Biometrika Advance Access published February 8, 2008

Biometrika (2008), *pp*. 1–7
doi: 10.1093/biomet/asm083

Printed in Great Britain

A note on path-based variable selection in the penalized proportional hazards model

BY HUI ZOU

School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, Minnesota 55455, U.S.A.

hzou@stat.umn.edu

SUMMARY

We propose an efficient and adaptive shrinkage method for variable selection in the Cox model. The method constructs a piecewise-linear regularization path connecting the maximum partial likelihood estimator and the origin. Then a model is selected along the path. We show that the constructed path is adaptive in the sense that, with a proper choice of regularization parameter, the fitted model works as well as if the true underlying submodel were given in advance. A modified algorithm of the least-angle-regression type efficiently computes the entire regularization path of the new estimator. Furthermore, we show that, with a proper choice of shrinkage parameter, the method is consistent in variable selection and efficient in estimation. Simulation shows that the new method tends to outperform the lasso and the smoothly-clipped-absolute-deviation estimators with moderate samples. We apply the methodology to data concerning nursing homes.

Some key words: Adaptive path; Lasso; Oracle property; Penalized partial likelihood; Smoothly-clipped-absolute deviation penalty; Variable selection.

1. INTRODUCTION

We consider the problem of variable selection in Cox's proportional hazards model. The traditional best-subset selection method works poorly, because it is unstable and is unable to handle a large number of covariates. Tibshirani (1997) applied the lasso method (Tibshirani, 1996) to perform automatic variable selection in the Cox model. The lasso outperforms the best-subset selection by reducing the estimation variability via continuous ℓ_1 shrinkage. In the same spirit, Fan & Li (2002) proposed another penalization method for variable selection in the Cox model, producing a penalized partial likelihood estimator using the smoothly-clipped-absolute-deviation penalty (Fan & Li, 2001).

A common drawback in both lasso and smoothly-clipped-absolute-deviation estimators is that they are not computationally efficient. Tibshirani (1997) and Fan & Li (2002) proposed iterative algorithms for their methods. This is not very satisfactory since we now have very efficient algorithms for solving the lasso in machine learning problems. In linear regression models, Efron et al. (2004) proposed the least-angle-regression algorithm to compute the entire lasso solution path in the same order of computations as a single least-squares fit. Zhu et al. (2003) extended the least-angle-regression algorithm to compute the 1-norm support vector machine. Zou & Hastie (2005) modified the least-angle-regression algorithm to calculate the solution path of the elastic net. Unfortunately, in the Cox model, the solution paths of the lasso and the smoothly-clipped-absolute-deviation estimator are not piecewise linear (Rosset & Zhu, 2007), and thus an algorithm of the least-angle-regression type is not directly applicable.

A question naturally arises: can we have a path-based efficient method for variable selection in the Cox model which also enjoys the optimal asymptotic properties? We achieve this with what we call an efficient-adaptive-shrinkage method to construct an adaptive sparse shrinkage path for variable selection in the Cox model.

2. BACKGROUND

Consider the standard survival-data set-up. For the sake of simplicity, we assume that there are no tied failure times and the censoring is noninformative. Tied failure times can be handled by tie-handling methods. The data are of the form $(y_i, x_i, d_i)_{i=1}^n$. Let y_i denote the survival time and let $1-d_i$ be the censoring indicator: $d_i = 1$ indicates no censoring and $d_i = 0$ indicates right censoring. Let $x_i = (x_{i1}, \ldots, x_{ip})$ represent the vector of predictors for the ith individual. Denote the distinct failure times by $t_1 < t_2 < \cdots < t_m$. Let R_r be the risk set at time $t_r - 0$. Cox's proportional hazards model (Cox, 1972, 1975) assumes that the hazard function at time t given predictor values x can be written as

$$h(t \mid x) = h_0(t) \exp\left(\beta_0^{\mathrm{T}} x\right), \tag{1}$$

where $h_0(t)$ is the baseline hazard function and β_0 is the true value of the regression coefficients. Cox's estimator of β_0, denoted by $\hat{\beta}^{\mathrm{C}}$, maximizes the partial likelihood

$$L(\beta) = \prod_{r=1}^{m} \frac{\exp\left(\beta^{\mathrm{T}} x_{j_r}\right)}{\sum_{j \in R_r} \exp\left(\beta^{\mathrm{T}} x_j\right)}, \tag{2}$$

where j_r is the index of the failure at time t_r.

