LASSO Method for Additive Risk Models 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

June 2005

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 347

SUMMARY. The additive risk model is a useful alternative to the proportional hazards model. It postulates that the hazard function is the sum of the baseline hazard function and the regression function of covariates. In this article, we investigate estimation in the additive risk model with right censored survival data and high dimensional covariates. A LASSO (least absolute shrinkage and selection operator) approach is proposed for estimating the regression parameters. We propose using the L_1 boosting algorithm, which is computationally affordable and relatively easy to implement, to compute the LASSO estimates. The V-fold cross validation is applied to select the tuning parameter and the weighted bootstrap is used to estimate the variances of the LASSO estimators. The proposed approach is illustrated with analysis of the PBC clinical data and the DLBCL genomic data. It is shown that this approach can provide interpretable and sparse predictive models with satisfactory predication and classification properties.

KEY WORDS: Additive risk model; Boosting; LASSO; Variable selection.

1. Introduction

Survival analysis for right censored data with high dimensional covariates has drawn extensive attentions in recent years. This article is partly motivated by the studies of linking gene expression profile to censored survival outcome, for example time to cancer relapse (Alizadeh et al., 2000; Garber et al., 2001; Rosenwald et al., 2002). Consider the study of genome-wide gene expression profiling analysis for diffuse large B-cell lymphoma (DLBCL) reported by Alizadeh et al. (2000), where the goal is to identify the statistical influence of molecular features of the tumor on the survival of patients. Gene expression levels of 13413 clones and survival information are measured for 40 patients. Statistically, standard estimation approaches cannot yield a unique estimator, when the dimension of covariates is greater than the sample size. Biologically, it is reasonable to assume that only a small number of genes are relevant to predicting the phenotype. Model reduction is necessary before any downstream analysis.

Dimension reduction has been extensively investigated for linear regression models (Miller, 1990; Helland and Almoy, 1994; Roecker, 1991; Jolliffe, 1986). One widely used approach is to use low dimensional projections of the covariates as surrogates for the true covariates. Examples include the ridge regression, the partial least squares (PLS) technique, and the principal component regression (for a detailed discussion, see Jolliffe, 1986). However, including all the covariate effects in the predictive models through projections may introduce noises which may lead to poor predictive performance, and it may difficult to interpret such models. An alternative approach is to use model selection techniques, for example the step-wise variable selection method, to choose important covariates. This can usually be accomplished by using penalization methods. A general approach is outlined in Fan and Li (2001). It has been noted that penalization methods may be numerically unstable with large numbers of covariates having small to moderate effects. Generally speaking, neither of these strategies dominates the

other, and how well a method works needs to be evaluated on case by case basis.

Modeling survival data with high dimensional covariates is more challenging due to the presence of censoring and the use of complicated semiparametric models. One approach used by Alizadeh et al. (2000) with microarray data is to cluster genes first, and then use the sample averages of the gene expression levels in a Cox model for right censored survival outcome. Another well developed clustering based algorithm is the gene harvesting procedure of Hastie et al. (2001). Nguyen and Rocke (2002) apply the standard PLS method and use the resulted PLS components in the Cox model. Li and Luan (2003) consider a penalized estimation procedure for the Cox model using kernels, under the assumption that the covariate effects are smooth functions of gene expression levels. Tibshirani (1997) and Gui and Li (2004) investigate LASSO (least absolute shrinkage and selection operator, Tibshirani 1996) type estimates for the Cox model with right censored data. In a recent study, Ma, Kosorok and Fine (2004) apply the principal component regression to additive risk models with right censored data.

An additive model is generally adopted when it is reasonable to assume that the covariate effects under consideration contribute additively to the conditional hazard. Consider one special form of the additive risk models studied in Lin and Ying (1994), where we model the conditional hazard at time t by

$$\lambda(t|Z) = \lambda_0(t) + \beta_0' Z, \tag{1}$$

given a d-dimensional vector of time-independent covariates Z. Here β_0 and $\lambda_0(\cdot)$ denote the unknown regression parameter and the unknown baseline hazard function, respectively. Previous studies have concluded its sound biological and empirical bases (Breslow and Day, 1987) and satisfactory statistical properties (Lin and Ying, 1994; Huffer and McKeague, 1991; McKeague and Sasieni, 1994).

Inspired by the special form of the estimating equations (Lin and Ying, 1994), we propose a LASSO type estimate, which minimizes a least-squares-type objective function subject to a L_1 constraint, for the additive risk model (1) when the dimension of the covariate is high. Because of the nature of the L_1 constraint, the LASSO method shrinks coefficients and produces some coefficients that are exactly zero, thus it can yield a sparse and interpretable model.

