Adaptive Lasso for Cox's Proportional Hazards Model

By HAO HELEN ZHANG and WENBIN LU

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.

hzhang@stat.ncsu.edu lu@stat.ncsu.edu

Summary

We investigate the variable selection problem for Cox's proportional hazards model, and propose a unified model selection and estimation procedure with desired theoretical properties and computational convenience. The new method is based on a penalised log partial likelihood with the adaptively-weighted L_1 penalty on regression coefficients, providing what we call the adaptive Lasso estimator. The method incorporates different penalties for different coefficients: unimportant variables receive larger penalties than important ones, so that important variables tend to be retained in the selection process, whereas unimportant variables are more likely to be dropped. Theoretical properties, such as consistency and rate of convergence of the estimator, are studied. We also show that, with proper choice of regularisation parameters, the proposed estimator has the oracle properties. The convex optimisation nature of the method leads to an efficient computation algorithm. Both simulated and real examples show that the method performs competitively.

Some key words: Adaptive Lasso; Penalised partial likelihood; Proportional hazards model; Variable selection.

1. Introduction

In the study of the dependence of survival time T on covariates $z = (z_1, \ldots, z_d)^{\mathrm{T}}$. Cox's proportional hazards model (Cox 1972, 1975) includes a hazard function $h(t|z)$ of a subject with covariates z of the form

$$h(t|z) = h_0(t) \exp(\beta^{\mathrm{T}} z), \tag{1}$$

where $h_0(t)$ is a completely unspecified baseline hazard function and $\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ is an unknown vector of regression coefficients.

In practice, not all the d covariates may contribute to the prediction of survival outcomes: some components of β may be zero in the true model. When the sample size goes to infinity, an ideal model selection and estimation procedure should be able to identify the true model with probability one, and provide consistent and efficient estimators for the relevant regression coefficients. In this paper, we propose a new procedure, the adaptive Lasso estimator, and show that it satisfies all theoretical properties.

Many variable selection techniques for linear regression models have been extended to the context of survival models. They include the best-subset selection, stepwise selection, asymptotic procedures based on score tests, Wald tests and other approximate chi-square testing procedures, bootstrap procedures (Sauerbrei & Schumacher, 1992), and Bayesian variable selection (Faraggi & Simon, 1998; Ibrahim et al., 1999). However, the theoretical properties of these methods are generally unknown (Fan & Li, 2002). Recently a family of penalised partial likelihood methods, such as the Lasso (Tibshirani, 1997) and the smoothly clipped absolute deviation method (Fan & Li, 2002) were proposed for Cox's proportional hazards model. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously. The Lasso estimator

does not possess the oracle properties (Fan & Li, 2002). The smoothly clipped absolute deviation estimator, proposed first by Fan & Li (2001) for linear models, has better theoretical properties than the Lasso, but the nonconvex form of its penalty makes its optimisation challenging in practice, and the solutions may suffer from numerical instability.

Our adaptive Lasso method is based on a penalised partial likelihood with adaptively weighted L_1 penalties on regression coefficients. Unlike the Lasso and smoothly clipped absolute deviation methods which apply the same penalty to all the coefficients, the adaptive Lasso penalty has the form $\lambda \sum_{j=1}^{d} |\beta_j| \tau_j$, with small weights τ_j chosen for large coefficients and large weights for small coefficients. In contrast to the Lasso, the new estimator enjoys the oracle properties. In contrast to the smoothly clipped absolute deviation method, the adaptive Lasso penalty has a convex form, which ensures the existence of global optimisers and can be efficiently solved by standard algorithms (Boyd & Vandenberghe, 2004).

2. Variable Selection Using Penalised Partial Likelihood

Suppose a random sample of n individuals is chosen. Let T_i and C_i be the failure time and censoring time of subject i $(i = 1, \ldots, n)$, respectively. Define $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. We use $z_i = (z_{i1}, \ldots, z_{id})^{\mathrm{T}}$ to denote the vector of covariates for the ith individual. Assume that T_i and C_i are conditionally independent given z_i, and the censoring mechanism is noninformative. The data then consists of the triplets $(\tilde{T}_i, \delta_i, z_i), i = 1, \ldots, n$.

The proportional hazards model (1) is assumed for the failure times T_i. For simplicity, assume that there are no ties in the observed failure times. When ties are present, we may use the technique in Breslow (1974). The log partial likelihood is then given by

$$l_n(\beta) \equiv \sum_{i=1}^{n} \delta_i [\beta^{\mathrm{T}} z_i - \log\{\sum_{j=1}^{n} I(\tilde{T}_j \geq \tilde{T}_i) \exp(\beta^{\mathrm{T}} z_j)\}]. \tag{2}$$

To select important variables under the proportional hazards model, Tibshirani (1997) and Fan & Li (2002) proposed to minimise the penalised log partial likelihood function,

$$-\frac{1}{n}l_n(\beta) + \lambda \sum_{j=1}^{d} J(\beta_j). \tag{3}$$

The Lasso penalty is $J(\beta_j) = |\beta_j|$, which shrinks small coefficients to zero and hence results in a sparse representation of the solution. However, estimation of large values of β_j's may suffer from substantial bias if λ is chosen too big, while the model may not be sufficiently sparse if λ is chosen too small. Fan & Li (2002) suggested the smoothly clipped absolute deviation penalty, which cleverly avoids excessive penalties on large coefficients and enjoys the oracle properties.