Denote the negative log-partial-likelihood by $\ell(\beta) = -\log L(\beta)$. The lasso (Tibshirani, 1997) estimator $\hat{\beta}^{\mathrm{L}}$ satisfies

$$\hat{\beta}^{\mathrm{L}} = \arg\min_\beta \left\{ \ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \qquad \lambda \geqslant 0,$$

where $\sum_{j=1}^{p} |\beta_j|$ is called the lasso penalty (Tibshirani, 1996). As λ increases, the elements of $\hat{\beta}^{\mathrm{L}}$ are continuously shrunk towards zero. Some elements will be shrunk to zero, and thus are automatically deleted. Calculation of the lasso estimator in the Cox model requires expensive iteratively reweighted least-squares algorithms (Tibshirani, 1997) for each value of the regularization parameter λ.

Fan & Li (2002)'s smoothly-clipped-absolute-deviation estimator $\hat{\beta}^{\mathrm{S}}$ satisfies

$$\hat{\beta}^{\mathrm{S}} = \arg\min_\beta \left\{ \ell(\beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\},$$

where $p_\lambda(\cdot)$ is defined by

$$p'_\lambda(\theta) = \lambda I(\theta \leqslant \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)} I(\theta > \lambda),$$

for some $a > 2$ and $\theta > 0$. Fan & Li (2001) suggested using $a = 3{\cdot}7$. Note that $p'_\lambda(0+) > 0$, which is responsible for the sparsity of $\hat{\beta}^{\mathrm{S}}$. Hence it performs simultaneous shrinkage and selection in a way similar to the lasso. In addition, the estimator $\hat{\beta}^{\mathrm{S}}$ is shown to possess an asymptotic oracle property: it works as well as if the true submodel were known. Cai et al. (2005) later extended the methodology and theory to the penalized pseudo-partial likelihood model. The estimator $\hat{\beta}^{\mathrm{S}}$ is calculated by an iterative local quadratic approximation algorithm (Fan & Li, 2001, 2002; Hunter & Li, 2005) for each value of λ.

3. THE EFFICIENT ADAPTIVE SHRINKAGE METHOD

3·1. *Main idea*

We write $\ell_r(\beta) = -\beta^{\mathrm{T}} x_{j_r} + \log\left\{\sum_{j \in R_r} \exp\left(\beta^{\mathrm{T}} x_j\right)\right\}$, so that $\ell(\beta) = \sum_{r=1}^{m} \ell_r(\beta)$. Let $\hat{I}(\beta) = \sum_{r=1}^{m} \nabla \ell_r(\beta)$ where ∇ indicates the Hessian operator. We define the efficient-adaptive-shrinkage estimator as follows:

$$\hat{\beta}^{\mathrm{E}} = \arg\min_\beta \left\{ \frac{1}{2}(\beta - \hat{\beta}^{\mathrm{C}})^{\mathrm{T}} \hat{I}(\hat{\beta}^{\mathrm{C}})(\beta - \hat{\beta}^{\mathrm{C}}) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\mathrm{C}}|^\gamma} \right\}, \tag{3}$$

in which γ is a positive constant and λ is the regularization parameter, and we have used the adaptive lasso penalty proposed by Zou (2006). Why not use the exact adaptive lasso estimator $\hat{\beta}^{\mathrm{A}}$, where

$$\hat{\beta}^{\mathrm{A}} = \arg\min_{\beta} \left\{ \ell(\beta) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\mathrm{C}}|^{\gamma}} \right\}? \tag{4}$$