The goal of the current study is to develop a computationally affordable and well-behaved estimating approach, which can effectively reduce the dimension of the covariates for right censored survival data, under the additive risk model assumption. In Section 2, we first give a brief description of the additive hazard model, describe the LASSO method in this model, and then use the V-fold cross validation for tuning parameter selection and the weighted bootstrap for inference. In Section 3, we propose using a L_1 boosting algorithm to compute the LASSO estimates in the additive hazard model. The proposed approach is demonstrated with two examples in section 4. Concluding remarks are in section 5.

2. LASSO estimate in additive risk model

2.1 Additive risk model

Consider a set of n independent observations $(T_i, C_i, Z_i), i = 1, \ldots, n$. Suppose that the ith subject's event time T_i is conditionally independent of the censoring time C_i, given the d-dimensional covariate vector Z_i. For simplicity of notations, we consider time-independent Z only throughout this paper unless otherwise specified. Let $X_i = min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ for right censored data. We assume the additive risk model (1). Other format of additive risk models have been studied in Huffer and McKeague (1991) and Mckeague and Sasieni (1994). For the ith subject, 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. The regression coefficient β_0 can be estimated by solving the following estimating equation:

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

where $\hat{\Lambda}(\beta, t)$ is the estimator of Λ_0 satisfying

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

The resulting estimator of β_0 satisfies the simple equation

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

where $\bar{Z}(t) = \sum_{i=1}^n Y_i(t) Z_i / \sum_{i=1}^n Y_i(t)$. Denote $L^i = \int_0^{\infty} Y_i(t)\{Z_i - \bar{Z}(t)\}^{\otimes 2} dt$ and $R^i = \int_0^{\infty} \{Z_i - \bar{Z}(t)\} dN_i(t)$. Ls are symmetric semi-positive-definite matrices with rank equal to 1.

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

2.2 The LASSO estimate

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

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

The validity of the estimating equation (4) does not depend on any assumption of d and n. The similarity between the estimating equations (5) and the normal equations for simple linear models motivates model reduction for the additive risk model with the following LASSO type estimator:

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

subject to the L_1 constraint that $|\beta|_1 = |\beta_1| + \ldots |\beta_d| \leq u$, for a data-dependent tuning parameter u, which indirectly determines how many covariates have zero coefficients. For $n \geq d$, denote $\hat{\beta}^*$ as the minimizer of $M(\beta)$ without the L_1 constraint. If $u \geq \sum |\hat{\beta}^*_s|$, the solution to (6) with constraint is the usual unbiased estimate. Otherwise, $\hat{\beta}$ may be biased. However, $\hat{\beta}$ may have "mean squared errors" smaller than $\hat{\beta}^*$ because of the bias variance trade off inherent to the penalization methods.

One unique characteristic of the LASSO estimate in the additive risk model is that the summation in (6) is over d, the dimension of covariates, not over the sample size as in the linear regression model. However, considering the equivalence of (6) and (4), the LASSO estimate defined in (6) can provide model reduction in the β space. The simplicity of the estimating equation in (6) for the additive risk model is not shared by other survival models. For the Cox model in Tibshirani (1997), a weighted least squares approximation to the partial likelihood function and an iterative computational algorithm are needed.

Occasionally, there may exist certain covariate effects that are known to be effective *a priori*. In this case interest lies in more accurate adjustment for other covariate effects and shrinkage of coefficients (of effective covariates) is not preferred. In such an instance, one may simply omit the corresponding β_s' from the L_1 constraint. The L_1 boosting algorithm discussed in section 3 can be applied to such situations with minor modifications.

2.3 Tuning parameter selection

We propose choosing the tuning parameter u with the following V-fold cross validation (Wahba, 1990) for a pre-defined integer V. Partition the data randomly into V non-overlapping subsets of equal sizes. Chose u to minimize the cross-validated objective function

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

where $\hat{\beta}^{(-v)}$ is the LASSO estimate of β based on the data without the v^{th} subset for a

fixed u and $M^{(-v)}$ is the function M defined in (6) evaluated without the v^{th} subset. Compared with the usual leave-one-out cross validation (Huang and Harrington, 2002), the V-fold cross-validation is preferred here because of its computational simplicity and sound theoretical properties (van der Laan, Dudoit and Keles, 2003). Another possible tuning parameter selection technique is the generalized cross validation as used in Tibshirani (1997). Relative efficacy of different validation techniques is of interest but is beyond the scope of the present article.

