Published in final edited form as:
Stat Comput. 2010 April 1; 20(2): 165–176. doi:10.1007/s11222-009-9126-y.

Rank-based variable selection with censored data

Jinfeng Xu,
Department of Statistics and Applied Probability, Risk Management Institute, National University of Singapore, 117546 Singapore, Singapore

Chenlei Leng, and
Department of Statistics and Applied Probability, Risk Management Institute, National University of Singapore, 117546 Singapore, Singapore

Zhiliang Ying
Department of Statistics, Columbia University, New York, NY 10027, USA

Jinfeng Xu: staxj@nus.edu.sg; Chenlei Leng: stalc@nus.edu.sg; Zhiliang Ying: zying@stat.columbia.edu

Abstract

A rank-based variable selection procedure is developed for the semiparametric accelerated failure time model with censored observations where the penalized likelihood (partial likelihood) method is not directly applicable.

The new method penalizes the rank-based Gehan-type loss function with the ℓ_1 penalty. To correctly choose the tuning parameters, a novel likelihood-based χ^2-type criterion is proposed. Desirable properties of the estimator such as the oracle properties are established through the local quadratic expansion of the Gehan loss function.

In particular, our method can be easily implemented by the standard linear programming packages and hence numerically convenient. Extensions to marginal models for multivariate failure time are also considered. The performance of the new procedure is assessed through extensive simulation studies and illustrated with two real examples.

Keywords

Accelerated failure time model; Adaptive Lasso; BIC; Gehan-type loss function; Lasso; Variable selection

1 Introduction

As an attractive alternative to the proportional hazards model (Cox 1972), the accelerated failure time model specifies directly a regression model of the log transformed survival time on a set of covariates,

$$\log T_i = \beta_0^T X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where T_i is the survival time, X_i is a p-dimensional covariate, β_0 is a p-vector of unknown regression parameters and ε_i ($i = 1, \ldots, n$) are independent error terms with a common, but completely unspecified, distribution. Because of this direct relationship between survival

Correspondence to: Jinfeng Xu, staxj@nus.edu.sg.

time and covariates, the accelerated failure time model is physically interpretable and in many ways more appealing than the proportional hazards model (Kalbfleisch and Prentice 2002, Chap. 7).

A common phenomenon of survival analysis is that data are subject to right censoring. Due to the censoring, the observed data are $(\tilde{T}_i, \Delta_i, X_i, i = 1, \ldots, n)$, where C_i is the censoring time, $\tilde{T}_i = T_i \wedge C_i$ is the observed time and $\Delta_i = 1_{\{T_i \le C_i\}}$ is the censoring indicator. Define $e_i(\beta) = \log \tilde{T}_i - \beta^T X_i$, $N_i(\beta; t) = \Delta_i 1_{\{e_i(\beta) \le t\}}$ and $Y_i(\beta; t) = 1_{\{e_i(\beta) \ge t\}}$. Note that N_i and Y_i are the counting processes and at-risk process on the time scale of the residual. Write

$$S^{(0)}(\beta; t) = n^{-1} \sum_{i=1}^{n} Y_i(\beta; t),$$
$$S^{(1)}(\beta; t) = n^{-1} \sum_{i=1}^{n} Y_i(\beta; t) X_i.$$

The weighted log-rank estimating function for β_0 takes the form

$$U_\phi(\beta) = \sum_{i=1}^{n} \Delta_i \phi\{\beta; e_i(\beta)\}[X_i - \bar{X}\{\beta; e_i(\beta)\}], \quad \text{or}$$
$$U_\phi(\beta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(\beta; t)\{X_i - \bar{X}(\beta; t)\} dN_i(\beta; t),$$

where $\bar{X}(\beta; t) = S^{(1)}(\beta; t)/S^{(0)}(\beta; t)$, and ϕ is a possibly data-dependent weight function, the choice of $\phi = S^{(0)}$ leads to the Gehan statistics. In this case, U_ϕ can be written as

$$U_G(\beta) = \sum_{i=1}^{n} \Delta_i S^{(0)}\{\beta; e_i(\beta)\}[X_i - \bar{X}\{\beta; e_i(\beta)\}],$$

or

$$U_G(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (X_i - X_j) 1_{\{e_i(\beta) \le e_j(\beta)\}},$$

which is the gradient of the convex function

$$L_G(\beta) := n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i \{e_i(\beta) - e_j(\beta)\}^-$$

for $|a|^- = |a| 1_{a<0}$. Based on this fact, Jin et al. (2003) provided simple and reliable methods to obtain the rank estimators. They showed that the rank estimator with the Gehantype (1965) weight function can be readily obtained by minimizing a convex objective function through a standard linear programming technique.

An important objective of survival analysis is to identify a subset of significant variables from a large number of covariates which are often collected to reduce possible modeling bias. Variable selection is fundamental to statistical modeling and recently, a number of approaches based on the penalized partial likelihood have been applied to the Cox proportional hazards model and gained increasing popularity. See, for example, Lasso

(Tibshirani, 1996, 1997), Scad (Fan and Li, 2001, 2002), Alasso (Zou, 2006, 2008; Zhang and Lu 2007; Lu and Zhang 2007; Wang et al. 2007b), and LSA (Wang and Leng 2007). Wang et al. (2007a) discussed the penalized least absolute deviation estimation without considering censoring. In the accelerated failure time model aforementioned, we are particularly interested in estimating the unknown vector β_0 and identifying any nonzero components. Zhang and Lu (2007) proposed the penalized partial likelihood method for variable selection in the Cox model. However, in the accelerated failure time model, partial likelihood is not available and semiparametric estimation of the regression coefficient vector relies on the rank-based methods. The estimates are usually obtained by minimizing a non-smooth objective function or solving the estimating equations which are step functions with potentially multiple roots (Jin et al. 2003). Johnson (2008) and Johnson et al. (2008) proposed a general variable selection procedure by penalizing estimation equations with broad and important applications especially including the censored accelerated failure time model. For uncensored data, Johnson and Peng (2008) developed a rank-based variable selection procedure in the linear model and the desirable properties such as the robustness and the oracle properties are obtained. In this article, we propose the ℓ_1 regularized Gehan estimator for simultaneous estimation and variable selection which yields advantages in two fronts. First, the shrinkage property of the ℓ_1 penalty and proper choice of tuning parameters build sparse models without sacrificing accuracy. Secondly, the single criterion function with both components being of ℓ_1-type reduces (numerically) the minimization to a strictly linear programming problem, making any resulting methodology extremely easy to implement.

The rest of the paper is organized as follows. Section 2 introduces the ℓ_1 regularized Gehan estimator and gives its asymptotic properties. A novel χ^2 type criterion is proposed to choose the tuning parameters in Sect. 3. Extensions to multivariate failure time data are considered in Sect. 4. The proposed methodology is illustrated with the applications to two datasets in Sect. 5. In Sect. 6, simulations are conducted to assess the finite-sample performance of the proposed methods. Section 7 concludes with a general discussion. All the proofs are relegated to the Appendix.

2 The ℓ_1 regularized Gehan estimator

Define the ℓ_1 regularized Gehan loss function as

$$\begin{aligned} L_P(\beta) &:= n^{-1} L_G(\beta) + \sum_{j=1}^{p} \lambda_{nj} |\beta_j| \\ &= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i \{e_i(\beta) - e_j(\beta)\}^- + \sum_{j=1}^{p} \lambda_{nj} |\beta_j|, \end{aligned} \tag{1}$$

where λ_{nj}, $j = 1, \ldots, p$ are tuning parameters and β_j is the j th component of the vector β, $j = 1, \ldots, p$. The regularized estimate of β_0 is a minimizer of $L_P(\beta)$ and denoted by $\hat{\beta}$.