According to Zou (2006), the above procedure will have the oracle properties of the smoothly-clipped-absolute-deviation estimator. However, calculating $\hat{\beta}^{\mathrm{A}}$ is not easier than calculating $\hat{\beta}^{\mathrm{L}}$. The idea behind the efficient-adaptive-shrinkage estimator is specifically to speed up the computations in (4) without losing any efficiency: the quadratic component is an efficient quadratic approximation of the partial likelihood, and this permits efficient computations without losing information about β contained in the partial likelihood; see Theorem 1 in § 3·3. Cox (1975) was the first to suggest the idea of approximating the partial likelihood.

3·2. *Computing the path*

Since the efficient-adaptive-shrinkage criterion is the sum of a quadratic loss and a weighted ℓ_1 penalty, the solution path is piecewise linear as a function of λ. The following algorithm of the least-angle-regression type obtains the entire efficient-adaptive-shrinkage solution path.

Step 1. Calculate $\hat{\beta}^{\mathrm{C}}$ and $\hat{I}(\hat{\beta}^{\mathrm{C}})$.

Step 2. Obtain the decomposition $\hat{I}(\hat{\beta}^{\mathrm{C}}) = V^{\mathrm{T}}V$ and write $V = [v_1|\cdots|v_p]$.

Step 3. Compute $u = V\hat{\beta}^{\mathrm{C}}$ and carry out the transformation $v_j = v_j|\hat{\beta}_j^{\mathrm{C}}|^{\gamma}$, $j = 1, 2, \ldots, p$.

Step 4. Use least-angle-regression to solve the following lasso-type problem:

$$\hat{\beta}^{*} = \arg\min_{\beta} \frac{1}{2}\|u - V\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Step 5. Compute $\hat{\beta}_j^{\mathrm{E}} = \hat{\beta}_j^{*}|\hat{\beta}_j^{\mathrm{C}}|^{\gamma}$, $j = 1, 2, \ldots, p$.

After $\hat{\beta}^{\mathrm{C}}$ is obtained, calculation of the efficient-adaptive-shrinkage solution path requires the same order of computations as a single least-squares fit. Therefore, the total computational cost is about that of obtaining $\hat{\beta}^{\mathrm{C}}$.

3·3. *Asymptotic theory*

It is well known that, under some regularity conditions, $\hat{\beta}^{\mathrm{C}}$ asymptotically has a normal distribution (Andersen & Gill, 1982). We assume the same conditions in our analysis.

Suppose $\beta_0 = (\beta_{1,0}^{\mathrm{T}}, \beta_{2,0}^{\mathrm{T}})^{\mathrm{T}}$ and denote by $\mathcal{A}$ the indices of nonzero coefficients. Assume without loss of generality that $\beta_{2,0} = 0$. Hence we write $\mathcal{A}$ as $(1, 2, \ldots, s)$ and s ($s < p$) is the size of the underlying submodel.

THEOREM 1. *Let* $\hat{\mathcal{A}} = \{j : \hat{\beta}_j^{\mathrm{E}} \neq 0,\ j = 1, 2, \ldots, p\}$. *Suppose that* $\lambda_n n^{-1/2} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$. *Then, as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{1,0}^{\mathrm{E}} - \beta_{1,0}) \to N\{0, I^{-1}(\beta_{1,0})\}, \tag{5}$$

in distribution, where $I(\beta_{1,0})$ *is the leading* $s \times s$ *submatrix of* $I(\beta_0)$, *and*

$$\sqrt{n}\hat{\beta}_{2,0}^{\mathrm{E}} \to 0, \tag{6}$$

in probability. Furthermore, $\lim_n \mathrm{pr}(\hat{\mathcal{A}} = \mathcal{A}) = 1$, *confirming consistency in variable selection.*

The proof of Theorem 1 is given in the Appendix, which is in line with the results in Zou (2006). With a proper choice of regularization parameter λ_n, the efficient-adaptive-shrinkage method consistently identifies the true submodel, and it is asymptotically efficient in estimating the nonzero coefficients. Thus the efficient-adaptive-shrinkage method works as well as if the true underlying model were given in advance, and in theoretical terms it performs as well as the smoothly-clipped-absolute-deviation estimator.