3. Adaptive Lasso Estimation

3·1. *The estimator*

Our adaptive Lasso estimator is the solution of

$$\min_{\beta}\{-\frac{1}{n}l_n(\beta) + \lambda \sum_{j=1}^{d} |\beta_j|\tau_j\}, \tag{4}$$

where the positive weights $\tau = (\tau_1, \ldots, \tau_d)^{\mathrm{T}}$ are chosen adaptively by data. The values chosen for the τ_j's are crucial for guaranteeing the optimality of the solution. Our proposal is to use $\tau_j = 1/|\tilde{\beta}_j|$, where $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)^{\mathrm{T}}$ is the maximiser of the log partial likelihood $l_n(\beta)$. Since $\tilde{\beta}$ are consistent estimator (Tsiatis, 1981; Andersen & Gill, 1982), their values well reflect the relative importance of the covariates. We therefore focus on the problem

$$\min_{\beta}\{-\frac{1}{n}l_n(\beta) + \lambda \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|\}. \tag{5}$$

Any consistent estimator of β can be used, and $\tilde{\beta}$ is just a convenient choice. Note that the adaptive penalty term in (5) is closely related to the L_0 penalty $\sum_{j=1}^{d} I(|\beta_j| \neq 0)$, also

called the entropy penalty in wavelet literature (Donoho & Johnstone 1998; Antoniadis & Fan 2001). As a result of the consistency of $\tilde{\beta}_j$, the term $|\beta_j|/|\tilde{\beta}_j|$ converges to $I(\beta_j \neq 0)$ in probability as n goes to infinity. Therefore the adaptive Lasso procedure can be regarded as an automatic implementation of the best-subset selection in some asymptotic sense.

3·2. *Theoretical properties of the estimator*

We study the asymptotic properties of the estimator from two perspectives. Consider the penalised log partial likelihood function based on n samples,

$$Q_n(\beta) = l_n(\beta) - n\lambda_n \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|. \tag{6}$$

Write the true parameter vector as $\beta_0 = (\beta_{10}^{\mathrm{T}}, \beta_{20}^{\mathrm{T}})^{\mathrm{T}}$, where β_{10} consists of all q nonzero components and β_{20} consists of the remaining zero components. Correspondingly, we write the maximiser of (6) as $\hat{\beta}_n = (\hat{\beta}_{1n}^{\mathrm{T}}, \hat{\beta}_{2n}^{\mathrm{T}})^{\mathrm{T}}$.

Define the counting and at-risk processes $N_i(t) = \delta_i I(\tilde{T}_i \leq t)$ and $Y_i(t) = I(\tilde{T}_i \geq t)$, respectively. In this section, the covariate z is allowed to be time-dependent, denoted by $z(t)$. Without loss of generality, assume that $t \in [0, 1]$. Then the Fisher information matrix is

$$I(\beta_0) = \int_0^1 v(\beta_0, t) s^{(0)}(\beta_0, t) h_0(t) dt,$$

where

$$v(\beta, t) = \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \left(\frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)}\right)\left(\frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)}\right)^{\mathrm{T}},$$

and $s^{(k)}(\beta, t) = E[z(t)^{\otimes k} Y(t) \exp\{\beta' z(t)\}]$, $k = 0, 1, 2$. The regularity conditions (A)-(D) used in Anderson & Gill (1982) are assumed in the whole section. Let $I(\beta_0)$ be the Fisher information matrix based on the log partial likelihood and let $I_1(\beta_{10}) = I_{11}(\beta_{10}, 0)$, where $I_{11}(\beta_{10}, 0)$ is the first $s \times s$ submatrix of $I(\beta_0)$ with $\beta_{20} = 0$. The following theorem shows

that $\hat{\beta}_n$ is root-n consistent if $\lambda_n \to 0$ at an appropriate rate.

THEOREM 1. *Assume that* $(z_1, T_1, C_1), ..., (z_n, T_n, C_n)$ *are independently and identically distributed, and that* T_i *and* C_i *are independent given* z_i. *If* $\sqrt{n}\lambda_n = O_p(1)$, *then the adaptive Lasso estimator satisfies* $\|\hat{\beta}_n - \beta_0\| = O_p(n^{-1/2})$.

NEXT WE SHOW, WHEN λ_n IS CHOSEN PROPERLY, THE ADAPTIVE LASSO ESTIMATOR HAS THE ORACLE PROPERTY (DONOHO & JOHNSTONE, 1994); THAT IS, AS n GOES TO INFINITY, THE ADAPTIVE LASSO CAN PERFORM AS WELL AS IF THE CORRECT SUBMODEL WERE KNOWN.

THEOREM 2. *Assume that* $\sqrt{n}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$, *then, under the conditions of Theorem 1, with probability tending to 1, the root-n consistent adaptive Lasso estimator* $\hat{\beta}_n$ *must satisfy the following conditions:*

(i) *(Sparsity)* $\hat{\beta}_{2n} = 0$;

(ii) *(Asymptotic normality)* $\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) \to N\{0, I_1^{-1}(\beta_{10})\}$ *in distribution as n goes to infinity.*

Proofs of both theorems are given in the Appendix 1. Since the proofs only require the root-n consistency of $\tilde{\beta}$, it is worth noting that, any root-n consistent estimator of β_0 can be used as the adaptive weights τ without changing the asymptotic properties of the adaptive Lasso solution.