Let $\beta_0 = \{(\beta_0^1)^T, (\beta_0^2)^T\}^T$, where β_0^1 is an s-vector and β_0^2 is a $(p - s)$-vector. Without loss of generality, assume β_0^2 is the zero vector and β_0^1 is the nonzero vector. Suppose further that $\{\lambda_{nj}\}$ satisfy the following conditions:

$$\text{(C)} \quad \begin{aligned} &\sqrt{n}\lambda_{nj} \to 0 \quad \text{for } 1 \le j \le s \text{ and} \\ &\sqrt{n}\lambda_{nj} \to \infty \quad \text{for } s+1 \le j \le p. \end{aligned}$$

This assumption states that the penalties applied on the zero entries in β dominate those on the nonzero entries. Intuitively, the requirement that $\sqrt{n}\lambda_{nj} \to 0$ for $1 \le j \le s$ enables the resulting estimates of β_0^1 to be $\sqrt{n}$-consistent; and the condition that $\sqrt{n}\lambda_{nj} \to \infty$ for $s + 1 \le j \le p$ shrinks the estimates of β_0^2 to zero. This observation is made rigorous by the following theorem.

Theorem 1

(Oracle properties) *Under the conditions* 1–4 *of* Ying (1993) *and* (C), *with probability tending to one, the penalized estimator* $\hat{\beta} = \{(\hat{\beta}^1)^T, (\hat{\beta}^2)^T\}^T$ *has the following properties*:

- **a.** $\hat{\beta}^2 = 0$;
- **b.** $n^{\frac{1}{2}}(\hat{\beta}^1 - \beta_0^1) \to N(0, V)$, *where*, $V = A_{G1}^{-1} B_{G1} A_{G1}^{-1}$, A_{G1} *and* B_{G1} *are defined in the* Appendix.

Remark 1—Theorem 1 states that the proposed estimator estimates the coefficients of the important variables as if the true were known in advance, which is referred to as the oracle properties by Fan and Li (2001). However, as noted by an anonymous referee, the model selection consistency of the adaptive Lasso is obtained through the large sample theory and for finite sample size and small signal-to-noise ratio, it can perform rather poorly, sometimes even worse than the ordinary Lasso. Furthermore, Leeb and Pötscher (2008) showed that the maximal risk of any sparse estimator goes to infinity as the sample size increases. Hence the minimax efficiency and the model selection consistency seem to be two irreconcilable properties. In practice, we should always be cautious about using which to choose a good estimator.

The limiting covariance matrix V involves the unknown hazard function, so that it is difficult to estimate the covariance matrix analytically. Here we apply the random perturbation method (Rao and Zhao 1992; Parzen et al. 1994; Jin et al. 2001) to approximate the distribution of $\hat{\beta}$. To be specific, define $\hat{\beta}^*$ as a minimizer of the perturbed ℓ_1 regularized Gehan loss function

$$L_p^*(\beta) := n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i \{e_i(\beta) - e_j(\beta)\}^- Z_i + \sum_{i=1}^{p} \lambda_{nj} |\beta_j|, \quad (2)$$

where the random variable Z_i satisfied $E(Z_i) = 1$ and $V(Z_i) = 1$. In the data analysis and simulation studies, the standard exponential distribution is used. The following theorem justifies the use of random perturbation method to distributional approximation of the estimator.

Theorem 2

(Asymptotic variance) *Under* (C) *and conditions* 1–4 *of Ying* (1993, *p*. 80), *with probability tending to one, conditional on the data* $(\hat{T}_i, \Delta_i, X_i)$ $(i = 1, \ldots, n)$, $\hat{\beta}^* = \{(\hat{\beta}_1^*)^T, (\hat{\beta}_2^*)^T\}$ *has the following properties*: (a) $\hat{\beta}_2^* = 0$; (b) $n^{\frac{1}{2}}(\hat{\beta}_1^* - \hat{\beta}^1) \to N(0, V)$.

In practice, we fix the tuning parameters which satisfy condition (C) in implementing the perturbation method.

3 Computation and tuning parameter selection

As pointed out in Jin et al. (2003), the minimization of $L_P(\beta)$ can be carried out by linear programming and is equivalent to the minimization of

$$n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i|e_i(\beta)-e_j(\beta)| + \left|M-\beta^T\sum_{k=1}^{n}\sum_{l=1}^{n}\Delta_k(X_l-X_k)\right| + \sum_{j=1}^{p}\lambda_{nj}|\beta_j|,$$

where M is a large constant. An implementation of the algorithm is discussed by Koenker and D'Orey (1987), with the code available in S-Plus and R and other softwares. The minimization of $L_p^*(\beta)$ can be implemented similarly. Condition (C) suggests pre-selection of $\{\lambda_{nj}\}$'s based on a preliminary estimate of β, and in this work we take

$$\lambda_{nj}=\lambda|\tilde{\beta}_j|^{-\tau}, \quad j=1,\ldots,p, \quad (3)$$

for $\tau>0$, as $\{\tilde{\beta}_j\}$ are $\sqrt{n}$-consistent. This choice is discussed by Zou (2006), Zhang and Lu (2007) and Wang et al. (2007b). For such λ_{nj} s, we have $a_n=\max\{\lambda_{nj}, j=1,\ldots, s\}=\lambda O_p(1)$ and $b_n=\min\{\lambda_{nj}, j=s+1, \ldots, p\}=\lambda O_p(n^{\tau/2})\to\infty$. It is then easy to see that once λ satisfies

$$\sqrt{n}a_n\to 0 \quad \text{and} \quad \sqrt{n}b_n\to\infty,$$

Theorems 1 and 2 hold. This simplification is attractive from a computational viewpoint, since we only need to choose one tuning parameter λ instead of p tuning parameters $\lambda_{nj}, j=1, \ldots, p$. For later exposition, we shall fix $\{\lambda_{nj}\}$ according to (3) with $\tau=1$.

In order to study the dependence of the selected model on the tuning parameter λ, we denote the model corresponding to $\hat{\beta}_\lambda$ as $S_\lambda=\{j:\hat{\beta}_{\lambda,j}\neq 0\}$. Write the derivative of Gehan loss function as

$$U_G(\beta)=n^{-1}\sum_{i=1}^{n}\Delta_i S^{(0)}\{\beta;e_i(\beta)\}\,[X_i-\overline{X}\{\beta;e_i(\beta)\}].$$

Then $\sqrt{n}U_G(\beta_0)$ is asymptotically normal with mean zero and variance

$$B_{Gn}=n^{-1}\sum_{i=1}^{n}\Delta_i S^{(0)}\{\beta;e_i(\beta)\}^2[X_i-\overline{X}\{\beta;e_i(\beta)\}]^{\otimes 2}.$$

Consider the following χ^2 type statistics:

$$T_\lambda=nU_G(\widehat{\beta}_\lambda)^T B_{Gn}^{-1}(\widehat{\beta}_G)U_G(\widehat{\beta}_\lambda),$$

and by applying arguments similar to those in Wei et al. (1990), when $S_\lambda\supseteq\{1,\ldots,s\}$ i.e. a correct model is identified (not necessarily the true model), T_λ follows the χ^2 distribution with degrees of freedom q which equals to the number of zero components in $\hat{\beta}_\lambda$. T_λ is

scale-invariant and has a likelihood interpretation and hence can be used for tuning parameter selection. An attractive property of T_λ is that B_{Gn} does not require density estimation and thus can be easily computed. Based on T_λ, we propose the Bayesian information criterion (BIC)

$$\text{BIC}_\lambda = T_\lambda + \log n \cdot df_\lambda,$$

where df_λ is the number of nonzero components in $\hat{\beta}_\lambda$. Similarly, the Akaike information criterion (AIC) can be defined as

$$\text{AIC}_\lambda = T_\lambda + 2 \cdot df_\lambda.$$

Replacing the ℓ_2 loss function in the generalized cross validation (GCV) with the Gehan loss function, GCV can be defined as

$$\text{GCV}_\lambda = \frac{L_G(\widehat{\beta}_\lambda)}{(1 - df_\lambda/n)^2}.$$

We define $R_0 = \{\lambda \geq 0 : S_\lambda = \{1, \ldots, s\}\}$ as the set of λ's such that the true model is identified. In addition, we define a reference tuning parameter sequence $\{\lambda_n = 1/[\sqrt{n}\log(n)]\}_{n=1}^{\infty}$. By Theorem 1, it follows that with probability one, $S_{\lambda_n} = \{1, \ldots, s\}$. We have the following consistency theorem for the BIC method.