3·4. *Tuning*

In this work we use AIC to select automatically the tuning parameter of the efficient-adaptive-shrinkage with finite samples. The AIC score is defined by

$$\text{AIC} = 2\ell(\hat{\beta}^{\text{E}}) + 2|\hat{\mathcal{A}}|, \tag{7}$$

where $|\hat{\mathcal{A}}|$ is the cardinality of $\hat{\mathcal{A}}$. It has been justified in Efron et al. (2004) and Zou et al. (2007) that the effective dimension of the ℓ_1-penalized model is well approximated by the number of nonzero coefficients.

We used λ as the regularization parameter in the definition of the efficient-adaptive-shrinkage estimator. However, given the form of the algorithm, we can also use the weighted ℓ_1-norm s,

$$s = \frac{\sum_{j=1}^{p} \left|\hat{\beta}_j^{\text{E}}\right| \left|\hat{\beta}_j^{\text{C}}\right|^{-\gamma}}{\sum_{j=1}^{p} \left|\hat{\beta}_j^{\text{C}}\right| \left|\hat{\beta}_j^{\text{C}}\right|^{-\gamma}},$$

as the tuning parameter. Often s could be more convenient, for its value is always between zero and one. The solution as a function of s is also piecewise linear. The selection of the tuning parameter is done by

$$\hat{\lambda} = \arg\min_{\lambda} \text{AIC}(\lambda) \quad \text{or} \quad \hat{s} = \arg\min_{s} \text{AIC}(s). \tag{8}$$

The AIC curve is easily computed after we have obtained the entire efficient-adaptive-shrinkage solution path so that the tuning step is effortless. In principle, γ is a second tuning parameter. In Zou (2006) two-dimensional crossvalidation was used to select a pair of (λ, γ) for optimal prediction. Since any positive γ leads to the asymptotically optimal estimator in Theorem 1, we can just use a fixed γ, such as $\gamma = 1$, for simplicity. According to our experience, $\gamma = 1$ works reasonably well in prediction with samples of moderate size.

4. NUMERICAL RESULTS

4·1. *A synthetic model*

We adopt the simulation set-up in Fan & Li (2002). We simulated 1000 datasets consisting of 80 and 120 observations from the exponential hazard model with

$$h(t \mid x) = \exp\left(\beta_0^{\text{T}} x\right),$$

where $\beta_0 = (0{\cdot}8, 0, 0, 1, 0, 0, 0{\cdot}6, 0)^{\text{T}}$. The x_i's were marginally standard normal and the correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0{\cdot}5$. The censoring time was simulated from an exponential distribution with mean $U \exp(\beta^{\text{T}} x)$, where $U \sim \text{Un}(1, 3)$. An oracle who knows the underlying model but not the actual values of the coefficients will apply partial likelihood to the data to estimate the coefficients of the variables x_1, x_4 and x_7. This will be called the oracle procedure.

Suppose a method, μ, gives an estimator $\hat{\beta}_\mu$. Fan & Li (2002) show that the model error, defined as

$$\text{ME}(\hat{\beta}_\mu) = E\left\{ \exp\left(-\hat{\beta}_\mu^{\text{T}} X\right) - \exp\left(-\beta_0^{\text{T}} X\right) \right\}^2,$$

is appropriate for measuring the goodness of fit of μ. For each method, we computed its relative model error compared to that of the oracle estimator.