2.4 Inference

Inference based on the LASSO estimator is done using the special form of the GCV score in Tibshirani (1997). It is not clear how to extend that approach to the estimate proposed here. As an alternative, we consider the following weighted bootstrap based inference.

Let W denote a positive random variable satisfying $E(W) = 1$ and $var(W) = 1$, and let $w_1, \ldots, w_n$ be n i.i.d. copies of W. Denote $L^i(w) = \left[\sum_{i=1}^n \int_0^\infty Y_i(t)\{Z_i - \bar{Z}_w(t)\}^{\otimes 2} dt\right]$ and $R^i(w) = \left[\sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}_w(t)\} dN_i(t)\right]$, where $\bar{Z}_w(t) = \sum_{i=1}^n w_i Y_i(t) Z_i / \sum_{i=1}^n w_i Y_i(t)$. Consider the weighted LASSO estimator $\hat{\beta}_w$ satisfying

$$\hat{\beta}_w = argmin \left\{ \sum_{s=1}^d \left\{ \left(\sum_{i=1}^n w_i L^i_{s,1}(w) \right) \beta_1 + \ldots + \left(\sum_{i=1}^n w_i L^i_{s,d}(w) \right) \beta_d - \sum_{i=1}^n w_i R^i_s(w) \right\}^2 \right\},$$

subject to the L_1 constraint that $|\beta|_1 = |\beta_1| + \ldots |\beta_d| \leq u$, where u is the same as the tuning parameter value for the estimator defined by (6).

To estimate the variance of $\hat{\beta}$, we generate positive random weights $w_i, i = 1, \ldots, n$. Then as described above, we compute the weighted estimator $\hat{\beta}_w$. This procedure is repeated K times. The variance of $\hat{\beta}$ can be estimated by the sample variance of $(\hat{\beta}_{w,1}, \ldots, \hat{\beta}_{w,K})$.

The weighted bootstrap technique has been shown to be effective in estimation of the additive risk model using principal components of the covariates (Ma, Kosorok and Fine,

2004). Related theoretical studies of weighted bootstrap for general M-estimators can be found in Ma and Kosorok (2004).

2.4 Model evaluation and comparison

For the estimators in the additive hazards model without constraints on the parameter proposed by Lin and Ying (1994), model evaluation can be built based on the martingale structure, which has not been proven to hold for the LASSO estimators. In Gui and Li (2004), time-dependent ROC (receiver-operator-characteristic) curves are used to compare performances of different models. Time-dependent ROC may be difficult to interpret, since it is a function of time and a single test statistic cannot be easily constructed.

Although the martingale structure may not hold, the "pseudo martingale residual" defined by $\sum_{i=1}^{n}(\delta_i - \exp(-\hat{\Lambda}(T_i) - \hat{\beta}' Z_i T_i))^2$ can still be used for model evaluation. A better model should have smaller pseudo martingale residual. The pseudo martingale residual defined here corresponds to the sum of squared errors in a linear regression model. An alternative model evaluation approach can be based on the linear risk scores $\hat{\beta}' Z_i$. Assume the censoring time C has density function $g(c)$ and is independent of Z. Then $Pr(\delta = 1|Z) = \int (1 - \exp(-\Lambda_0(c) - \beta_0' Zc))g(c)dc$. So $Pr(\delta = 1|Z)$ is a monotone function of $\beta_0' Z$ 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 curves, in which case the AUC (area under curve) can be used as a single comparison criteria.

It is also of interest to quantify the similarity between different models with possibly different set of covariates. We measure this by the correlation between the estimated linear scores based on different set of covariates. A strong correlation suggests similar classification performance.

3. Computational algorithm

The L_1 constraint is equivalent to adding a L_1 penalty to the objective function and ignoring the constraint (Tibshirani, 1997). Since the L_1 penalty is not differentiable, usual derivative-based minimization techniques (for example the Newton-Raphson) cannot be used to obtain the estimate in (6). In most previous studies, the minimization relies on the quadratic programming (QP) or general non-linear program which are known to be computationally intensive. Moreover, the quadratic programming procedure cannot be applied directly to the settings when the sample size is much smaller than the number of predictors.