4. COMPUTATIONAL ALGORITHM

4·1. *The optimisation routine*

The optimisation problem (5) is strictly convex and therefore can be solved by many standard software packages such as MATLAB, R and MINOS, and algorithms like the interior point algorithm (Boyd & Vandenberghe, 2004). Here we present our algorithm for finding

the maximiser $\hat{\beta}$ of (5). We approximate the partial likelihood function using the Newton-Raphson update through an iterative least squares procedure, at each iteration solving the least squares problem subject to the weighted L_1 penalty. Define the gradient vector $\nabla l(\beta) = -\partial l_n(\beta)/\partial\beta$ and the Hessian matrix $\nabla^2 l(\beta) = -\partial^2 l_n(\beta)/\partial\beta\partial\beta^{\mathrm{T}}$. Consider the Cholesky decomposition of $\nabla^2 l(\beta)$, i.e. $\nabla^2 l(\beta) = X^{\mathrm{T}}X$, and set the pseudo response vector $Y = (X^{\mathrm{T}})^{-1}\{\nabla^2 l(\beta)\beta - \nabla l(\beta)\}$. By second-order Taylor expansion, $-l_n(\beta)$ can be approximated by the quadratic form $\frac{1}{2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta)$. Thus at each iterative step, we need to minimise

$$\frac{1}{2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta) + \lambda\sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|. \tag{7}$$

To solve the standard Lasso, Tibshirani (1996) suggested two algorithms based on quadratic programming techniques, and Fu (1998) proposed the shooting algorithm. Recently Efron et al. (2004) showed that, in the least squares setting, the whole solution path of Lasso can be obtained by a modified least angle regression algorithm. To minimise (7), we have slightly modified Fu's shooting algorithm; see Appendix 2. For any fixed λ, the following is the complete algorithm for solving (5).

Step 1. Obtain $\tilde{\beta}$ by minimising the negative log partial likelihood $-l_n(\beta)$.

Step 2. Initialise by setting $k = 1$ and $\hat{\beta}_{[1]} = 0$.

Step 3. Compute $\nabla l, \nabla^2 l, X$ and Y based on the current value $\hat{\beta}_{[k]}$.

Step 4. Minimise (7) using the modified shooting algorithm, denoting the solution by $\hat{\beta}_{[k+1]}$.

Step 5. Let $k = k + 1$. Go back to Step 3 until the convergence criterion is met.

This algorithm gives exact zeros for some coefficients and it converges fast based on our empirical experience. Similarly to Theorem 3 in Fu (1998), we can show that the modified shooting algorithm is guaranteed to converge to the global minimiser of (7). The

Lasso optimisation, as a special case with all weights equal to 1, can also be solved by this algorithm.

4·2. *Variance estimation and parameter tuning*

For their estimations, Tibshirani (1996) and Fan & Li (2002) proposed standard error formulae based on their approximated ridge solutions, and we follow their methods here. Define $A(\beta) = \mathrm{diag}\{1/\beta_1^2, \ldots, 1/\beta_d^2\}$,

$$D(\beta) = \mathrm{diag}\left\{\frac{I(\beta_1 \neq 0)}{\beta_1^2}, \ldots, \frac{I(\beta_d \neq 0)}{\beta_d^2}\right\}, \quad b(\beta) = \left(\frac{\mathrm{sign}(|\beta_1|)}{|\tilde{\beta}_1|}, \ldots, \frac{\mathrm{sign}(|\beta_d|)}{|\tilde{\beta}_d|}\right)'.$$

At the $(k+1)$th step, the adaptive Lasso solution can be approximated by

$$\hat{\beta}_{[k+1]} = \hat{\beta}_{[k]} - \left\{\nabla^2 l(\hat{\beta}_{[k]}) + \lambda A(\hat{\beta}_{[k]})\right\}^{-1} \left\{\nabla l(\hat{\beta}_{[k]}) + \lambda b(\hat{\beta}_{[k]})\right\}.$$

Using similar to those in Fan & Li (2002), we can approximate the covariance matrix of the adaptive Lasso estimator $\hat{\beta}$ by the following sandwich formula:

$$\left\{\nabla^2 l(\hat{\beta}) + \lambda A(\hat{\beta})\right\}^{-1} \Sigma(\hat{\beta}) \left\{\nabla^2 l(\hat{\beta}) + \lambda A(\hat{\beta})\right\}^{-1},$$

where $\Sigma(\hat{\beta}) = \{\nabla^2 l(\hat{\beta}) + \lambda D(\hat{\beta})\}\{\nabla^2 l(\hat{\beta})\}^{-1}\{\nabla^2 l(\hat{\beta}) + \lambda D(\hat{\beta})\}$.

Let $\hat{\beta} = (\hat{\beta}_1^{\mathrm{T}}, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\beta}_1$ corresponds to the r nonzero components. Correspondingly, we may decompose the Hessian matrix as

$$G = \nabla^2 l(\hat{\beta}) = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} corresponds to the first $r \times r$ submatrix of G. Similarly, let A_{11} be the first $r \times r$ submatrix of $A \equiv A(\hat{\beta})$. Define $E = G_{22} - G_{21}G_{11}^{-1}G_{12}$ and $\tilde{G}_{11} = G_{11} + \lambda A_{11}$. It is easy to show that the estimator of the covariance matrix of $\hat{\beta}_1$ is

$$\hat{\mathrm{cov}}(\hat{\beta}_1) = G_{11}^{-1} + \left(G_{11}^{-1} - \tilde{G}_{11}^{-1}\right) G_{12} E^{-1} G_{21} \left(G_{11}^{-1} - \tilde{G}_{11}^{-1}\right). \tag{8}$$

If λ is small, then (8) can be well approximated by G_{11}^{-1}, the inverse of the observed Fisher information matrix $\hat{I}_1$. This is consistent with the asymptotic covariance matrix of $\hat{\beta}_1$ in Theorem 2.

To estimate the tuning parameter λ, we use generalised crossvalidation (Craven & Wahba, 1979). At convergence, the minimiser of (6) in Step 4 can be approximated by a ridge solution $(G + \lambda A)^{-1} X^{\mathrm{T}} Y$. Therefore, the number of effective parameters in the adaptive Lasso estimator can be approximated by $p(\lambda) = \mathrm{tr}\{(G + \lambda A)^{-1} G\}$, and the generalised crossvalidation constructed is $\mathrm{GCV}(\lambda) = -l_n(\hat{\beta})/\{n\{1 - p(\lambda)/n\}^2\}$.