The median relative model errors based on the 1000 simulated datasets are summarized in Table 1. We also report the variable selection results in Table 1. The smoothly-clipped-absolute-deviation estimator, the efficient-adaptive-shrinkage estimator and the adaptive lasso all outperform the lasso in terms of model error. Not surprisingly, the adaptive lasso is comparable to the smoothly-clipped-absolute-deviation estimator. Note that $\hat{\beta}^{\text{E}}$ does better than $\hat{\beta}^{\text{A}}$. This indicates that efficient quadratic approximation of the log-partial likelihood can improve the prediction accuracy with samples of moderate sizes. 2

4·2. *Nursing home data*

The nursing home data, analyzed by Morris et al. (1994), came from a study sponsored by the National Centre for Health Services in 1980–82. The data were used to investigate the impact of certain financial

Table 1. *Simulation results for Cox's proportional hazards model. Values of the median relative model error,* MRME*, the average number of correctly selected variables,* C*, and the average number of incorrectly selected variables,* IC

	$n = 80$			$n = 120$		
Estimator	MRME	C	IC	MRME	C	IC
$\hat{\beta}^{\mathrm{L}}$	1·56	2·97	1·59	2·24	2·99	1·65
$\hat{\beta}^{\mathrm{S}}$	1·11	2·91	0·4	1·25	2·96	0·22
$\hat{\beta}^{\mathrm{A}}$	1·15	2·81	0·1	1·22	2·92	0·11
$\hat{\beta}^{\mathrm{E}}$	1·07	2·96	0·99	1·11	2·99	0·84

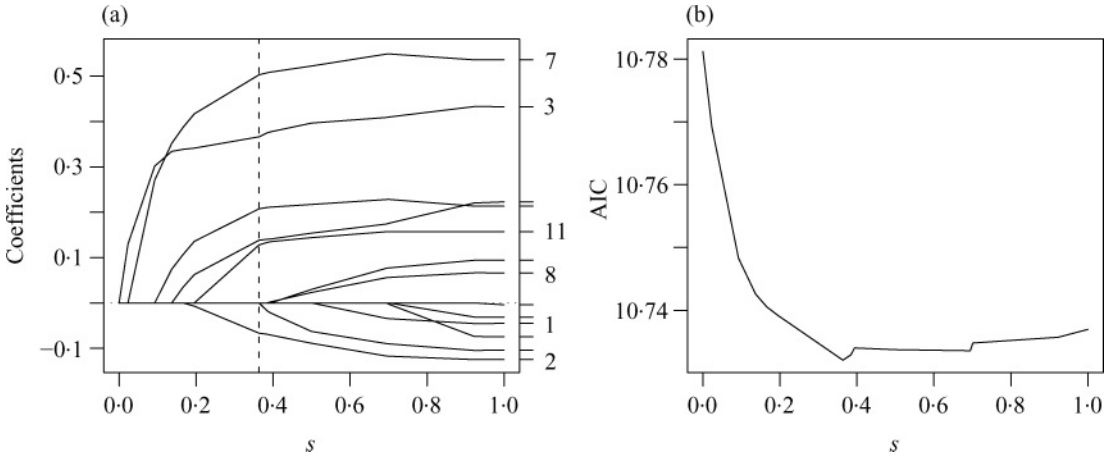

Fig. 1. Nursing home data example. Panel (a) displays the estimated coefficients as a function of s. The vertical dashed line indicates the optimal efficient-adaptive-shrinkage fit in which six variables are selected. Here $\hat{s}$ is chosen based on the AIC curve shown in panel (b).

