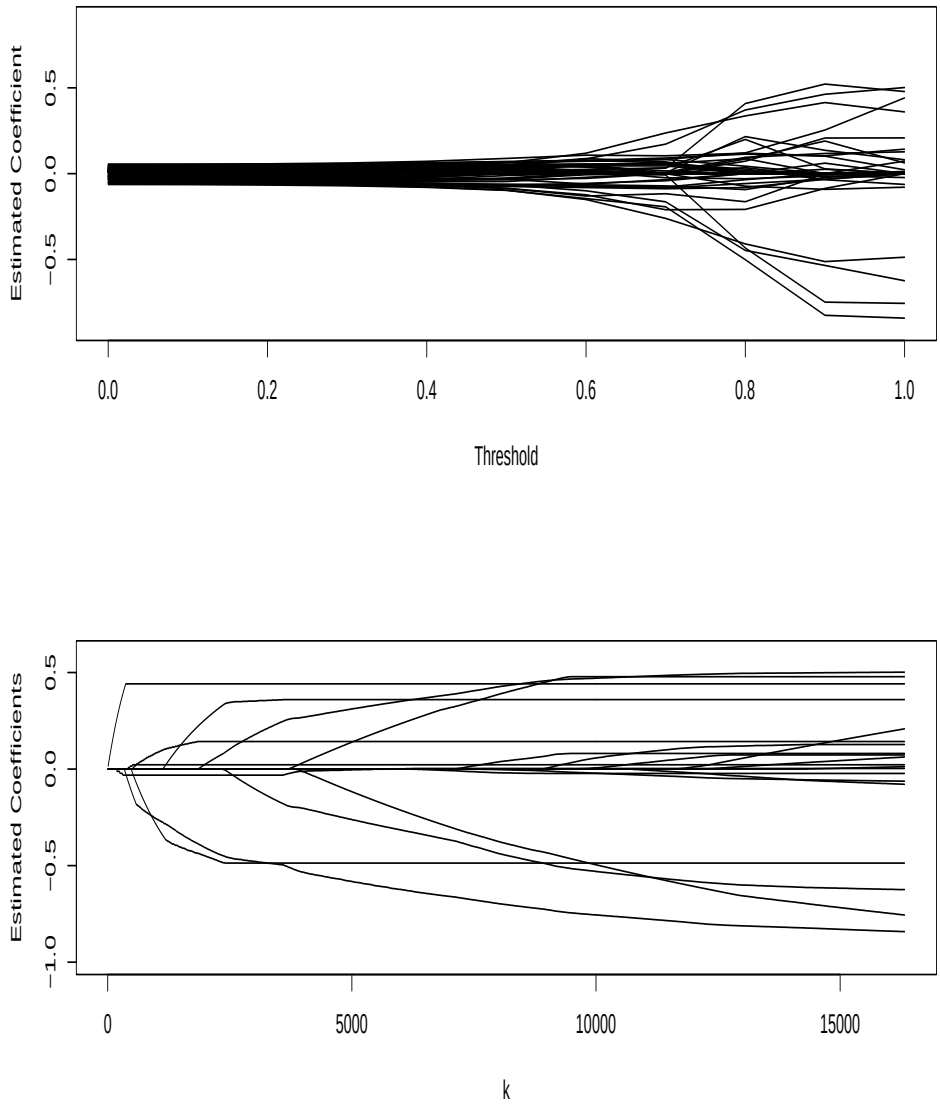

Figure 2: DLBCL data. Upper panel: estimated coefficients as a function of the threshold value. Lower panel: estimated coefficients as a function of step for $\tau = 1.0$.