5. Numerical Studies

5·1. *Simulations*

We compare the performance of the maximum partial likelihood estimators, Lasso, smoothly clipped absolute deviation and adaptive Lasso, under Cox's proportional hazards model. We report the average numbers of correct and incorrect zero coefficients in the final model over 100 replicates. To measure prediction accuracy, we follow Tibshirani (1997) and summarise the average mean squared errors (MSE) $(\hat{\beta} - \beta)^{\mathrm{T}} V (\hat{\beta} - \beta)$ over 100 runs. Here V is the population covariance matrix of the covariates. Generalised crossvalidation is used to estimate the tuning parameter λ in the Lasso, smoothly clipped absolute deviation, and adaptive Lasso.

The failure times are generated from the proportional hazards model (1) in two settings.

Model 1: $\beta = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^{\mathrm{T}}$, corresponding to large effects.

Model 2: $\beta = (-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^{\mathrm{T}}$, corresponding to small effects.

The nine covariates $z = (z_1, ..., z_9)$ are marginally standard normal with the pairwise correlations $\mathrm{corr}(z_j, z_k) = \rho^{|j-k|}$. We assume moderate correlation between the covariates by

taking $\rho = 0.5$. Censoring times are generated from a $\mathrm{Un}(0, c_0)$ distribution, where c_0 is chosen to obtain the desired censoring rate. We used two censoring rates, 25% and 40%, and three sample sizes, $n = 100, 200$ and 300.

Table 1 summarises the mean squared errors and variable selection results for Model 1, where important variables have large effects. Standard errors are given in parentheses. Overall, the adaptive Lasso performs best in terms of both variable selection and prediction accuracy. For example, when $n = 100$ and the censoring rate is 25%, the adaptive Lasso selects important covariates most accurately; the true model size is 3, whereas the average size from maximum likelihood is 9, from Lasso is 4.13, from smoothly clipped absolute deviation is 3.62 and from adaptive Lasso is 3.27. Adaptive Lasso also gives the smallest mean squared error. Between the two procedures with oracle properties, the adaptive Lasso performs consistently better in term of variable selection. With regard to mean squared error, the adaptive Lasso is better when $n = 100$; as n increases to 200 or 300, two methods become equally good. In Table 2, we present the frequency with which each variable was selected among 100 runs for the 25% censoring case. The adaptive Lasso chooses unimportant variables much less often than the other two procedures in all the settings, and the Lasso is the worst. Similar results are observed for the 40% censored case.

(Insert Tables 1 and 2 here)

To test the accuracy of the proposed standard error formula given in §3·2, we compare the sample standard errors with their estimates. For the Lasso estimates, we use the standard error formula in Tibshirani (1997). For the smoothly clipped absolute deviation estimates, the formula in Fan & Li (2002) is used. Table 3 gives the mean of the estimated standard errors and the sample standard errors from Monte Carlo simulations for the 25% censoring case. Similar results are found for 40% censoring case. For all methods, there

is a discrepancy between the estimated standard errors and the sample standard errors when n is small, but it decreases when n gets large. Also, we observed that the smoothly clipped absolute deviation solutions were not as robust as the other estimators, showing large variability among replicated solutions. Therefore the values in Table 3 corresponds to the robust estimator of the sample standard error (Fan & Li, 2002), which is calculated as the median absolute deviation of the estimates divided by 0.6745, and compare it with the median of the estimated standard errors. For the other procedures, standard variance estimators are used.

(Insert Table 3 here)

In Model 2, important variables have smaller effects than in Model 1 and the coefficients are of different magnitudes. Table 4 shows that, for variable selection, the adaptive Lasso is best in terms of selecting correct zeros. With regard to prediction accuracy, the Lasso, adaptive Lasso and maximum likelihood give similar mean squared errors, with Lasso slightly better, and the smoothly clipped absolute deviation method is consistently worse than the others.

(Insert Tables 4 here)

5·2. *Primary biliary cirrhosis data*

Data gathered in the Mayo Clinic trial in primary biliary cirrhosis of liver conducted between 1974 and 1984, are provided in Therneau & Grambsch (2000); a more detailed account can be found in Dickson et al. (1989). In this study, 312 out of 424 patients who agreed to participate in the randomised trial are eligible for the analysis. For each patient, clinical, biochemical, serological and histological parameters are collected. In all, 125 patients died before the end of follow-up. We study the dependence of the survival time on seventeen

covariates: continuous variables are age (in years), alb (albumin in g/dl), alk (alkaline phosphatase in units/liter), bil (serum bilirunbin in mg/dl), chol (serum cholesterol in mg/dl), cop (urine copper in μg/day), plat (platelets per cubic ml/1000), prot (prothrombin time in seconds), sgot (liver enzyme in units/ml), trig (triglycerides in mg/dl); categorical variables are asc (0 denotes absence of ascites and 1 denotes presence of ascites), oed (0 denotes no oedema, 0.5 denotes untreated or successfully treated oedema and 1 denotes unsuccessfully treated oedema), hep (0 denotes absence of hepatomegaly and 1 denotes presence of hepatomegaly), sex (0 denotes male and 1 denotes female), spid (0 denotes absence of spiders and 1 denotes presence of spiders), stage (histological stage of disease, graded 1, 2, 3 or 4) and trt (1 for control and 2 for treatment).

We restrict our attention to the 276 observations without missing values. All seventeen variables are included in the model. Table 5 gives the estimated coefficients by three methods, together with the corresponding standard errors. As reported in Tibshirani (1997), the stepwise selection chooses eight variables, namely age, oed, bil, alb, cop, sgot, prot and stage, sets of variables similar to that set are chosen by each of the Lasso, smoothly clipped absolute deviation and adaptive Lasso methods (details clear in Table 5).