Theorem 3

Under the assumptions in Theorem 1, $P(\inf_{\lambda \notin R_0} \text{BIC}_\lambda > \text{BIC}_{\lambda_n}) \to 1$.

This theorem states that for any λ which can not choose the true model, its associated BIC is larger than the one identified by the reference sequence. Therefore, the optimal λ which minimizes BIC must correspond to the true model. The proof of this theorem is similar to that of Theorem 4 in Wang and Leng (2007) and is therefore omitted. Note that neither AIC nor GCV yields consistent model selection results if a true sparse model exists (Wang et al. 2007c).

4 Extensions to multivariate failure time data

Following Jin et al. (2006), in this section, we consider the extension of ℓ_1 regularized method to multivariate failure time data. The oracle properties for the estimators defined in this section, similar to those in Theorems 1–3, can be shown by using the similar techniques and are therefore omitted. Additionally, computing the estimates relies on the linear programming technique, thus can be easily implemented.

4.1 Multiple events data

Multiple events data arise when a subject can potentially experience several types of event or failure. For $k = 1, \ldots, K$ and $i = 1, \ldots, n$, let T_{ki} be the time to the kth failure of the ith subject, let C_{ki} be the censoring time on T_{ki}, and let X_{ki} be the corresponding p_k-vector of covariates. We assume that $(T_{1i}, \ldots, T_{Ki})$ is independent of $(C_{1i}, \ldots, C_{Ki})$ conditional on $(X_{1i}, \ldots, X_{Ki})$. The marginal accelerated failure time models take the form

$$\log T_{ki}=X_{ki}^T\beta_k+\varepsilon_{ki} \quad (k=1,\ldots,K;\ i=1,\ldots,n),$$

where β_k is a p_k-vector of unknown regression parameters, and $(\varepsilon_{1i}, \ldots, \varepsilon_{Ki})$, $i = 1, \ldots, n$ are independent random vectors from an unspecified joint distribution with marginal distribution functions $F_1, \ldots, F_K$. The data consists of $(\tilde{T}_{ki}, \delta_{ki}, X_{ki})$, $k = 1, \ldots K$; $i = 1, \ldots, n$, where $\tilde{T}_{ki} = \min(T_{ki}, C_{ki})$ and $\delta_{ki} = I_{\{T_{ki} \le C_{ki}\}}$.

Let $e_{ki}(\beta) = \log \tilde{T}_{ki} - \beta^T X_{ki}$ and the Gehan-type loss function for β_k is then

$$L_{k,G}(\beta)=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ki}\{e_{ki}(\beta)-e_{kj}(\beta)\}^-.$$

For each $k = 1, \ldots, K$, correspondingly, the ℓ_1 regularized loss function is

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ki}\{e_{ki}(\beta)-e_{kj}(\beta)\}^-+\lambda\sum_{j=1}^{p_k}\lambda_j|\beta_{jk}|.$$

Similarly, the minimization problem can be reduced to K standard linear programming problems and the distribution of the estimator can be approximated by the perturbation method.

4.2 Clustered failure time data

The clustered data arise when we have a random sample of n clusters and there are K_i members in the ith cluster. Let T_{ik} and C_{ik} be the failure time and censoring time for the kth member of the ith cluster, and let X_{ik} be the corresponding $p \times 1$ vector of covariates. We assume that $(T_{i1}, \ldots, T_{i K_i})$ and $(C_{i1}, \ldots, C_{i K_i})$ are independent conditional on $(X_{i1}, \ldots, X_{i K_i})$. The data consist of $(\tilde{T}_{ik}, \delta_{ik}, X_{ik})$, $k = 1, \ldots, K_i$; $i = 1, \ldots, n$, where $\tilde{T}_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = I_{\{T_{ik} \le C_{ik}\}}$.

Suppose that the marginal distribution of the T_{ik} satisfy the accelerated failure time model

$$\log T_{ik}=X_{ik}^T\beta_0+\varepsilon_{ik} \quad (k=1,\ldots,K_i;\ i=1,\ldots,n),$$

where β_0 is a p-vector and $(\varepsilon_{i1}, \ldots, \varepsilon_{i K_i})$, $i = 1, \ldots, n$ are independent random vectors. Define, $e_{ik}(\beta)=\log \tilde{T}_{ik} - X_{ik}^T\beta$, the Gehan type loss function for β is

$$L_G(\beta)=n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{j=1}^{n}\sum_{\ell=1}^{K_j}\delta_{ik}\{e_{ik}(\beta)-e_{j\ell}(\beta)\}^-.$$

The ℓ_1 regularized loss function is

$$n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{j=1}^{n}\sum_{\ell=1}^{K_j}\delta_{ik}\{e_{ik}(\beta)-e_{j\ell}(\beta)\}^-+\lambda\sum_{j=1}^{p}\lambda_j|\beta_j|.$$

4.3 Recurrent events data

With a random sample of n subjects, let T_{ki} be the time to the kth recurrent event on the ith subject; let C_i and X_i be the censoring time and the $p \times 1$ vector of covariates for the ith subject. Assume that C_i is independent of T_{ki} $(k = 1, \ldots)$ conditional on X_i. Let

$$N_i^*(t) = \sum_{k=1}^{\infty} I(T_{ki} \le t).$$

We specify the following accelerated time model for the mean frequency function:

$$E\{N_i^*(t)|X_i\} = \mu_0(te^{-\beta_0^T X_i}),$$

where β_0 is a $p \times 1$ vector and $\mu_0(\cdot)$ is an unspecified baseline mean function. The Gehan type loss function for β is

$$L_G(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{\infty} I(T_{ki} \le C_i) \times \{\log T_{ki} - \log C_j - \beta^T (X_i - X_j)\}^-.$$

The ℓ_1 regularized Gehan loss function is

$$L_G(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{\infty} I(T_{ki} \le C_i) \{\log T_{ki} - \log C_j - \beta^T (X_i - X_j)\}^- + \lambda \sum_{j=1}^{p} \lambda_j |\beta_j|.$$

5 Two real examples

In this section, we apply our method to analyze two well known datasets.