incentives on the nursing home care of Medicaid patients. The original dataset consists of 1601 observations and five predictors, namely treatment, age, gender, marital status and health status. The number of days in the nursing home is taken as the response variable. Morris et al. (1994) fitted the Cox model to this dataset without including any possible interactions, but, as pointed out by Fan & Li (2002), one can reduce possible modelling biases by exploring interactions. Following this suggestion, we considered using both main effects and first-order interactions in the Cox model. Figure 1(a) shows the estimated coefficients from the efficient-adaptive-shrinkage fit as a function of s. The AIC score is also plotted against s in Fig. 1(b), and the minimum AIC occurs at $s = 0{\cdot}364$. Six variables have nonzero coefficients. Table 2 shows the estimated coefficients. The efficient-adaptive-shrinkage method and the lasso selected the same set of covariates, while the smoothly-clipped-absolute-deviation model included two more, namely the treatment–gender interaction and the age–marital-status interaction. The treatment variable and its interactions with others were deleted from the efficient-adaptive-shrinkage and the lasso fits. Table 2 shows that the age variable is important in all three models considering interactions. Its positive coefficient indicates that elderly patients are more likely to stay at a nursing home. However, age is not statistically significant in the analysis without interactions, as pointed out by Morris et al. (1994). The age variable shows its influence perhaps through its interaction with gender. Note that the gender variable has a very strong impact on the length of stay at a nursing home, and the age–gender interaction is also strong. In this example, the amount of shrinkage imposed on the nonzero coefficients appears to be largest in the lasso, while the smoothly-clipped-absolute-deviation method slightly shrinks the coefficients. We see that the efficient-adaptive-shrinkage method behaves like a compromise between the lasso and the smoothly-clipped-absolute-deviation method. It imposes appreciable shrinkage but controls the shrinkage bias.

Table 2. *Nursing home data. Estimated coefficients of the efficient-adaptive-shrinkage estimates, $\hat{\beta}^{\mathrm{E}}$, the smoothly-clipped-absolute-deviation estimates, $\hat{\beta}^{\mathrm{S}}$, and the lasso estimates, $\hat{\beta}^{\mathrm{L}}$. The notation 'A ∗ B' indicates the interaction between A and B. Here* HS3, HS4 *and* HS5 *are health services variables*

	$\hat{\beta}^{\mathrm{C}}$	$\hat{\beta}^{\mathrm{E}}$	$\hat{\beta}^{\mathrm{S}}$	$\hat{\beta}^{\mathrm{L}}$
Treatment	−0·045	0	0	0
Age	−0·125	−0·067	−0·094	−0·050
Gender	0·433	0·366	0·441	0·307
Marital status	0·223	0·138	0·185	0·085
HS3	−0·031	0	0	0
HS4	0·213	0·208	0·231	0·143
HS5	0·536	0·503	0·543	0·347
Treatment ∗ Age	0·066	0	0	0
Treatment ∗ Gender	−0·104	0	−0·159	0
Treatment ∗ Marital status	−0·004	0	0	0
Age ∗ Gender	0·157	0·128	0·157	0·068
Age ∗ Marital status	0·094	0	0·085	0
Gender ∗ Marital status	−0·075	0	0	0

ACKNOWLEDGEMENT

The author sincerely thanks the editor, an associate editor and referees for their helpful comments which greatly improved the presentation of the paper.

APPENDIX

Proof of Theorem 1. We first prove the asymptotic normality of $\hat{\beta}^{\mathrm{E}}_{1,0}$. Define

$$Z_n(u) = \frac{1}{2}\left(\beta_0 + \frac{u}{\sqrt{n}} - \hat{\beta}^{\mathrm{C}}\right)^{\mathrm{T}} \hat{I}(\hat{\beta}^{\mathrm{C}})\left(\beta_0 + \frac{u}{\sqrt{n}} - \hat{\beta}^{\mathrm{C}}\right) + \lambda_n \sum_{j=1}^{p} \frac{\left|\beta_{0,j} + \frac{u_j}{\sqrt{n}}\right|}{\left|\hat{\beta}^{\mathrm{C}}_j\right|^{\gamma}},$$

and let $\hat{u}_n = \arg\min_u\{Z_n(u) - Z_n(0)\}$. Then it is easy to see that $\hat{\beta}^{\mathrm{E}} = \beta_0 + \hat{u}_n/\sqrt{n}$, and $\hat{u}_n = \sqrt{n}(\hat{\beta}^{\mathrm{E}} - \beta_0)$. For any p-dimensional vector u we write $u = \left(u_1^{\mathrm{T}}, u_2^{\mathrm{T}}\right)^{\mathrm{T}}$, where $u_1 = u_{\mathcal{A}}$. Following the analysis in Zou (2006) and using the asymptotic normality of $\hat{\beta}^{\mathrm{C}}$, we can show that, for each fixed u, in distribution,