(Insert Table 5 here)

6. Discussion

For the adaptive Lasso procedure, the choice of the weights τ_j's is very important. We have used $\tau_j = 1/|\tilde{\beta}_j|$. However, β_j's may not be estimable in cases such as high-dimensional gene expression data where the number of covariates d is much larger than the sample size n, or the $\tilde{\beta}_j$'s may be unstable if strong collinearity exists among covariates. In such cases, we suggest using robust estimators such as ridge regression estimators in order to determine

the weights.

Acknowledgment

The authors are grateful to the referees, associate editor and editor for their constructive comments. The research of both authors was partially supported by grants from the U.S. National Science Foundation.

Appendix 1

Proof of Theorems

We follow steps similar to the proofs of Fan & Li (2002). Throughout the paper, we define $s_n(\beta) = \partial l_n(\beta)/\partial\beta$ and $\nabla s_n(\beta) = \partial s_n(\beta)/\partial\beta^{\mathrm{T}}$.

Proof of Theorem 1. The log partial likelihood $l_n(\beta)$ can be written as

$$l_n(\beta) = \sum_{i=1}^{n} \int_0^1 \beta^{\mathrm{T}} z_i(s) dN_i(s) - \int_0^1 \log \left[\sum_{i=1}^{n} Y_i(s) \exp\{\beta^{\mathrm{T}} z_i(s)\} \right] d\bar{N}(s), \qquad \text{(A1)}$$

where $\bar{N} = \sum_{i=1}^{n} N_i$. By Theorem 4.1 and Lemma 3.1 of Anderson & Gill (1982), it follows that, for each β in a neighbourhood of β_0,

$$\begin{aligned} \frac{1}{n}\{l_n(\beta) - l_n(\beta_0)\} &= \int_0^1 \left[(\beta - \beta_0)^T s^{(1)}(\beta_0, t) - \log \left\{ \frac{s^{(0)}(\beta, t)}{s^{(0)}(\beta_0, t)} \right\} s^{(0)}(\beta_0, t) \right] \lambda_0(t) dt \\ &+ O_p(\frac{\|\beta - \beta_0\|}{\sqrt{n}}). \qquad \text{(A2)} \end{aligned}$$

Consider the C-ball $B_n(C) = \{\beta : \beta = \beta_0 + n^{-1/2}u, ||u|| \leq C\}$, $C > 0$, and denote its boundary by $\partial B_n(C)$. Note that $Q_n(\beta)$ is strictly convex when n is large. Thus, there exists a unique maximiser $\hat{\beta}_n$ of $Q_n(\beta)$ for large n. It is sufficient to show that, for any given $\epsilon > 0$, there exists a large constant C so that

$$\Pr \left\{ \sup_{\beta \in \partial B_n(C)} Q_n(\beta) < Q_n(\beta_0) \right\} \geq 1 - \epsilon. \qquad \text{(A3)}$$

This implies that, with probability at least $1 - \epsilon$, there exists a local maximiser of $Q_n(\beta)$ in the ball $B_n(C)$. Hence, the maximiser $\hat{\beta}_n$ must satisfy $\|\hat{\beta}_n - \beta_0\| = O_p(n^{-1/2})$.

Furthermore, we have $s_n(\beta_0)/\sqrt{n} = O_p(1)$ and $\nabla s_n(\beta_0)/n = I(\beta_0) + o_p(1)$. For any $\beta \in \partial B_n(C)$, by the second-order Taylor expansion of the log partial likelihood, we have

$$\begin{aligned} & \frac{1}{n}\{l_n(\beta_0 + n^{-1/2}u) - l_n(\beta_0)\} \\ = & \frac{1}{n}s_n^{\mathrm{T}}(\beta_0)n^{-1/2}u - \frac{1}{2n}u^{\mathrm{T}}\{\nabla s_n(\beta_0)/n\}u + \frac{1}{n}u^{\mathrm{T}}o_p(1)u \\ = & -\frac{1}{2n}u^{\mathrm{T}}\{I(\beta_0) + o_p(1)\}u + \frac{1}{n}O_p(1)\sum_{j=1}^{d}|u_j|, \end{aligned}$$

where $u = (u_1, \ldots, u_d)^{\mathrm{T}}$. Then we have

$$\begin{aligned} D_n(u) \equiv & \frac{1}{n}\{Q_n(\beta_0 + n^{-1/2}u) - Q_n(\beta_0)\} \\ = & \frac{1}{n}\{l_n(\beta_0 + n^{-1/2}u) - l_n(\beta_0)\} - \lambda_n\sum_{j=1}^{d}\left(\frac{|\beta_{j0} + n^{-1/2}u_j|}{|\tilde{\beta}_j|} - \frac{|\beta_{j0}|}{|\tilde{\beta}_j|}\right) \\ \leq & \frac{1}{n}\{l_n(\beta_0 + n^{-1/2}u) - l_n(\beta_0)\} - \lambda_n\sum_{j=1}^{s}(|\beta_{j0} + n^{-1/2}u_j| - |\beta_{j0}|)/|\tilde{\beta}_j| \\ \leq & \frac{1}{n}\{l_n(\beta_0 + n^{-1/2}u) - l_n(\beta_0)\} + n^{-1/2}\lambda_n\sum_{j=1}^{s}|u_j|/|\tilde{\beta}_j| \\ = & -\frac{1}{2n}u^{\mathrm{T}}\{I(\beta_0) + o_p(1)\}u + \frac{1}{n}O_p(1)\sum_{j=1}^{d}|u_j| + \frac{1}{\sqrt{n}}\lambda_n\sum_{j=1}^{s}|u_j|/|\tilde{\beta}_j|. \end{aligned} \tag{A4}$$