5.1 Primary biliary cirrhosis data

The primary biliary cirrhosis (PBC) data, provided in Therneau and Grambsch (2001), came from the Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between 1974 and 1984. The data contain information about the survival time and 17 prognostic variables for 424 PBC patients who met eligibility criteria for the randomized placebo-controlled trial of the drug D-penicillamine. We considered the 276 patients with the complete information of all 17 variables and used the accelerated failure time model to study the relationship of the survival time and the prognostic variables. In our analysis, the 17 variables are drug, age, sex, ascites, hepatomegaly, spiders, edema, bilirubin, cholesterol, albumin, urine copper, alkaline phosphotase, SGOT, triglycerides, platelets, prothrombin time and histologic stage of disease. Albumin, alkaline phosphatase, bilirubin and prothrombin time have all been transformed on the natural logarithmic scale. The variables were also standardized to have mean zero and unit variance. The R package quantreg was used in both data analysis and simulation studies. Table 1 summarizes the estimated coefficients of the Gehan estimate, the Lasso estimate and the adaptive Lasso estimate. The AIC, BIC or GCV criteria were used to choose the tuning parameters respectively. For this data set, GCV and AIC yield the same estimates for the same penalty function (Lasso or adaptive Lasso) and thus only the estimated coefficients via AIC were reported. The standard errors were computed via 1000

random perturbations for the standard exponential distribution. However, we only recorded the standard errors of those which were identified as nonzero coefficients. To appreciate the relationship between various estimates and the tuning parameters, for a shrinkage estimator, we calculated the shrinkage parameter as

$$s=\frac{\sum_{j=1}^{p}|\widehat{\beta}|}{\sum_{j=1}^{p}|\widehat{\beta^G}|},$$

where $\hat{\beta}^G$ is the unpenalized Gehan estimator, $\hat{\beta}$ is either the Lasso ($\hat{\beta}^L$) or the adaptive Lasso ($\hat{\beta}^A$) estimator.

In Figs. 1 and 2, both the coefficients and the criterion function (AIC/BIC/GCV) are plotted against the shrinkage parameter for Lasso and adaptive Lasso estimates respectively.

A few observations can be made from Table 1. First, GCV and AIC perform similarly for Lasso and adaptive Lasso penalties. Secondly, BIC yields smaller models than AIC and GCV. The combination of adaptive Lasso and BIC gives the model with the most zero coefficients. From Figs. 1 and 2, we see that BIC tends to shrink more than AIC and GCV and hence gives sparser models.

As noted by an anonymous referee, BIC usually has less favorable out-of-sample prediction performance than the cross-validation method. Here we use a tenfold crossvalidation scheme to compare the out-of-sample performance of AIC, BIC, GCV and cross-validation. To be specific, we left out one tenth of the data and use the remaining data to obtain the Lasso and adaptive Lasso estimates while choosing the tuning parameters via AIC, BIC, GCV or a tenfold cross-validation. Then we estimate the prediction error by evaluating the Gehan loss function on the one-tenth leftout sample. Table 2 shows the results for the PBC data. It indicates BIC performs a bit worse than the other criteria but the sacrificed prediction accuracy is not much.

5.2 Framingham heart data

The Framingham Heart Study (Dawber 1980) was initiated in 1948, with 2336 men and 2873 women aged between 30 and 62 years at their baseline examination. Individuals were examined every 2 years after participating in the 30-year follow-up study. Multiple events, e.g., times to coronary heart disease (CHD), denoted by T_1; and cerebrovascular accident (CVA), T_2, were observed from the same individual. Data used here included the participants in the study who had an examination at age 44 or 45 and were disease-free at that examination in the sense that there existed no history of hyper-tension or glucose intolerance and no previous experiences of a CHD or CVA. The original dataset contains a total of 1571 disease-free individuals. The risk factors of interest were age, x_1; body mass index, x_2; cholesterol level, x_3; systolic blood pressure, x_4; cigarette smoking, x_5; and gender, x_6. Since modeling biases can possibly be reduced by introducing interactions, we consider the marginal bivariate accelerated failure time model using both the main effects and first-order interactions. The variables were standardized to have zero mean and unit variance. Table 3 summarizes the estimated coefficients of the Gehan estimate, the Lasso estimate and the adaptive Lasso estimate for both CHD and CVA. The BIC criterion was used to choose the tuning parameters. It shows that there are many interactions among the risk factors on both CHD and CVA and as expected, adaptive Lasso yields sparser models than Lasso.

6 Simulations

We conducted extensive simulation study for univariate and multivariate failure time data in this session. All the simulations were conducted using R package quantreg.

6.1 Univariate failure time data

We simulated datasets consisting of 100 observations from the accelerated failure time model

$$\log T_i = \beta^T X_i + \varepsilon_i,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, the x_i s were marginally standard normal and the correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0.5$. The censoring times were generated from $Un(0, \tau)$ distribution, where τ was chosen to be 142 or 50 yielding the censoring level 30% or 50% respectively. The distribution of ε was set to be $N(0, 1)$, t_3 and $0.5\ N(0, 1) + 0.5\ N(0, 9)$ respectively to assess the robustness of our proposed method. For the estimator $\hat{\beta}$, its performance is gauged by the model error which is defined as

$$E(\widehat{\beta} - \beta_0)^T X^T X(\widehat{\beta} - \beta_0).$$

The ideal oracle estimator which knows the true nonzero co-efficients but not their exact values will apply the rank-based estimation procedure by considering only covariates x_1, x_2 and x_5. The Lasso and adaptive Lasso were used to penalize the Gehan-loss function and the tuning parameters were chosen by AIC, BIC or GCV as defined in Sect. 3. For each method, its relative model error compared to that of the oracle estimator was computed.

In Table 4, the median relative model errors (MRME) based on 1000 simulated datasets as well the average number of correctly selected (C) and incorrectly selected (IC) variables are summarized for the three error distributions. It shows that adaptive Lasso with BIC yields estimates with smaller models and more accurate estimates. Furthermore, the adaptive Lasso outperforms the Lasso in terms of variable selection and mean squares errors and the proposed estimator is very robust to the heavy tailed distribution t_3 and contamination normal distribution $0.5\ N(0, 1) + 0.5\ N(0, 9)$.

To investigate the performance of the adaptive Lasso and the Lasso when the signal-to-noise ratio is small, following Leeb and Pötscher (2008), we multiply the regression coefficient vector by $1/\sqrt{n}$. With sample size 100 and censoring 30%, the results are summarized in Table 5. It can be seen that the adaptive Lasso performs very poorly and even a bit worse than the Lasso.

6.2 Multivariate failure time data

For multiple events and clustered data, two failure times T_1 and T_2 were generated from Gumbel (1960) bivariate distribution:

$$F(t_1, t_2) = F_1(t_1) F_2(t_2) [1 + \theta\{1 - F_1(t_1)\}\{1 - F_2(t_2)\}],$$

where $-1 \le \theta \le 1$. The correlation between T_1 and T_2 is $\theta/4$. The two marginal distributions $F_k(t_k)$, $k = 1, 2$, were generated from the exponential distribution with hazard rate $\lambda_k = e^{\beta_k^T X_k}$. We simulated 100 datasets consisting of 100 observations from the model where $X_k = (x_{1k},$

…, x_{pk}), $p = 8$, the x_{ik} s were marginally standard normal and the correlation between x_{ik} and x_{jk} was $\rho^{|i-j|}$ with $\rho = 0.5$. The censoring times were generated from $Un(0, \tau)$ distribution, where τ was chosen to be 1.5 yielding the censoring level 50%. For multiple events, we set $\beta_{10} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\beta_{20} = (0, 0, 2, 0, 0, 1.5, 3, 0)^T$ and $X_k = X$. For clustered data, X_1 and X_2 were generated inependently and $\beta_{k0} = \beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. For recurrent events, we set $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the covariates were generated in the same manner as in the case of multiple events. The gap times between successive events were generated from the aforementioned Gumbel's bivariate exponential distribution. The resultant recurrent event process is Poisson under $\theta = 0$ and non-Poisson under $\theta \neq 0$. The follow-up time was an independent $Un(0, 2.5)$ random variable, which on average yielded approximately 2.60 and 2.86 events per subject for the Poisson and non-Poisson cases respectively.

