On the robustness of the adaptive lasso to model misspecification

By W. LU

*Department of Statistics, North Carolina State University, 2311 Stinson Drive, 5212 SAS Hall,
Raleigh, North Carolina 27695-8203, U.S.A.*

lu@stat.ncsu.edu

Y. GOLDBERG AND J. P. FINE

*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill,
North Carolina 27599-7420, U.S.A.*

yair.goldy@gmail.com jfine@email.unc.edu

SUMMARY

Penalization methods have been shown to yield both consistent variable selection and oracle parameter estimation under correct model specification. In this article, we study such methods under model misspecification, where the assumed form of the regression function is incorrect, including generalized linear models for uncensored outcomes and the proportional hazards model for censored responses. Estimation with the adaptive least absolute shrinkage and selection operator, lasso, penalty is proven to achieve sparse estimation of regression coefficients under misspecification. The resulting estimators are selection consistent, asymptotically normal and oracle, where the selection is based on the limiting values of the parameter estimators obtained using the misspecified model without penalization. We further derive conditions under which the penalized estimators from the misspecified model may yield selection consistency under the true model. The robustness is explored numerically via simulation and an application to the Wisconsin Epidemiological Study of Diabetic Retinopathy.

Some key words: Least false parameter; Model misspecification; Oracle property; Penalization; Selection consistency; Shrinkage estimation; Variable selection.

1. INTRODUCTION

Variable selection has attracted much attention recently due to its importance in constructing parsimonious models with superior predictive performance. Traditional variable selection algorithms include forward selection, backward elimination and best subset selection. Model selection criteria such as Mallows's C_p (Mallows, 1973), Akaike's information criterion (Akaike, 1973), and the Bayesian information criterion (Schwarz, 1978) have been heavily utilized in conjunction with such procedures. While these methods are conceptually simple and widely used in practice, the lack of a general theoretical justification across a range of algorithms and the unstable empirical performance of the algorithms have motivated the development of alternative model selection techniques. Penalization methods are available in general regression settings which simultaneously yield sparse models which shrink some coefficients to zero and parameter estimation for the nonzero coefficients. Methods exist for generalized linear models, e.g., the

least absolute shrinkage and selection operator (Tibshirani, 1996), the smoothly clipped absolute deviation (Fan & Li, 2001) and the adaptive lasso (Zou, 2006), and have been extended to the proportional hazards model for censored data (e.g., Tibshirani, 1997; Fan & Li, 2002; Zhang & Lu, 2007). With the proper choice of the tuning parameter, the smoothly clipped absolute deviation and adaptive lasso estimators can be shown to achieve selection consistency and asymptotic normality, with asymptotic variance equalling that of the oracle procedure in which the zero coefficients are known a priori.

A key assumption for the shrinkage methods is that the regression model is correctly specified. To our knowledge, the robustness of the variable selection to misspecification has not been rigorously examined. The conditions yielding selection consistency in the true model are unclear, as is the extent to which penalization procedures may be oracle under misspecification. Our approach builds on earlier work for misspecified models, adapting theoretical results for unpenalized estimation using likelihood (White, 1982) and partial likelihood (Lin & Wei, 1989; Sasieni, 1993) to establish the asymptotic properties of the corresponding penalized estimators from the adaptive lasso. The quantities being estimated, the so-called least false parameters, are defined implicitly as the maximizers of the asymptotic limits of the penalized likelihoods. In §§ 2 and 3, we show that selection consistency may be achieved for the nonzero least false parameters. Moreover, under certain conditions (Li & Duan, 1989; Kosorok et al., 2004), selection consistency for the least false parameters provides selection consistency for the parameters in the true model. The oracle property may be achieved, in the sense that the variances of the penalized estimators equal that, for the oracle procedure, in which the nonzero least false parameters are known a priori.

2. VARIABLE SELECTION FOR MISSPECIFIED GENERALIZED LINEAR MODELS

Let $V_1, \dots, V_n$ be independently and identically distributed random vectors, where $V_i = (X_i, Y_i)$, X_i is a p -dimensional vector of covariates and Y_i is a response variable. Here p is assumed to be fixed. In generalized linear models (McCullagh & Nelder, 1989), it is assumed that the true model has density function $f\{y; g(x'\beta)\}c(x)$, where β is a p -dimensional vector of unknown regression coefficients, $c(x)$ is a p -variate function of X , $f\{y; g(x'\beta)\}$ is a conditional density function of Y given $X = x$ and g is a specified link function. The loglikelihood is $L_n^{(1)}(V_1, \dots, V_n; \beta) = \sum_{i=1}^n \ell(V_i; \beta)$, where $\ell(v; \beta) = \log f\{y; g(x'\beta)\}$. Let $\tilde{\beta}^{(1)}$ denote the maximizer of $L_n^{(1)}$.