Recent study by Kim and Kim (2004), which relates the minimization step for the LASSO estimate to the L_1 boosting algorithm, a regularized boosting algorithm proposed by Mason et al. (2000), provides a computationally more feasible solution. The L_1 boosting algorithm can be applied to general objective functions (other than least squares criterion) with L_1 constraints. For the current L_1 constrained estimator defined in (6) with a fixed u, this algorithm can be implemented in the following steps:

1. Initialization $\beta_s = 0$ for $s = 1 \dots d$ and $m = 0$.

2. With the current estimate of $\beta = (\beta_1, \dots, \beta_d)$, compute

$$\phi_k(\beta) = \sum_{s=1}^{d} \left\{ \left(\sum_{i=1}^{n} L_{s,1}^i \right) \beta_1 + ... + \left(\sum_{i=1}^{n} L_{s,d}^i \right) \beta_d - \sum_{i=1}^{n} Y_s^i \right\} \times \left(\sum_{i=1}^{n} L_{s,k}^i \right)$$

   for $k = 1 \dots d$.

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

4. Otherwise denote $\gamma = -sign(\phi_{k^*}(\beta))$. Find $\hat{\alpha}$ that

$$\hat{\alpha} = argmin_{\alpha \in [0,1]} M((1-\alpha)(\beta_1, \dots, \beta_d) + \alpha \times u \times \gamma \eta_{k^*}),$$

   where η_{k^*} has the k^{*th} element equals to 1 and the rest equal to 0.

5. Let $\beta_k = (1 - \hat{\alpha})\beta_k$ for $k \neq k^*$ and $\beta_{k^*} = (1 - \hat{\alpha})\beta_{k^*} + \gamma u\hat{\alpha}$. 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 in (6). We conclude convergence if the absolute value of $\phi_{k^*}(\beta)$ computed in step 3 is less than a pre-defined criteria, and/or if $M(\beta)$ is less than a pre-defined threshold.

Compared with traditional algorithms, the L_1 boosting only involves evaluations of simple functions. Data analysis experiences show the computational burden for the L_1 boosting is minimal. As pointed out in Kim and Kim (2004), one attractive feature of the L_1 boosting algorithm is that the convergence rate is independent of the dimension of input. This property of convergence rate is essential to the proposed approach since weighted bootstrap is used. Kim and Kim (2004)provide an example where convergence can be achieved after 50 iterations for a data set with a 36-dimensional covariate vector. On the other hand, it has been known that for general boosting methods, over-fitting usually does not pose a serious problem (Friedman, Hastie and Tibshirani, 2000). So the overall iteration N can be taken to be a large number to ensure convergence.

4. Examples

4.1 PBC data

Between 1974 and 1984, the Mayo Clinic conducted a double-blinded randomized clinical trial in primary cirrhosis of the liver (PBC). Values of 18 biological covariates were measured for 312 randomized patients. We focus on the 276 randomized patients with complete records only. Descriptions of the experiment and data analysis can be found in Fleming and Harrington (1991), where the Cox model is assumed.

As an alternative, we consider the additive risk model for the PBC data. *log* transformations of the covariate *alkphos*, *bili*, *chol*, *copper*, *platelet*, *protime*, *sgot* and *trig* are first

made, so that the marginal distributions of those covariates are closer to the Normal distribution. Another reason for transformation is that the LASSO method usually requires initial standardization of the regressors, so that the penalization scheme is comparable for all regressors.

Model (1) is assumed for the relationship between survival time and the transformed covariates. The LASSO estimate is obtained from (6), with u selected by the 10-fold cross validation. After β is estimated, the cumulative baseline can be estimated using (3). For the purpose of comparison, we also consider the full additive model, where all covariates (even if not significant) are included in the additive model, and the simplified model where only significant covariates identified by the backwards step-wise approach are included.

The LASSO approach with $u = 0.195$ (Figure 1) yields an estimate with only four non-zero components (Table 1). It is interesting that the non-zero components from the LASSO are not necessarily significant in the full model. The estimates from the LASSO and from the step-wise approach may differ significantly. This has also been noticed in Tibshirani (1997). Inference results with the weighted bootstrap proposed in section 2.4 and exponential weights are also shown in Table 1. We can see that the stepwise procedure mostly inflates the Z scores of chosen variables relative to the full model, and the LASSO mainly shrinks them towards zero. Similar phenomena has been observed for the Cox model (Tibshirani, 1997). This similarity observation partly supports the validity of the proposed inference procedure.

The pseudo martingale residuals are 78.36 (for the full model), 70.03 (for the step-wise approach) and 67.58 (for the LASSO estimate). The LASSO approach has smaller "sum of squared errors", although the differences are small. We generate two hypothetical risk groups based on the estimated linear risk scores in a manner that there are equal number of subjects in two risk groups. The survival curves for the two groups are shown in Figure 1, which demonstrates the effectiveness of the proposed LASSO approach for classifying subjects into

different risk groups. This argument is supported by the boxplot of the linear risk scores for subjects with $\delta = 1$ and $\delta = 0$ (Figure 1). The classification performance is evaluated using ROC (Figure 1). We can see that for the purpose of classification the differences between the LASSO, the full model and the step-wise model are negligible for the PBC data.