For multiple events, the performance of the estimator is gauged by the model error which is defined as

$$\sum_{k=1}^{2} E\,(\widehat{\beta}_k - \beta_{0k})^T\, X^T X\,(\widehat{\beta}_k - \beta_{0k}).$$

The ideal oracle estimator which knows the true nonzero coefficients but not their exact values will apply the rank-based estimation procedure by considering only covariates x_1, x_2 and x_5 for T_1 and x_3, x_6 and x_7. For clustered data, the performance of the estimator is gauged by the model error

$$\sum_{k=1}^{2} E\,(\widehat{\beta} - \beta_0)^T\, X_k^T X_k\,(\widehat{\beta} - \beta_0).$$

For recurrent events, the performance of the estimator is gauged by the model error

$$E\,(\widehat{\beta} - \beta_0)^T\, X^T X\,(\widehat{\beta} - \beta_0).$$

For clustered data and recurrent events, the ideal oracle estimator just considers variables x_1, x_2 and x_5.

The Lasso and adaptive Lasso were used to penalize the Gehan-loss function and the tuning parameters were chosen by AIC, BIC or GCV as defined in Sect. 4. For each method, its relative model error compared to that of the oracle estimator was computed.

In Table 6, the median relative model errors (MRME) based on 100 simulated datasets as well the average number of correctly selected (C) and incorrectly selected (IC) variables are summarized for multiple events, clustered and recurrent data respectively. We can see again that adaptive Lasso with BIC performs the best.

7 Discussion

We propose in this paper an ℓ_1 regularized procedure for variable selection in the accelerated failure time model. The resulting estimates possess the oracle properties if the adaptive Lasso penalty is used for penalization and BIC is used for tuning parameter selection. Additionally, we extend the ℓ_1 regularized procedure to multivariate failure time models, including multiple events data, clustered survival data and recurrent events data. Extensive

simulation study and a real data analysis of primary biliary cirrhosis data illustrate the usefulness of our approach in terms of both variable selection and coefficient estimation. Although we only considered the Gehan statistics based loss function, it is rather straightforward to extend our approach to other weighting schemes discussed in Jin et al. (2003). Our current implementation via quantreg uses a grid of tuning parameter values and alternatively, we could implement the path following algorithm as detailed in Li and Zhu (2008), which, as noted by an anonymous referee, has been recently investigated in Cai et al. (2009). We modeled multivariate survival data via the marginal approach. It would be interesting to investigate how to conduct variable selection while accounting for correlations among multiple failure times.

Acknowledgments

This research was supported by grants from the U.S. National Institutes of Health, the U.S. National Science Foundation and the National University of Singapore (R-155-000-075-112 and R-155-000-080-112). We are very grateful to the editor, the associate editor and the referees for their helpful comments which have greatly improved the paper.

Appendix

Proofs of Theorems 1 and 2

Proof of Theorem 1 proceeds by first establishing the local quadratic property (Proposition 1) of the Gehan loss function and an inequality (Proposition 2) which relates the minimizer of the penalized loss function to the minimizer of a penalized quadratic function. Then the $\sqrt{n}$ consistency of the penalized estimator and the oracle properties follow by applying the arguments similar to those of Fan and Li (2001).

Define

$$A_G = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{\infty} S^{(0)}(\beta_0;t) \times\{X_i - \overline{X}(\beta_0;t)\}^{\otimes 2}\{\dot{\lambda}(t)/\lambda(t)\}dN_i(\beta_0;t),$$

$$B_G = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{\infty}[S^{(0)}(\beta_0;t)]^2 \times\{X_i - \overline{X}(\beta_0;t)\}^{\otimes 2} dN_i(\beta_0;t).$$

Proposition 1

Under conditions 1–4 *of Ying* (1993, *p.* 80), *for every sequence* $d_n > 0$ *with* $d_n \to 0$ *a.s., we have*

$$\begin{aligned} &n^{-1}L_G(\beta) - n^{-1}L_G(\beta_0) \\ =&n^{-1}U_G^T(\beta_0)(\beta-\beta_0) + (\beta-\beta_0)^T A_G(\beta-\beta_0)/2 \\ &+ o(\|\beta-\beta_0\|^2 + n^{-1}) \end{aligned} \quad (4)$$

holds uniformly in $\|\beta - \beta_0\| \le d_n$.

Proof—It follows from Theorem 2 of Ying (1993) that almost surely, uniformly in $\|\beta - \beta_0\| \le d_n$, we have

$$U_G(\beta) = U_G(\beta_0) + nA_G(\beta - \beta_0) + o(n^{1/2} + n\|\beta - \beta_0\|) \quad (5)$$

and denote $U_G = (U_{G1}, \ldots, U_{Gp})^T$, $A_G = (a_{ij})$, $1 \le i, j \le p$, $\beta = (\beta_1, \ldots, \beta_p)^T$, $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T$, and $\beta^j = (\beta_1, \ldots, \beta_{p-j}, \beta_{0(p-j+1)}, \ldots, \beta_{0p})^T$, $1 \le j \le p$, $\beta^0 = \beta$, noticing that $\beta^p = \beta_0$, we have

$$\begin{aligned} & L_G(\beta) - L_G(\beta_0) \\ & = L_G(\beta^0) - L_G(\beta^p) \\ & = \sum_{k=1}^{p} [L_G(\beta^{k-1}) - L_G(\beta^k)], \\ & = \sum_{k=1}^{p} \int_{\beta_{0(p-k+1)}}^{\beta_{p-k+1}} U_{G(p-k+1)}(\beta_1, \ldots, \beta_{p-k}, s_{p-k+1}, \\ & \beta_{0(p-jk+2)}, \ldots, \beta_{0p}) ds_{p-k+1}. \end{aligned}$$

By (5),

$$\begin{aligned} & U_{G(p-k+1)}(\beta_1, \ldots, \beta_{p-k}, s_{p-k+1}, \beta_{0(p-jk+2)}, \ldots, \beta_{0p}) \\ & = U_{G(p-k+1)}(\beta_0) + n \sum_{l=1}^{p-k} a_{(p-k+1)l}(\beta_l - \beta_{0l}) \\ & + n a_{(p-k+1)(p-k+1)}(s_{p-k+1} - \beta_{0(p-k+1)}) \\ & + o(n^{1/2} + n\|\beta - \beta_0\|) \end{aligned}$$

then

$$\begin{aligned} & L_G(\beta) - L_G(\beta_0) \\ & = \sum_{k=1}^{p} U_{G(p-k+1)}(\beta_0)(\beta_{p-k+1} - \beta_{0(p-k+1)}) \\ & + \frac{n}{2}(\beta_{p-k+1} - \beta_{0(p-k+1)})^2 \\ & + n \sum_{k=1}^{p} \sum_{l=1}^{p-k} a_{(p-k+1)l}(\beta_l - \beta_{0l})(\beta_{p-k+1} - \beta_{0(p-k+1)}) \\ & + o(n^{1/2}\|\beta - \beta_0\| + n\|\beta - \beta_0\|^2) \\ & = U_G^T(\beta_0)(\beta - \beta_0) + n(\beta - \beta_0)^T A_G(\beta - \beta_0)/2 \\ & + o(n^{1/2}\|\beta - \beta_0\| + n\|\beta - \beta_0\|^2). \end{aligned}$$

Hence

$$\begin{aligned} & n^{-1}L_G(\beta) - n^{-1}L_G(\beta_0) \\ & = n^{-1}U_G^T(\beta_0)(\beta - \beta_0) + (\beta - \beta_0)^T A_G(\beta - \beta_0)/2 \\ & + o(n^{-1/2}\|\beta - \beta_0\| + \|\beta - \beta_0\|^2) \\ & = n^{-1}U_G^T(\beta_0)(\beta - \beta_0) + (\beta - \beta_0)^T A_G(\beta - \beta_0)/2 \\ & + o(\|\beta - \beta_0\|^2 + n^{-1}). \end{aligned}$$

This completes the proof.