To select variables, we adopt the adaptive lasso (Zou, 2006). Specifically, we consider the penalized loglikelihood function

$$Q_n^{(1)}(\beta) \equiv L_n^{(1)}(V_1, \dots, V_n; \beta) - \lambda_n \sum_{j=1}^p \hat{w}_{nj} |\beta_j|,$$

where $\hat{w}_{nj} = |\tilde{\beta}_j^{(1)}|^{-\gamma}$ for some $\gamma > 0$, $\tilde{\beta}_j^{(1)}$ is the j th component of $\tilde{\beta}^{(1)}$ and $\lambda_n > 0$ is a tuning parameter. Let $\hat{\beta}^{(1)}$ denote the maximizer of $Q_n^{(1)}(\beta)$.

Next, we study the theoretical properties of the adaptive lasso estimator $\hat{\beta}^{(1)}$ under model misspecification. We first introduce some notation. Define $\dot{\ell}(v; \beta) = \partial \ell(v; \beta) / \partial \beta$, $\ddot{\ell}(v; \beta) = \partial^2 \ell(v; \beta) / (\partial \beta \partial \beta')$, $A^{(1)}(\beta) = E\{\ddot{\ell}(V; \beta)\}$ and $B^{(1)}(\beta) = E\{\dot{\ell}(V; \beta) \dot{\ell}(V; \beta)'\}$. In addition, following White (1982), we assume the following conditions.

Condition 1. The functions $f\{y; g(x'\beta)\}$ are continuous in β for all $\beta \in \mathcal{B}$, where $\mathcal{B}$ is a compact set.

Condition 2. The functions satisfy (i) $E\{\log c(X)\} < \infty$; (ii) $E\{\log f_o(Y|X)\}$ exists, where f_o is the true conditional density; (iii) $E\{\ell(V; \beta)\}$ has a unique maximizer β^* and β^* is interior to $\mathcal{B}$.

Condition 3. The likelihood function $\ell(v; \beta)$ is twice differentiable with respect to β for all $\beta \in \mathcal{B}$, and the components of $|\ell(v; \beta)|$, $|\dot{\ell}(v; \beta)\dot{\ell}(v; \beta)'|$ and $|\ddot{\ell}(v; \beta)|$ are dominated by integrable functions.

Condition 4. The matrices $A^{(1)} \equiv A^{(1)}(\beta^*)$ are nonsingular and negative definite, and $B^{(1)} \equiv B^{(1)}(\beta^*)$ is nonsingular.

Condition 5. For all $i, j, k = 1, \dots, p$, and all β in some neighbourhood of β^* , $|\partial^3 \ell(v; \beta) / \partial \beta_j \partial \beta_k \partial \beta_l|$ are dominated by integrable functions.

Conditions 1–5 ensure the consistency and asymptotic normality of the unpenalized estimator $\tilde{\beta}^{(1)}$. The quantity β^* is the least false parameter to which $\tilde{\beta}^{(1)}$ converges. This value corresponds to the maximizer of the asymptotic limit of the unpenalized loglikelihood. In general, β^* does not correspond to a well-defined parameter in the true conditional distribution f_o . Additional assumptions, like those described below, are needed to link the least false parameter for the misspecified model to the true regression parameter in f_o .

Without loss of generality, we can write $\beta^* = (\beta_1^{*'}, \beta_2^{*'})'$, where β_1^* is a p_1 -dimensional vector of nonzero parameters and $\beta_2^* = 0$ is a $p_2 = (p - p_1)$ -dimensional vector of zero parameters. Accordingly, we write $\hat{\beta}^{(1)} = (\hat{\beta}_1^{(1)'}, \hat{\beta}_2^{(1)'})'$. In the Appendix, we establish the following theorem.

THEOREM 1. Assume that Conditions 1–5 hold, and $n^{-1/2}\lambda_n \rightarrow 0$ and $n^{(\gamma-1)/2}\lambda_n \rightarrow \infty$ as $n \rightarrow \infty$. Then $\text{pr}(\hat{\beta}_2^{(1)} = 0) \rightarrow 1$ and

$$n^{1/2}(\hat{\beta}_1^{(1)} - \beta_1^*) \rightarrow N\{0, (A_{11}^{(1)})^{-1} B_{11}^{(1)} (A_{11}^{(1)})^{-1}\}$$

in distribution, where $A_{11}^{(1)}$ and $B_{11}^{(1)}$ are the upper left $p_1 \times p_1$ submatrices of $A^{(1)}(\beta^*)$ and $B^{(1)}(\beta^*)$, respectively.