Since the maximum partial likelihood estimator $\tilde{\beta}$ satisfies $||\tilde{\beta} - \beta_0|| = O_p(n^{-1/2})$, we have, for $1 \leq j \leq s$,

$$\frac{1}{|\tilde{\beta}_j|} = \frac{1}{|\beta_{j0}|} - \frac{\mathrm{sign}(\beta_{j0})}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0}) + o_p(|\tilde{\beta}_j - \beta_{j0}|) = \frac{1}{|\beta_{j0}|} + \frac{O_p(1)}{\sqrt{n}}.$$

In addition, since $\sqrt{n}\lambda_n = O_p(1)$, we have

$$\begin{aligned} \frac{1}{\sqrt{n}}\lambda_n\sum_{j=1}^{s}|u_j|/|\tilde{\beta}_j| = & \frac{1}{\sqrt{n}}\lambda_n\sum_{j=1}^{s}\left(\frac{|u_j|}{|\beta_{j0}|} + \frac{|u_j|}{\sqrt{n}}O_p(1)\right) \\ \leq & Cn^{-1/2}\lambda_n O_p(1) = Cn^{-1}(\sqrt{n}\lambda_n)O_p(1) = Cn^{-1}O_p(1). \end{aligned}$$

Therefore in (A4), if we choose a sufficiently large C, the first term is of the order C^2n^{-1}. The second and third terms are of the order Cn^{-1}, which are dominated by the first term. Therefore (A3) holds and it completes the proof.

Proof of Theorem 2. (i) Here we show that $\hat{\beta}_{2n} = 0$.

It is sufficient to show that, for any sequence β_1 satisfying $\|\beta_1 - \beta_{10}\| = O_p(n^{-1/2})$ and for any constant C,

$$Q_n(\beta_1, 0) = \max_{\|\beta_2\| \le Cn^{-1/2}} Q_n(\beta_1, \beta_2).$$

We will show that, with probability tending to 1, for any β_1 satisfying $\|\beta_1 - \beta_0^{(1)}\| = O_p(n^{-1/2})$, $\partial Q(\beta)/\partial\beta_j$ and β_j have different signs for $\beta_j \in (-Cn^{-1/2}, Cn^{-1/2})$ with $j = s+1, \ldots, d$. For each β in a neighbourhood of β_0, by (A1) and Taylor expansion,

$$l_n(\beta) = l_n(\beta_0) + nf(\beta) + O_P(\sqrt{n}\|\beta - \beta_0\|),$$

where $f(\beta) = -\frac{1}{2}(\beta - \beta_0)'\{I(\beta_0) + o(1)\}(\beta - \beta_0)$. For $j = s+1, \ldots, d$, we have

$$\frac{\partial Q_n(\beta)}{\partial \beta_j} = \frac{\partial l_n(\beta)}{\partial \beta_j} - n\lambda_n \frac{\mathrm{sign}(\beta_j)}{|\tilde{\beta}_j|} = O_p(n^{1/2}) - (n\lambda_n)n^{1/2}\frac{\mathrm{sign}(\beta_j)}{|n^{1/2}\tilde{\beta}_j|}.$$

Note that $n^{1/2}(\tilde{\beta}_j - 0) = O_p(1)$, so that we have

$$\frac{\partial Q_n(\beta)}{\partial \beta_j} = n^{1/2}\left\{O_p(1) - n\lambda_n \frac{\mathrm{sign}(\beta_j)}{|O_p(1)|}\right\}. \tag{A5}$$

Since $n\lambda_n \to \infty$, the sign of $\frac{\partial Q_n(\beta_j)}{\partial \beta_j}$ in (A5) is completely determined by the sign of β_j when n is large, and they always have different signs.

(ii) Here we show the asymptotic normality of $\hat{\beta}_{1n}$.

From the proof of Theorem 1, it is easy to show that there exists a root-n consistent maximiser $\hat{\beta}_{1n}$ of $Q_n(\beta_1, 0)$, i.e. $\frac{\partial Q_n(\beta)}{\partial \beta_1}|_{\beta = \{\hat{\beta}_{1n}^{\mathrm{T}}, 0^{\mathrm{T}}\}^{\mathrm{T}}} = 0$. Let $s_{1n}(\beta)$ be the first q elements of $s_n(\beta)$ and let $\hat{I}_{11}(\beta)$ be the first $q \times q$ submatrix of $\nabla s_n(\beta)$. Then

$$\begin{aligned}
0 &= \frac{\partial Q_n(\beta)}{\partial \beta_1}|_{\beta = \{\hat{\beta}'_{1n}, 0'\}'} \\
&= \frac{\partial l_n(\beta)}{\partial \beta_1|}|_{\beta = \{\hat{\beta}'_{1n}, 0'\}'} - n\lambda_n \left(\frac{\mathrm{sign}(\hat{\beta}_1)}{\tilde{\beta}_1}, \ldots, \frac{\mathrm{sign}(\hat{\beta}_q)}{\tilde{\beta}_q}\right)' \\
&= s_{1n}(\beta_0) - \hat{I}_{11}(\beta^*)(\hat{\beta}_{1n} - \beta_{10}) - n\lambda_n \left(\frac{\mathrm{sign}(\beta_{10})}{\tilde{\beta}_1}, \ldots, \frac{\mathrm{sign}(\beta_{q0})}{\tilde{\beta}_q}\right)',
\end{aligned}$$

where β^* is between $\hat{\beta}_n$ and β_0. The last equation is implied by $\text{sign}(\hat{\beta}_{jn}) = \text{sign}(\beta_{j0})$ when n is large, since $\hat{\beta}_n$ is a root-n consistent estimator of β_0. Using Theorem 3.2 of Anderson & Gill (1982), we can prove that $s_{1n}(\beta_0)/\sqrt{n} \to N\{0, I_1(\beta_{10})\}$ in distribution and $\hat{I}_{11}(\beta^*)/n \to I_1(\beta_{10})$ in probability as $n \to \infty$. Furthermore, if $\sqrt{n}\lambda_n \to \lambda_0$, a nonnegative constant, we have