$$Z_n(u) - Z_n(0) \to Z(u) \equiv \begin{cases} \frac{1}{2}u_1^{\mathrm{T}} I(\beta_{1,0})u_1 - u_1^{\mathrm{T}} W_1 & \text{if} \quad u_2 = 0 \\ \infty & \text{otherwise,} \end{cases}$$

where W is a p-dimensional random vector distributed as $N\{0, I(\beta_0)\}$ and $W_1 = W_{\mathcal{A}}$. The unique minimum of $Z(u)$ is $\hat{u}_1 = I^{-1}(\beta_{1,0})W_1$ and $\hat{u}_2 = 0$. Here $Z_n(u) - Z_n(0)$ is a convex function of u. By epiconvergence (Geyer, 1994; Knight & Fu, 2000), we have that, in distribution,

$$\hat{u}_{n,1} \to \hat{u}_1 = I^{-1}(\beta_{1,0})W_1, \quad \hat{u}_{n,2} \to \hat{u}_2 = 0.$$

Finally, $W_1 \sim N\{0, I(\beta_{1,0})\}$, so that (5) and (6) are proven.

For the variable selection consistency result, note that the asymptotic normality result indicates that $\mathrm{pr}(\mathcal{A} \in \hat{\mathcal{A}}) \to 1$. Thus it suffices to show that $\mathrm{pr}(\mathcal{A}^c \cap \hat{\mathcal{A}} \neq \emptyset) \to 0$. As shown in Zou (2006), we only need to show that, for any $j_0 \in \mathcal{A}^c$,

$$\mathrm{pr}\left(\frac{1}{\sqrt{n}}\{\hat{I}(\hat{\beta}^{\mathrm{C}})(\hat{\beta}^{\mathrm{E}} - \hat{\beta}^{\mathrm{C}})\}_{j_0} = \frac{1}{\sqrt{n}}\lambda_n \left|\hat{\beta}^{\mathrm{C}}_{j_0}\right|^{-\gamma}\right) \to 0,$$

which follows because $n^{-1/2}\{\hat{I}(\hat{\beta}^{\mathrm{C}})(\hat{\beta}^{\mathrm{E}} - \hat{\beta}^{\mathrm{C}})\}_{j_0} = O_P(1)$ but

$$\frac{1}{\sqrt{n}}\lambda_n \left|\hat{\beta}^{\mathrm{C}}_{j_0}\right|^{-\gamma} = \frac{\lambda_n n^{(\gamma-1)/2}}{\left|\sqrt{n}\hat{\beta}^{\mathrm{C}}_{j_0}\right|^{\gamma}} \to \infty,$$

in probability. This completes the proof.

References

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

Cai, J., Fan, J., Li, R. & Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303–16.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **74**, 187–220.

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

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407–99.

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.

Geyer, C. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.* **22**, 1993–2010.

Hunter, D. & Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617–42.

Knight, K. & Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.

Morris, C., Norton, E. & Zhou, X.H. (1994). Parametric duration analysis of nursing home usage. In *Case Studies in Biometry,* Ed. N. Lange, L. Ryan, D. Billard, D. Brillinger, L. Conquest and J. Greenhouse, pp. 231–48. New York: Wiley.

Rosset, S. & Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35**, 1012–30.

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

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385–95.

Zhu, J., Rosset, S., Hastie, T. & Tibshirani, R. (2004). 1-norm support vector machines. In *Advances in Neural Information Processing Systems* **16**, Ed. S. Thrun, L. Saul and B. Schölkopf. Cambridge, MA: MIT Press.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

Zou, H. & Hastie, T. (2005). Regression and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173–92.

[*Received May* 2006*, Revised June* 2007]