With fixed p , sparseness implies that there are zero components in the least false parameter, although the theory permits $p_2 = 0$, where all components are nonzero. The notion of an oracle estimator is defined in terms of such sparseness. With growing p , sparseness generally refers to the proportion of zero parameters, which typically grows as some function of n . Such sparseness is critical to the consistency of penalized estimation procedures, which break down if there are too many nonzero parameters. The complications that arise with growing p are described in § 5.

For fixed p , Theorem 1 gives that if the least false parameter β^* is sparse, then the adaptive lasso procedure is selection consistent and achieves sparsity with probability going to one asymptotically. Moreover, the estimator of the nonzero parameters in β^* , $\hat{\beta}_1^{(1)}$ is asymptotically normal with variance equal to the oracle estimator, in which $\beta_2^* = 0$ is known a priori. The variance generally has the sandwich form described in White (1982) and may be estimated by replacing the unknown quantities with empirical estimates. If the generalized linear model is correctly specified, then under mild regularity conditions (e.g., White, 1982, Assumption A7) $-A^{(1)}(\beta_0) = B^{(1)}(\beta_0) = I$, where I is the Fisher information matrix of the true model with $\beta^* = \beta_0$, the true regression coefficients. Hence, Theorem 1 generalizes Zou (2006, Theorem 4).

Next, we study the relationship between the true model f_o and β^* when the true model satisfies a generalized linear model but the link function is misspecified when the model is fitted using the

loglikelihood functions. To be specific, we reformulate the posited model using the exponential family with canonical parameter θ , that is, the conditional distribution of Y given $X = x$ is given by $f(y | \theta) = h(y) \exp\{y\theta - \psi(\theta)\}$ with $\theta = \beta'x$. Note that $E(Y | X = x) = \dot{\psi}(\theta)$ in the posited model, where $\dot{\psi}$ is the derivative of ψ . Moreover, assume that the true conditional distribution of Y given $X = x$, f_o , also follows a one-parameter family: $H_{\theta_0}(y)$ with $\theta_0 = \beta'_0 x$, where β_0 is the true regression parameter. Here the family $\{H_\theta\}$ can be arbitrary and unknown.

We make the following additional assumptions (Li & Duan, 1989).

Condition 6. The p -variate density function $c(x)$ of X is nondegenerate in $\mathbb{R}^p$.

Condition 7. The conditional expectation $E(\beta'X | \beta'_0 X = \beta'_0 x)$ exists and is linear in $\beta'_0 x$ for all $\beta \in \mathbb{R}^p$.

COROLLARY 1. Assume that $\psi(\theta)$ is strictly convex, $\mathcal{B}$ is nonempty and convex in $\mathbb{R}^p$, and Conditions 1–5 and 6–7 hold. Then $\beta^* = \alpha^* \beta_0$ for some nonzero scalar α^* .

Corollary 1 is a consequence of Li & Duan (1989, Theorem 2.1). It gives that the zero coefficients in β^* are the same as those in β_0 . Condition 7 requires that the covariate distribution is elliptically symmetric, as occurs under the multivariate normal distribution. Combining this result with Theorem 1 yields that the adaptive lasso estimator achieves sparsity and selection consistency for the true model, i.e., correctly identifies the zero and nonzero parameters in β_0 with probability tending to unity, even under model misspecification.

3. VARIABLE SELECTION FOR THE MISSPECIFIED COX MODEL

Let T_i be the survival time and C_i be the censoring time of subject i . Define $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. The observed data consist of $V_1, \dots, V_n$, where $V_i = (X_i, Y_i, \Delta_i)$ and the covariate vector X_i is defined as before. We assume that T_i is independent of C_i given X_i . The proportional hazards model (Cox, 1972) assumes that the conditional hazard function satisfied $\lambda(t | X) = \lambda(t) \exp(\beta'X)$, where $\lambda(t)$ is an unspecified baseline hazard function and β is a p -dimensional vector of unknown regression coefficients.