$$\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) = I_1^{-1}(\beta_{10}) \left\{ \frac{1}{\sqrt{n}} s_{1n}(\beta_0) - \lambda_0 b_1 \right\} + o_p(1),$$

with $b_1 = (\text{sign}(\beta_{10})/|\beta_{10}|, \ldots, \text{sign}(\beta_{q0})/|\beta_{q0}|)^{\mathrm{T}}$, since $\tilde{\beta}_j \to \beta_{j0} \neq 0$ for $1 \leq j \leq q$. Then, by Slutsky's Theorem, $\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) \to N\{-\lambda_0 I_1^{-1}(\beta_{10}) b_1, I_1^{-1}(\beta_{10})\}$ in distribution as $n \to \infty$. In particular, if $\sqrt{n}\lambda_n \to 0$, we have

$$\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) \to N\{0, I_1^{-1}(\beta_{10})\}$$

in distribution as $n \to \infty$.

Appendix 2

Modified shooting algorithm for adaptive Lasso

We present the modified shooting algorithm for minimising

$$\sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda \sum_{j=1}^{d} |\beta_j| / |\tilde{\beta}_j|.$$

Define $F(\beta) = \sum_{i=1}^{n} (y_i - \beta' x_i)^2$, $\dot{F}_j(\beta) = \partial F(\beta)/\partial \beta_j$, $j = 1, \ldots, d$, and write β as $(\beta_j, (\beta^{-j})^{\mathrm{T}})^{\mathrm{T}}$, where β^{-j} is the $(d-1)$-dimensional vector consisting of all β_i's other than β_j. The modified shooting algorithm is then initialised by taking $\hat{\beta}_0 = \tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)^{\mathrm{T}}$ and letting $\lambda_j = \lambda/|\tilde{\beta}_j|$ for $j = 1, \ldots, d$. The mth iterative stage involves, for each $j = 1, \ldots, p$,

letting $F_0 = \dot{F}_j(0, \hat{\beta}_{m-1}^{-j})$ and setting

$$\hat{\beta}_j = \begin{cases} \frac{\lambda_j - F_0}{2(x^j)'x^j} & \text{if} \quad F_0 > \lambda_j \\ \frac{-\lambda_j - F_0}{2(x^j)'x^j} & \text{if} \quad F_0 < -\lambda_j \\ 0 & \text{if} \quad |F_0| \leq \lambda_j, \end{cases}$$

where $x^j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$. A new estimator $\hat{\beta}_m = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^{\mathrm{T}}$ is formed after updating all the $\hat{\beta}_j$'s. This is repeated until $\hat{\beta}_m$ converges.

References

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100–20.

Antoniadis, A. & Fan, J. (2001). Regularization of wavelet approximations. *J. Am. Statist. Assoc.* **96**, 939–63.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.*, B, Methodological **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403.

Dickson, E., Grambsch, P., Fleming, T., Fisher, L. & Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* **10**, 1–7.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Donoho, D. L. & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–51.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

Faraggi, D. & Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54**, 1475–85.

Fu, W. (1998). Penalized regression: the bridge versus the lasso. *J. of Comp. Graph. Statist.* **7**, 397–416.

Ibrahim, J. G., Chen, M.-H. & MacEachern, S. N. (1999). Bayesian variable selection for proportional hazards models. *Can. J. Statist.* **27**, 701–17.

Sauerbrei, W. & Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the cox regression model. *Statist. Med.* **11**, 2093–109.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox model.* New York: Springer-Verlag Inc.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.,* B **58**, 267–88.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9**, 93–108.

Table 1: Simulation study. Mean squared error and model selection results for Model 1. The numbers in parentheses are standard errors.

		25% Censored			40% Censored		
n	Method	Corr. (6)	Incorr. (0)	MSE	Corr. (6)	Incorr. (0)	MSE
100	MLE	0	0	0.25 (0.02)	0	0	0.31 (0.03)
	LASSO	4.87	0.00	0.19 (0.01)	4.67	0.00	0.20 (0.01)
	SCAD	5.38	0.01	0.20 (0.02)	5.47	0.08	0.25 (0.03)
	ALASSO	5.73	0.01	0.16 (0.01)	5.63	0.04	0.17 (0.01)
200	MLE	0.00	0.00	0.10 (0.01)	0.00	0.00	0.11 (0.01)
	LASSO	4.94	0.00	0.10 (0.01)	4.69	0.00	0.11 (0.01)
	SCAD	5.68	0.00	0.07 (0.01)	5.55	0.00	0.07 (0.01)
	ALASSO	5.91	0.00	0.07 (0.01)	5.86	0.00	0.07 (0.01)
300	MLE	0	0	0.06 (0.00)	0	0	0.08 (0.00)
	LASSO	4.82	0.00	0.06 (0.01)	4.72	0.00	0.07 (0.01)
	SCAD	5.79	0.00	0.04 (0.00)	5.79	0.00	0.05 (0.01)
	ALASSO	5.91	0.00	0.04 (0.00)	5.85	0.00	0.04 (0.00)

MSE, mean squared error; MLE, maximum partial likelihood; LASSO, Lasso method; SCAD, smoothly clipped absolute deviation method; ALASSO, adaptive Lasso method; Corr., average number of correct zeros; Incorr., average number of incorrect zeros.