Consider the object function:

$$C(u)=u^T Du/2 - a^T u+\sum_{j=1}^{s} \lambda_j u_j+ \sum_{j=s+1}^{p} \lambda_j|u_j|$$

where $u \in R^p$, D is a positive definite matrix, $\lambda_1, \ldots, \lambda_s$ are constants, $\lambda_{s+1},\ldots, \lambda_p$ are nonnegative constants, and suppose that $\hat{u}$ is a minimizer of $c(u)$, we have the following proposition.

Proposition 2

For any u, we have $C(u) - C(\hat{u}) \geq (u - \hat{u})^T D(u - \hat{u})/2$.

Suppose that $\hat{u}_n$ is a minimizer of the objective function

$$\begin{aligned} B_n(u) \quad = \quad & n^{-1/2} U_G^T(\beta_0)u+u^T A_G\, u/2 \\ & +\sum_{i=1}^{s} \sqrt{n}\lambda_{ni}\,\mathrm{sgn}(\beta_{0i})\, u_i+ \sum_{i=s+1}^{p} \sqrt{n}\lambda_{ni}|u_i|. \end{aligned} \quad (6)$$

By Propositions 1 and 2, it can be shown that $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)$ and $\hat{u}_n$ have the same asymptotic distribution. Then it is straightforward to obtain the $\sqrt{n}$ consistency of the penalized estimator and the oracle properties by following the same arguments in Fan and Li (2001).

Proof of Theorem 2 can be established similarly by looking at the perturbed penalized loss function and applying the conditional arguments as in Jin et al. (2003) and is thus omitted.

References

Cai T, Huang J, Lu T. Regularized estimation for the accelerated failure time model. Biometrics. 2009 to appear.

Cox DR. Regression models and life-tables (with Discussion). J R Stat Soc B. 1972; 34:187–220.

Dawber, TR. The Epidemiology of Atherosclerotic Disease. Harvard University Press; Cambridge: 1980. The Framingham Study.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001; 96:1348–1360.

Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. Ann Stat. 2002; 30:74–99.

Gehan EA. A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika. 1965; 52:203–223. [PubMed: 14341275]

Gumbel EJ. Bivariate exponential distributions. J Am Stat Assoc. 1960; 55:698–707.

Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. Biometrika. 2001; 88:381–390.

Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. Biometrika. 2003; 90:341–353.

Jin Z, Lin DY, Ying Z. Rank regression analysis of multivariate failure time data based on marginal linear models. Scand J Stat. 2006; 33:1–23.

Johnson BA. Variable selection in semiparametric linear regression with censored data. J R Stat Soc Ser B. 2008; 70:351–370.

Johnson BA, Peng LM. Rank-based variable selection. J Nonparametric Stat. 2008; 20:241–252.

Johnson BA, Lin DY, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. J Am Stat Assoc. 2008; 103:672–680. [PubMed: 20376193]

Kalbfleisch, J.; Prentice, R. The Statistical Analysis of Failure Time Data. 2. Wiley; New York: 2002.

Koenker R, D'Orey V. Computing regression quantiles. Appl Stat. 1987; 36:383–393.

Leeb H, Pötscher BM. Sparse estimators and the oracle property, or the return of Hodges' estimator. J Econom. 2008; 142:201–211.

Li Y, Zhu J. L1-norm quantile regression. J Comput Graph Stat. 2008; 17:163–185.

Lu W, Zhang HH. Variable selection for proportional odds model. Stat Med. 2007; 26:3771–3781. [PubMed: 17266170]

Parzen MI, Wei LJ, Ying Z. A resampling method based on pivotal estimating functions. Biometrika. 1994; 81:341–350.

Rao CR, Zhao LC. Approximation to the distribution of *M*-estimates in linear models by randomly weighted bootstrap. Sankhyā A. 1992; 54:323–331.

Therneau, TM.; Grambsch, PM. Introduction to Nonparametric Regression. Springer; New York: 2001.

Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc B. 1996; 58:267–288.

Tibshirani R. The Lasso method for variable selection in the cox model. Stat Med. 1997; 16:385–395. [PubMed: 9044528]

Wang H, Leng C. Unified Lasso estimation via least squares approximation. J Am Stat Assoc. 2007; 102(479):1039–1048.

Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. J Bus Econ Stat. 2007a; 25:347–355.

Wang H, Li G, Tsai CL. Regression coefficients and autoregressive order shrinkage and selection via the Lasso. J R Stat Soc Ser B. 2007b; 69:63–78.

Wang H, Li R, Tsai CL. Tuning parameter selector for SCAD. Biometrika. 2007c; 94:553–568. [PubMed: 19343105]

Wei LJ, Ying Z, Lin DY. Linear regression analysis for censored observations based on rank tests. Biometrika. 1990; 77:845–851.

Ying Z. A large sample study of rank estimation for censored regression data. Ann Stat. 1993; 21:76–99.

Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. Biometrika. 2007; 94:691–703.

Zou H. The adaptive Lasso and its oracle properties. J Am Stat Assoc. 2006; 101:1418–1429.

Zou H. A note on path-based variable selection in the penalized proportional hazards model. Biometrika. 2008; 95:241–247.

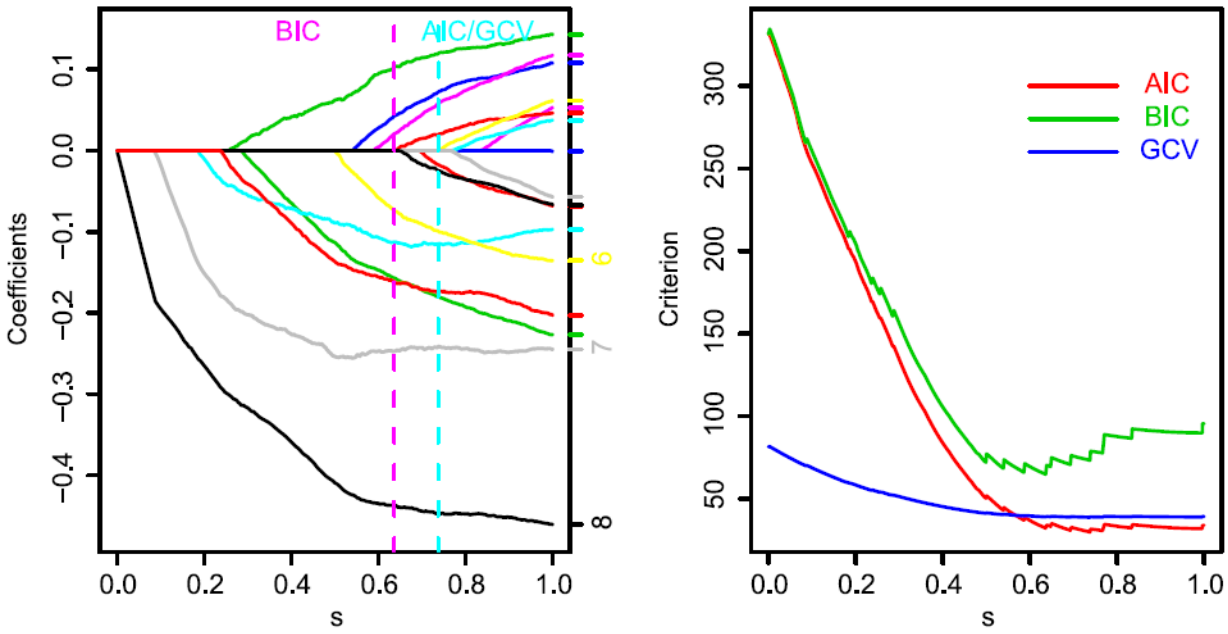

Fig. 1.