Define $S^{(r)}(\beta, t) = n^{-1} \sum_{j=1}^n I(Y_j \geq t) \exp\{\beta'X_j\} X_j^{\otimes r}$ and $s^{(r)}(\beta, t) = E\{S^{(r)}(\beta, t)\}$ for $r = 0, 1, 2$, where for a column vector a , $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa'$. The adaptive lasso estimator based on the penalized log partial likelihood is $\hat{\beta}^{(2)} = \arg\max_{\beta} Q_n^{(2)}(\beta)$, where

$$\begin{aligned} Q_n^{(2)}(\beta) &= \sum_{i=1}^n \Delta_i [\beta'X_i - \log\{S^0(\beta, X_i)\}] - \lambda_n \sum_{j=1}^p \hat{w}_{nj} |\beta_j| \\ &\equiv L_n^{(2)}(V_1, \dots, V_n; \beta) - \lambda_n \sum_{j=1}^p \hat{w}_{nj} |\beta_j|, \end{aligned}$$

where $\hat{w}_{nj} = |\tilde{\beta}_j^{(2)}|^{-\gamma}$. The unpenalized partial likelihood estimator $\tilde{\beta}^{(2)} = (\tilde{\beta}_1^{(2)}, \dots, \tilde{\beta}_p^{(2)})'$ maximizes $L_n^{(2)}$.

Let $\tau > 0$ be a constant such that $\text{pr}(Y \geq \tau) > 0$. Define

$$A^{(2)}(\beta) = \int_0^\tau \left\{ \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \frac{s^{(1)}(\beta, t)^{\otimes 2}}{s^{(0)}(\beta, t)^2} \right\} dP^{(2)}(t), \quad B^{(2)}(\beta) = E\{W(\beta)^{\otimes 2}\},$$

where $P^{(2)}(t) = \text{pr}(Y \leq t, \Delta = 1)$,

$$W(\beta) = \Delta \left\{ X - \frac{s^{(1)}(\beta, Y)}{s^{(0)}(\beta, Y)} \right\} - \int_0^\tau \left\{ X - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} I(Y \geq t) \exp(\beta^{*'} X) \frac{dP^{(2)}(t)}{s^{(0)}(\beta, t)},$$

and β^* is the least false parameter that maximizes $E(L_n^{(2)})$. To account for misspecification of the proportional hazards model, we make the following assumptions (Sasieni, 1993).

Condition 8. The expectation $E\{\exp(X'\beta)\} < \infty$ for all $\beta \in \mathcal{B}$.

Condition 9. There is no pair (α, ϕ) with $\alpha \neq 0 \in \mathbb{R}^p$ and $\phi: \mathbb{R} \mapsto \mathbb{R}$, a monotone decreasing function, such that for $P^{(2)}$ -almost all $t < \tau$, $\alpha'X\Delta = \phi(Y)\Delta$ and $\alpha'XI(Y \geq t) \leq \phi(t)$ almost surely.

Conditions 8–9 ensure the consistency and asymptotic normality of the estimator $\tilde{\beta}^{(2)}$ under model misspecification. In particular, Condition 9 was originally proposed by Sasieni (1993), who studied the proportional hazards model under model misspecification. It ensures that as the sample size gets larger, there exists a unique maximizer of the partial likelihood under model misspecification. The condition is needed because of the potential lack of convexity of the partial likelihood loss function under model misspecification. This differs from simpler M-estimation problems, like those in §2, in which the loss function is guaranteed to be convex. One should recognize that if the proportional hazards model does not hold, then the estimand β^* will depend on the distribution of T given X , the distribution of C given X and the marginal distribution of X . Establishing the relationship between the parameters in β^* and those in the true model is thus more challenging than for the generalized linear model.

Similarly to before, we can write $\beta^* = (\beta_1^{*'}, \beta_2^{*'})'$, where β_1^* is a p_1 -dimensional vector of nonzero parameters and $\beta_2^* = 0$ is a $p_2 = (p - p_1)$ -dimensional vector of zero parameters. Analogously, we write $\hat{\beta}^{(2)} = (\hat{\beta}_1^{(2)'}, \hat{\beta}_2^{(2)'})'$. In the Appendix, we prove the following.

THEOREM 2. Assume that Conditions 8–9 hold, and $n^{-1/2}\lambda_n \rightarrow 0$ and $n^{(\gamma-1)/2}\lambda_n \rightarrow \infty$ as $n \rightarrow \infty$. Then $\text{pr}(\hat{\beta}_2^{(2)} = 0) \rightarrow 1$ and

$$n^{1/2}(\hat{\beta}_1^{(2)} - \beta_1^*) \rightarrow N\{0, (A_{11}^{(2)})^{-1} B_{11}^{(2)} (A_{11}^{(2)})^{-1}\}$$

in distribution, where $A_{11}^{(2)}$ and $B_{11}^{(2)}$ are the upper left $p_1 \times p_1$ submatrices of $A^{(2)}(\beta^*)$ and $B^{(2)}(\beta^*)$, respectively.