4.2 Application to DLBCL data

Alizadeth et al. (2000) reported a genome-wide gene expression profiling analysis for diffuse large B-cell lymphoma (DLBCL), in which a total of 96 normal and malignant lymphocytes samples were profiled over 17856 cDNA clones. None of the patients included in this study had been treated before obtaining the biopsy samples. 40 patients had follow-up survival information, including 22 deaths with survival time ranging from 1.3 to 71.3 months and 18 alive, with followup time ranging from 51.2 to 129.9 months. The goal is to identify genes whose expression levels are significantly associated with survival. Global normalization of gene expressions is carried out, so that different genes are comparable.

We assume the additive risk model (1) and apply the proposed LASSO approach. The underlying biological assumption is that multiple genes contribute to the hazard of event independently in an additive manner. The proposed approach has no computational or methodological limitation on the number of genes that can be used in the prediction of patients' event times. However, if we apply the proposed approach directly, manipulation of 13413-dimensional matrices is needed. To avoid the instability caused by using existing software, we 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. Similar approach has been extensively used in previous studies (see Li and Luan, 2003 for reference). 122 out of 13413 genes are identified to be marginally significant at the 0.01

level.

Apply the proposed LASSO approach with $n = 40$ and $d = 122$. With $u = 0.28$ (estimated from the 10-fold cross validation), only 9 out of 122 coefficients are not zero. We can see from Table 2 that the genes identified by the LASSO approach are not necessarily marginally most significant. Inference is carried out with the proposed weighted bootstrap and the exponentially distributed weights. The non-zero coefficients estimated from LASSO are not necessarily statistically significant. Classification based on the linear risk scores estimated from the LASSO approach is shown in Figure 2. The difference between the survival curves for the two risk groups (created based on the LASSO estimated linear risk scores) are significant. We can also see the obvious difference between the linear risk scores for the groups with $\delta = 1$ (uncensored) and $\delta = 0$ (censored) in Figure 2.

For comparison, we consider an additive model with the nine marginally most significant genes (so that two models will have the same degrees of freedom). The estimates are also shown in Table 2. The ROC curves based on the linear risk scores from LASSO and from the nine most significant genes are shown in Figure 2. Two approaches have similar classification performance (measured by the AUC). Roughly speaking, the LASSO estimate for the additive model has similar classification power (measured by the AUC) as the LARS-LASSO estimate for the Cox model in Gui and Li (Figure 3 of Gui and Li, 2004). The corresponding pseudo martingale residuals are 5.198 (nine most significant) and 5.152 (LASSO). The correlation coefficient between the two sets of estimated linear risk scores is 0.73, which suggests a moderate similarity.

In the above DLBCL data analysis, we assume that linearity is reasonable for all gene effects. The validity of this assumption may need to be checked. We leave this important, but not quite relevant, issue to future study.

5. Concluding remarks

Right censored survival data with high dimensional covariates is analyzed with LASSO type estimates in this article, under the additive risk model assumption. This procedure can be applied to identify important covariates that are related to patients' survival outcome. For simplicity of notations, we assume time-independent covariate Z with only minor modifications. It can be seen that the same technique is applicable to time-dependent covariate Z.

In Lin and Ying (1994), inference for the estimate of β_0 can be constructed based on the martingale structure. For the LASSO estimate with the Cox model, Tibshirani (1997) proposes an approximated variance estimation based on a linear estimate of the regression parameters. Since the summation in the objective function $M(\beta)$ is over d instead of n, it is unlikely that Tibshirani's approach will hold here. We propose using the weighted bootstrap for inference in this article. Limited data analysis show the weighted bootstrap can produce reasonable estimates.

With the proposed LASSO approach, we are able to identify individual covariate effects. However, the tradeoff is that the number of covariate effects can be evaluated is limited by the sample size. In principal, this method can identify up to $min(n-1, d)$ covariate effects. This limitation is especially important for data like the DLBCL, where the dimension of the covariates is much larger than the sample size. If it is biologically reasonable to believe the number of covariates significantly related to survival is comparable to or larger than the sample size, then transformation of the covariates will be firstly needed. One possibility is to use the principal components (as in Ma, Kosorok and Fine, 2004), where the transformed covariates correspond to "super genes" for the genetic data.