The *left panel* displays the Lasso estimates as a function of *s*. The *right panel* shows the AIC, BIC and GCV curves plotted against *s*

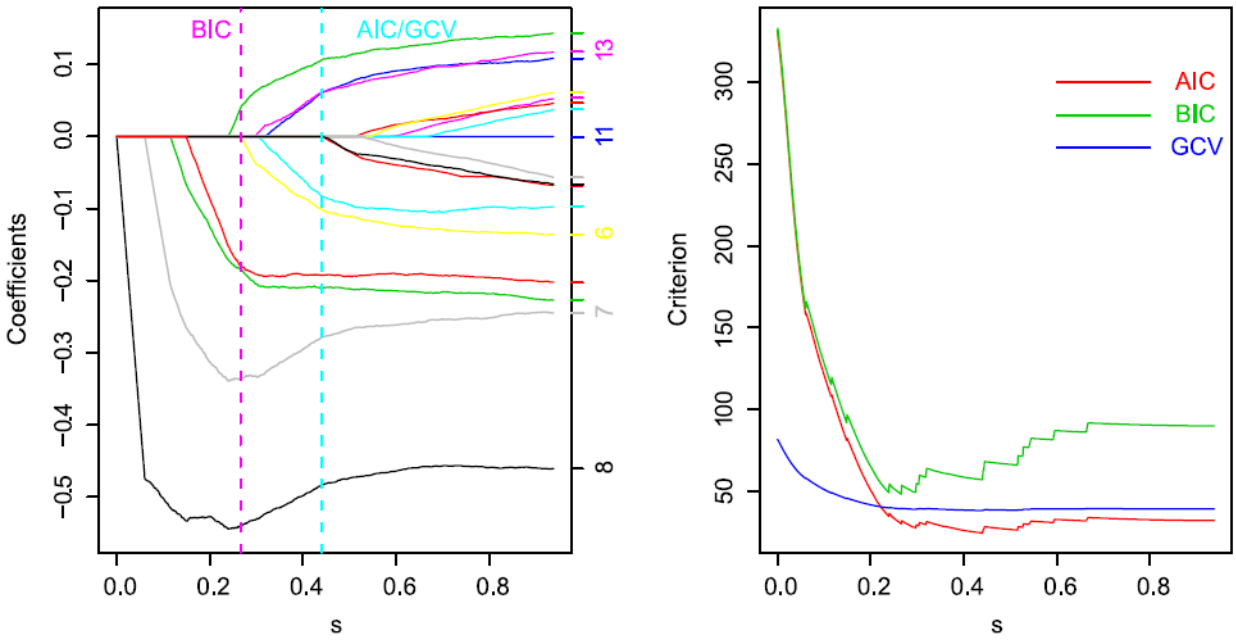

Fig. 2.

The *left panel* displays the adaptive Lasso estimates as a function of *s*. The *right panel* shows the AIC, BIC and GCV curves plotted against *s*

Table 1

Primary biliary cirrhosis data.

	$\hat{\beta}^G$	$\hat{\beta}^L$(AIC)	$\hat{\beta}^L$(BIC)	$\hat{\beta}^A$(AIC)	$\hat{\beta}^A$(BIC)
Drug	−0.068(0.058)	−0.018(0.039)	0(−)	0(−)	0(−)
Age	−0.227(0.060)	−0.180(0.064)	−0.157(0.062)	−0.210(0.066)	−0.185(0.062)
Sex	0.108(0.045)	0.073(0.045)	0.042(0.040)	0.060(0.045)	0(−)
Asc	−0.097(0.077)	−0.115(0.080)	−0.113(0.086)	−0.083(0.071)	0(−)
Hep	0.053(0.064)	0(−)	0(−)	0(−)	0(−)
Spid	−0.136(0.070)	−0.099(0.065)	−0.073(0.061)	−0.101(0.066)	0(−)
Ede	−0.245(0.087)	−0.241(0.088)	−0.247(0.095)	−0.278(0.085)	−0.336(0.085)
Logbil	−0.461(0.073)	−0.448(0.069)	−0.437(0.067)	−0.484(0.070)	−0.541(0.072)
Chol	0.046(0.059)	0.020(0.040)	0(−)	0(−)	0(−)
Logalb	0.143(0.072)	0.120(0.071)	0.103(0.072)	0.106(0.072)	0.039(0.056)
Cop	−0.001(0.065)	0(−)	0(−)	0(−)	0(−)
Logalk	0.037(0.055)	0(−)	0(−)	0(−)	
Sgot	0.117(0.056)	0.058(0.048)	0.021(0.039)	0.061(0.050)	0(−)
Trig	0.062(0.059)	0(−)	0(−)	0(−)	0(−)
Plat	−0.057(0.065)	0(−)	0(−)	0(−)	0(−)
Logprot	−0.067(0.061)	−0.025(0.042)	0(−)	0(−)	0(−)
Stage	−0.203(0.077)	−0.174(0.060)	−0.161(0.058)	−0.192(0.062)	−0.180(0.064)

Estimated coefficients of the Gehan estimate: $\hat{\beta}^G$, the Lasso estimates: $\hat{\beta}^L$, and the adaptive Lasso estimates: $\hat{\beta}^A$

Table 2

AIC, BIC, GCV and cross validation's predictive performance

Prediction error	Lasso	Adaptive Lasso
AIC	30.50	31.64
BIC	30.60	31.69
GCV	30.31	31.47
Cross-validation	30.28	31.42

Table 3

Framingham heart data.

	CHD			CVA		
	$\hat{\beta}^G$	$\hat{\beta}^L$	$\hat{\beta}^A$	$\hat{\beta}^G$	$\hat{\beta}^L$	$\hat{\beta}^A$
x_1	0.035	0.025	0.028	−0.131	−0.052	−0.045
x_2	0.047	0.037	0.037	−0.356	−0.091	−0.081
x_3	−0.052	−0.045	−0.046	0.071	0.044	0.024
x_4	−0.048	−0.041	−0.041	0.098	0.022	0.017
x_5	−0.046	−0.038	−0.039	−0.694	−0.160	−0.140
x_6	0.023	0.019	0.018	0.012	0	0
$x_1 * x_2$	0.003	0	0	−0.039	−0.028	−0.007
$x_1 * x_3$	−0.016	−0.013	−0.013	−0.013	0	0
$x_1 * x_4$	−0.033	−0.025	−0.024	−0.006	0.007	0
$x_1 * x_5$	−0.082	−0.071	−0.074	0.273	0.080	0.116
$x_1 * x_6$	0.015	0.011	0.007	−0.044	−0.033	−0.005
$x_2 * x_3$	0.037	0.029	0.024	−0.041	0	0
$x_2 * x_4$	0.008	0.008	0	0.050	0.006	0
$x_2 * x_5$	−0.006	0	0	0.415	0.066	0.097
$x_2 * x_6$	−0.026	−0.023	−0.021	0.004	−0.030	0
$x_3 * x_4$	0.022	0.015	0.010	0.100	0.060	0.029
$x_3 * x_5$	−0.010	−0.011	−0.005	−0.217	−0.118	−0.083
$x_3 * x_6$	−0.003	0	0	−0.019	0.010	0
$x_4 * x_5$	−0.004	−0.002	0	−0.249	−0.104	−0.096
$x_4 * x_6$	0.012	0.006	0	0.073	0.043	0.021
$x_5 * x_6$	0.041	0.032	0.031	−0.006	0	0