Table 2: Simulation study. Frequency of variable selection for Model 1 and 25% censoring.

n	Method	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9
100	LASSO	100	100	29	15	16	100	16	15	12
	SCAD	99	100	11	15	9	100	7	12	8
	ALASSO	99	100	5	7	3	100	2	4	6
200	LASSO	100	100	14	16	24	100	19	17	16
	SCAD	100	100	4	4	6	100	5	7	6
	ALASSO	100	100	1	1	2	100	2	3	0
300	LASSO	100	100	22	23	21	100	18	12	22
	SCAD	100	100	4	4	4	100	0	2	7
	ALASSO	100	100	2	1	2	100	0	2	2

LASSO, Lasso method; SCAD, smoothly clipped absolute deviation method; ALASSO, adaptive Lasso method

Table 3: Simulation study. Estimated and actual standard errors for estimates for Model 1 and 25% censoring. The numbers in parentheses are standard deviations of estimated standard errors.

		$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_6$	
n	Method	SE	ASE	SE	ASE	SE	ASE
100	MLE	0.192	0.169 (0.019)	0.214	0.186 (0.020)	0.216	0.185 (0.021)
	LASSO	0.154	0.105 (0.017)	0.153	0.104 (0.016)	0.158	0.096 (0.016)
	SCAD	0.196	0.153 (0.028)	0.224	0.150 (0.024)	0.155	0.134 (0.016)
	ALASSO	0.206	0.155 (0.023)	0.201	0.155 (0.016)	0.175	0.138 (0.015)
200	MLE	0.119	0.111 (0.008)	0.121	0.121 (0.009)	0.148	0.123 (0.009)
	LASSO	0.109	0.081 (0.007)	0.096	0.081 (0.006)	0.116	0.075 (0.007)
	SCAD	0.119	0.106 (0.009)	0.094	0.105 (0.010)	0.120	0.095 (0.008)
	ALASSO	0.128	0.107 (0.007)	0.116	0.106 (0.007)	0.131	0.096 (0.007)
300	MLE	0.095	0.088 (0.006)	0.108	0.098 (0.005)	0.102	0.098 (0.006)
	LASSO	0.087	0.069 (0.005)	0.092	0.070 (0.004)	0.085	0.065 (0.005)
	SCAD	0.091	0.086 (0.005)	0.093	0.087 (0.005)	0.068	0.078 (0.006)
	ALASSO	0.100	0.086 (0.005)	0.107	0.087 (0.005)	0.094	0.078 (0.005)

SE, sample standard error; ASE, the average of estimated standard error; MLE, maximum partial likelihood; LASSO, Lasso method; SCAD, smoothly clipped absolute deviation method; ALASSO, adaptive Lasso method.

Table 4: Simulation study. Mean squared error and model selection results for Model 2. The numbers in parentheses are standard errors.

		25% Censored			40% Censored		
n	Method	Corr. (6)	Incorr. (0)	MSE	Corr. (6)	Incorr. (0)	MSE
200	MLE	0.00	0.00	0.08 (0.00)	0.00	0.00	0.10 (0.00)
	LASSO	5.43	0.38	0.08 (0.00)	5.32	0.43	0.08 (0.01)
	SCAD	5.64	0.61	0.13 (0.00)	5.64	0.82	0.15 (0.01)
	ALASSO	5.82	0.75	0.08 (0.00)	5.80	0.73	0.09 (0.01)
300	MLE	0.00	0.00	0.05 (0.00)	0.00	0.00	0.06 (0.00)
	LASSO	5.35	0.15	0.05 (0.00)	5.25	0.19	0.05 (0.00)
	SCAD	5.58	0.43	0.10 (0.01)	5.46	0.42	0.08 (0.00)
	ALASSO	5.86	0.45	0.06 (0.00)	5.73	0.44	0.06 (0.00)

MSE, mean squared error; MLE, maximum partial likelihood; LASSO, Lasso method; SCAD, smoothly clipped absolute deviation method; ALASSO, adaptive Lasso method; Corr., average number of correct zeros; Incorr., average number of incorrect zeros.

Table 5: Primary biliary cirrhosis data. Estimated coefficients and standard errors given in parentheses.

Covariate	MLE	LASSO	SCAD	ALASSO
trt	-0.124 (0.215)	0 (-)	0 (-)	0 (-)
age	0.029 (0.012)	0.033 (0.004)	0.033 (0.009)	0.019 (0.010)
sex	-0.366 (0.311)	0 (-)	0 (-)	0 (-)
asc	0.088 (0.387)	0.107 (0.052)	0 (-)	0 (-)
hep	0.026 (0.251)	0 (-)	0 (-)	0 (-)
spid	0.101 (0.244)	0 (-)	0 (-)	0 (-)
oed	1.011 (0.394)	0.648 (0.177)	1.250 (0.341)	0.671 (0.377)
bil	0.080 (0.025)	0.084 (0.013)	0.065 (0.018)	0.095 (0.020)
chol	0.001 (0.000)	0 (-)	0.001 (0.000)	0 (-)
alb	-0.742 (0.308)	-0.548 (0.133)	-0.684 (0.274)	-0.612 (0.280)
cop	0.003 (0.001)	0.003 (0.001)	0.003 (0.001)	0.002 (0.001)
alk	0.000 (0.000)	0 (-)	0 (-)	0 (-)
sgot	0.004 (0.002)	0.001 (0.000)	0.003 (0.002)	0.001 (0.000)
trig	-0.001 (0.001)	0 (-)	0 (-)	0 (-)
plat	0.001 (0.001)	0 (-)	0 (-)	0 (-)
prot	0.233 (0.106)	0.125 (0.040)	0 (-)	0.103 (0.108)
stage	0.455 (0.175)	0.265 (0.064)	0.519 (0.152)	0.367 (0.142)

MLE, maximum partial likelihood; LASSO, Lasso method; SCAD, smoothly clipped absolute deviation method; ALASSO, adaptive Lasso method.