We used the same tuning parameter u for different bootstraps. An alternative is to choose the tuning parameter for each bootstrap based on a weighted cross validation score. For a different penalization model, Ma and Kosorok (2004) argue that the tuning parameters selected

from minimizing the weighted and the usual cross validation scores should be asymptotically of the same order. We expect similar results to hold here: fixing the tuning parameter should give the same asymptotic results as allowing for different tuning parameters for each bootstrap.

In this article, we propose using the L_1 boosting for computing the L_1 constrained estimator in the additive hazard model. The computational efficacy of the L_1 boosting technique for LASSO has been thoroughly discussed in Kim and Kim (2004). Our own experience indicates that the L_1 boosting technique may converge very slowly near the optimum. Even for the case when $n > d$ and u is large enough, that is when exact solution to (6) is expected, it may take many iterations for the L_1 boosting to achieve the exact solution. However, the difference between the L_1 boosting solution and the exact solution estimates is negligible.

There exist several promising methodologies for linking high dimensional covariates to survival type responses. At this point, simulation studies and data analysis are still too limited to draw conclusions about the relative efficacy of different approaches. A comprehensive comparison of different methods is of interest for future study.

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.

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.

Breslow, N.E. and Day, N.E. (1987). *Statistical Models in Cancer Research, 2.* Lyon: IARC.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties.*Journal of the American Statistical Association* **96**, 1348–1360.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28**, 337–374.

Garber, M.E., Troyanskaya, O.C., Schluens, K., Petersen, S., Thaesler, Z., Pacyna Gengelbach, M., van de Rijn, M., Rosen, G.D., Perou, C.M., Whyte, R.I., Altman, R.B., Brown, P.O., Bostein, D. and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of National Academy of Science USA* **98**, 13784–13789.

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*

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A.A., Levy, R., Staudt, L.,

Chan, W.C., Bostein, D. and Brown, P. (2001) Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **2**, 1–21.

Helland, I.S. and Almoy, T. (1994) Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association* **89**, 583–591.

Huang, J. and Harrington, D. (2002) Penalized partial likelihood regression for right-censored data with bootstrap selection of the Penalty Parameter. *Biometrics* **58**, 781–791.

Huffer, F.D. and McKeague, I.W (2003) Weighted least squares estimation for Aalen's additive risk model. *Journal of the American Statistical Association* **86**, 114–129.

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

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

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

Li, H.Z. and Gui, J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* **20**, 208–215.

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

Ma, S. and Kosrok, M.R. (2004) Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis, In press.*

Ma, S., Kosorok, M.R., and Fine, J.P. (2004) Additive risk models for survival data with high dimensional covariates. *UW Madison Biostatistics and Medical Informatics TR186.*

Mason, L., Baxter, L., Bartlett, P. and Frean, M. (2000) Functional gradient techniques for combining hypotheses. *Advances in Large Margin Classifiers* Cambridge: MIT press.

McKeague, I.W. and Sasieni, P.D. (1994) A partly parametric additive risk model. *Biometrika* **81**, 501–514.

Miller, A. (1990) *Subset Selection in Regression.* London: Chapman and Hall.

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.*

Roecker, E.B. (1991) Prediction error and its estimation for subset selected models. *Technometrics* **33** 459–468.

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.

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.

van der Laan, M.J., Dudoit, S. and Keles, S. (2003) Asymptotic optimality of likelihood-based cross validation. *Technical Report, Division of Biostatistics, University of California.*

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

Table 1. *Estimation of regression parameters for the PBC data with the additive risk model.*

	Full Model			Step-wise			LASSO		
Covariate	Estimate (×10)	SE (×10)	z-score	Estimate (×10)	SE (×10)	z-score	Estimate (×10)	SE (×10)	z-score
age	0.025	0.008	3.109	0.027	0.008	3.452	0.033	0.015	2.200
alb	-0.436	0.276	-1.582	–	–	–	0.000	0.000	0.000
$log(alkphos)$	-0.055	0.115	-0.477	–	–	–	0.000	0.183	0.000
$ascites$	2.736	1.260	2.172	3.074	1.201	2.559	0.000	0.000	0.000
$log(bili)$	0.597	0.153	3.884	0.654	0.130	5.041	0.769	0.206	3.733
$log(chol)$	-0.222	0.298	-0.744	–	–	–	0.000	0.000	0.000
$edtrt$	0.165	0.075	2.214	1.906	0.709	2.687	0.000	0.000	0.000
$hepmeg$	-0.047	0.186	-0.253	–	–	–	0.000	0.000	0.000
$log(platelet)$	0.128	0.233	0.549	–	–	–	0.000	0.106	0.000
$log(protime)$	1.569	1.039	1.510	–	–	–	0.000	0.000	0.000
sex	-0.067	0.309	-0.217	–	–	–	0.000	0.000	0.000
$log(sgot)$	0.302	0.224	1.347	–	–	–	0.000	0.000	0.000
$spiders$	0.154	0.258	0.698	–	–	–	0.000	0.000	0.000
$stage$	0.068	0.089	0.768	–	–	–	0.177	0.114	1.553
trt	0.035	0.150	0.233	–	–	–	0.000	0.000	0.000
$log(trig)$	0.034	0.200	0.170	–	–	–	0.000	0.000	0.000
$log(copper)$	0.183	0.107	1.717	0.217	0.103	2.111	0.062	0.148	0.419