Estimated coefficients of the Gehan estimate: $\hat{\beta}^G$, the Lasso estimates: $\hat{\beta}^L$, and the adaptive Lasso estimates: $\hat{\beta}^A$

Table 4

Simulation results for the univariate failure time data. Mix denotes the mixture error distribution $0.5N(0,1) + 0.5N(0,9)$. Sample size is 100

Error	Censoring	Method	MRME	C	IC
$N(0,1)$	30%	Lasso(AIC)	2.32(2.45)	3(0)	2.43(1.38)
		Lasso(BIC)	2.37(2.75)	3(0)	1.46(1.15)
		Lasso(GCV)	2.28(2.80)	3(0)	1.72(1.26)
		ALasso(AIC)	1.48(1.55)	3(0)	1.00(1.12)
		ALasso(BIC)	1.13(0.77)	3(0)	0.27(0.57)
		ALasso(GCV)	1.27(0.97)	3(0)	0.61(0.86)
	50%	Lasso(AIC)	2.27(2.53)	3(0)	2.62(1.39)
		Lasso(BIC)	2.61(2.46)	3(0)	1.63(1.25)
		Lasso(GCV)	2.41(2.37)	3(0)	1.84(1.35)
		ALasso(AIC)	1.46(1.58)	3(0)	1.04(1.21)
		ALasso(BIC)	1.16(0.84)	3(0)	0.24(0.51)
		ALasso(GCV)	1.22(1.11)	3(0)	0.63(0.91)
t_3	30%	Lasso(AIC)	2.41(2.07)	3(0)	2.39(1.43)
		Lasso(BIC)	2.55(2.45)	3(0)	1.39(1.05)
		Lasso(GCV)	2.60(2.37)	3(0)	1.46(1.08)
		ALasso(AIC)	1.53(1.39)	3(0)	0.98(1.07)
		ALasso(BIC)	1.30(0.93)	3(0)	0.28(0.65)
		ALasso(GCV)	1.24(0.79)	3(0)	0.35(0.67)
	50%	Lasso(AIC)	2.61(2.27)	3(0)	2.40(1.38)
		Lasso(BIC)	2.78(2.47)	3(0)	1.56(1.06)
		Lasso(GCV)	2.68(2.31)	3(0)	1.58(1.04)
		ALasso(AIC)	1.75(1.70)	3(0)	1.02(1.06)
		ALasso(BIC)	1.35(1.04)	3(0)	0.27(0.58)
		ALasso(GCV)	1.38(1.08)	3(0)	0.41(0.73)
Mix	30%	Lasso(AIC)	2.30(2.70)	3(0)	2.30(1.37)
		Lasso(BIC)	2.37(2.57)	3(0)	1.52(1.26)
		Lasso(GCV)	2.38(2.75)	3(0)	1.52(1.28)
		ALasso(AIC)	1.64(1.71)	3(0)	1.06(1.15)

Error	Censoring	Method	MRME	C	IC
		ALasso(BIC)	1.46(1.47)	3(0)	0.47(0.82)
		ALasso(GCV)	1.50(1.47)	3(0)	0.56(0.88)
	50%	Lasso(AIC)	2.15(2.61)	3(0)	2.34(1.30)
		Lasso(BIC)	2.16(2.99)	3(0)	1.55(1.21)
		Lasso(GCV)	2.38(2.95)	3(0)	1.68(1.27)
		ALasso(AIC)	1.82(1.77)	3(0)	1.25(1.20)
		ALasso(BIC)	1.44(1.56)	3(0)	0.49(0.80)
		ALasso(GCV)	1.59(1.68)	3(0)	0.65(0.88)

Table 5

Simulation results when the signal-to-noise is small

Method	MRME	C	IC
Lasso(AIC)	1.92(3.71)	2.26(0.76)	1.79(1.41)
Lasso(BIC)	2.43(4.15)	1.44(0.94)	0.58(0.98)
Lasso(GCV)	2.03(3.58)	1.82(0.87)	0.82(1.03)
ALasso(AIC)	2.27(3.30)	2.01(0.69)	1.32(1.18)
ALasso(BIC)	2.44(3.78)	1.46(0.77)	0.52(0.80)
ALasso(GCV)	2.28(3.50)	1.71(0.73)	0.66(0.84)

Table 6

Simulation results for the multivariate failure time data. Sample size is 100. Censoring level is 50%

	θ	Method	MRME	C	IC
Multiple events	0	Lasso(AIC)	2.36(1.58)	6(0)	5.83(1.86)
		Lasso(BIC)	2.79(1.65)	6(0)	3.92(1.74)
		Lasso(GCV)	2.43(1.61)	6(0)	5.29(2.08)
		ALasso(AIC)	1.77(1.38)	6(0)	2.67(1.83)
		ALasso(BIC)	1.41(0.67)	6(0)	0.64(0.89)
		ALasso(GCV)	1.68(1.18)	6(0)	2.45(1.82)
	1	Lasso(AIC)	2.42(1.78)	6(0)	5.63(1.86)
		Lasso(BIC)	2.69(1.95)	6(0)	3.89(1.75)
		Lasso(GCV)	2.44(1.79)	6(0)	5.32(2.02)
		ALasso(AIC)	1.81(1.44)	6(0)	2.68(1.80)
		ALasso(BIC)	1.41(0.70)	6(0)	0.68(0.89)
		ALasso(GCV)	1.74(1.55)	6(0)	2.46(1.90)
Clustered data	0	Lasso(AIC)	2.23(2.42)	3(0)	2.57(1.36)
		Lasso(BIC)	2.56(2.53)	3(0)	1.55(1.17)
		Lasso(GCV)	2.36(2.32)	3(0)	1.79(1.30)
		ALasso(AIC)	1.28(1.45)	3(0)	0.94(1.27)
		ALasso(BIC)	1.04(0.86)	3(0)	0.21(0.40)
		ALasso(GCV)	1.08(1.10)	3(0)	0.54(0.84)
	1	Lasso(AIC)	2.33(2.02)	3(0)	2.40(1.32)
		Lasso(BIC)	2.70(3.70)	3(0)	1.24(1.10)
		Lasso(GCV)	2.54(2.46)	3(0)	1.68(1.14)
		ALasso(AIC)	1.12(1.36)	3(0)	0.72(0.96)
		ALasso(BIC)	1.00(0.53)	3(0)	0.11(0.25)
		ALasso(GCV)	1.03(0.80)	3(0)	0.32(0.78)
Recurrent events	0	Lasso(AIC)	2.63(2.29)	3(0)	2.42(1.39)
		Lasso(BIC)	2.76(2.48)	3(0)	1.58(1.04)
		Lasso(GCV)	2.64(2.32)	3(0)	1.61(1.06)

θ	Method	MRME	C	IC
	ALasso(AIC)	1.74(1.68)	3(0)	1.04(1.09)
	ALasso(BIC)	1.38(1.02)	3(0)	0.29(0.56)
	ALasso(GCV)	1.42(1.10)	3(0)	0.43(0.76)
1	Lasso(AIC)	2.36(2.54)	3(0)	2.76(1.34)
	Lasso(BIC)	2.80(3.90)	3(0)	1.26(1.08)
	Lasso(GCV)	2.50(2.76)	3(0)	1.78(1.36)
	ALasso(AIC)	1.48(1.60)	3(0)	1.16(1.26)
	ALasso(BIC)	1.26(0.86)	3(0)	0.18(0.42)
	ALasso(GCV)	1.32(1.26)	3(0)	0.50(0.69)