Table 2. *Estimation of regression parameters for the DLBCL data with the additive risk model. ID: gene ID. Rank: rank based on marginal p-values.*

9 Most Significant					LASSO				
ID	Rank	Estimate (×10)	SE (×10)	Z-score	ID	Rank	Estimate (×10)	SE (×10)	Z-score
14837	1	0.582	0.264	2.210	14837	1	0.058	0.105	0.553
19384	2	-2.087	1.572	-1.327	17879	12	0.056	0.086	0.647
4899	3	-1.096	0.514	-2.132	12822	16	0.319	0.188	1.698
14689	4	1.795	1.067	1.682	19274	73	-0.479	0.369	-1.297
15914	5	0.574	1.293	0.444	2059	88	-0.740	0.389	-1.899
15463	6	0.495	0.545	0.907	19307	91	0.417	0.227	1.833
21309	7	-0.484	0.636	-0.761	14140	95	0.271	0.141	1.922
2808	8	0.373	0.546	0.683	19282	102	-0.332	0.220	-1.510
515	9	0.536	0.181	2.969	14049	103	0.0211	0.087	0.244

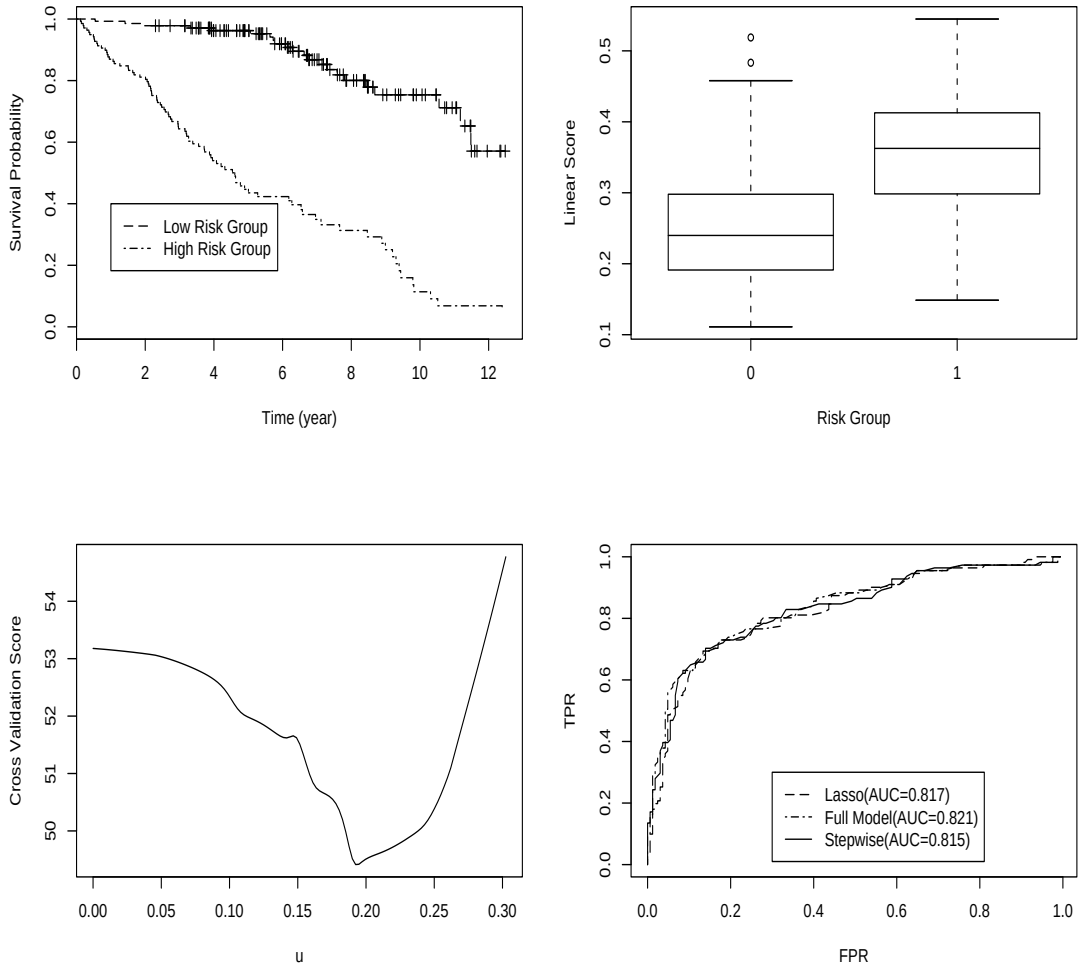

Figure 1: PBC data: model evaluation. Upper-left: survival functions (classified by the LASSO estimated risk scores) for two risk groups. Upper-right: linear risk scores for the groups with $\delta = 0$ and $\delta = 1$. Lower-right: cross validation score plot. Lower-right: ROC curves.

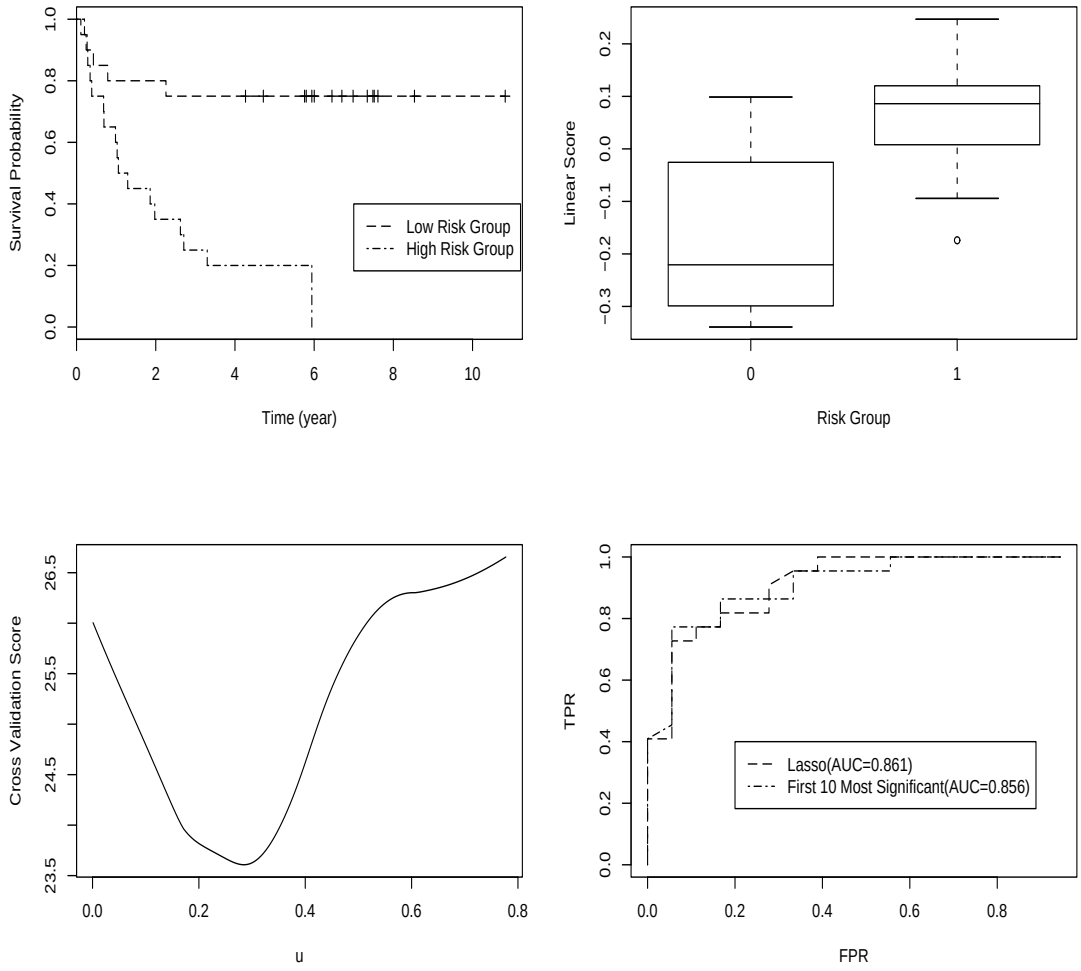

Figure 2: DLBCL data: model evaluation. Upper-left: survival functions (classified by the LASSO estimated risk scores) for two risk groups. Upper-right: linear scores for the groups with $\delta = 0$ and $\delta = 1$. Lower-right: cross validation score plot. Lower-right: ROC curves.