Theorem 2 implies that the penalized partial likelihood estimator is sparse with probability one when $p_1 < p$. Moreover, the nonzero parameter estimators are asymptotically normal, with the same variance as the oracle partial likelihood procedure with the zero components in β^* known. A robust variance estimator may be obtained using the plug-in formulas in Lin & Wei (1989).

If the Cox model is correctly specified, then $\beta^* = \beta_0$, and $-A^{(2)}(\beta_0) = B^{(2)}(\beta_0) = I$, the log partial likelihood information matrix of the Cox model. Theorem 2 thus generalizes Zhang & Lu (2007, Theorem 2) for the case with $\gamma = 1$. This follows since Sasieni (1993, Lemmas 2.3 and 7.3–7.5) and Conditions 6–7 imply that conditions of Andersen & Gill (1982) hold. Hence, by Andersen & Gill (1982, Theorem 3.2), $A^{(2)}(\beta_0) = -B^{(2)}(\beta_0)$ and the result follows.

Under a misspecified model, the relationship between β^* and β_0 is quite complicated, depending on the joint distribution of the failure time, the censoring time and the covariate.

We now consider a special case where the true failure time distribution satisfies a one-parameter proportional hazards frailty model (Kosorok et al., 2004). The true conditional hazard function is

$$\lambda(t|X, W) = \lambda_0(t) \exp\{\log(W) + \beta'_0 X\},$$

where $\lambda_0(t)$ is an unspecified base hazard function and W is a positive continuous frailty variable with density $f_W(\cdot)$ that is independent of X and $E(W) = 1$. Let $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ and $L_W(t) = \int_0^\infty e^{-wt} f_W(w) dw$ be the Laplace transform of W .

Additional conditions beyond Conditions 6–7 on the covariate distribution are needed. As in Kosorok et al. (2004), we assume the following conditions.

Condition 10. The censoring variable C is independent of T .

Condition 11. The true density of T given X exists and is bounded over $[0, \tau]$ almost surely, and $\text{pr}(T > \tau|X) > 0$ almost surely.

Condition 12. Both $L_W(ct)$ and $c/L_W(ct)$ are decreasing functions of c for each finite $t > 0$.

Condition 10 strengthens the usual conditional independence assumption. Condition 12 is a regularity condition on the frailty distribution which relaxes those in Kosorok et al. (2004), which have been shown to hold for many common distributions.

COROLLARY 2. Assume that $\mathcal{B}$ is nonempty and convex in $\mathbb{R}^p$, and that Conditions 6–7 and 8–12 hold. Then $\beta^* = \alpha^* \beta_0$ for some nonzero scalar α^* .

Corollary 2 follows immediately from Kosorok et al. (2004, Proposition 5). Theorem 2 and Corollary 2 imply that under model misspecification the adaptive lasso estimator may achieve selection consistency for the regression coefficients β_0 in the true model.

4. NUMERICAL STUDIES

4.1. Simulations

We conducted simulations to study the variable selection performance of the adaptive lasso estimators under various model misspecification scenarios. Specifically, we consider the following scenarios.

Scenario 1. The true model is given by $Y \sim N\{\exp(\beta'_0 X), 0.5^2\}$, but we fit a linear model of Y on X with the identity link function.

Scenario 2. The true model is given by $\text{pr}(Y = 1|X) = \Phi(0.5 + \beta'_0 X)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable, but we fit a logistic regression of Y on X .

Scenario 3. The true model for the failure time is from a class of linear transformation models: $H(T) = -\beta'_0 X + \epsilon$, where $H(\cdot)$ is an unspecified monotone increasing function and the error term ϵ has a known density function that is independent of X . Here we consider three error distributions: e^ϵ follows a Pareto(r) distribution with $r = 1$ and $r = 2$, and a standard log-normal, where Pareto(1) error corresponds to the proportional odds model. The linear transformation model with Pareto(r) distribution is equivalent to the one-dimensional frailty model considered

Table 1. Variable selection results for normal covariates: scenarios 1 and 2

n	Method	CZ	IZ	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Scenario 1													
50	L	5.59	0.12	487	460	148	130	495	94	81	74	90	89
	AL	6.49	0.12	487	471	44	46	483	33	28	34	35	34
	FR _A	5.55	0.03	498	494	116	115	495	99	92	92	102	111
	FR _B	6.47	0.04	496	490	53	39	492	37	27	40	35	35
100	L	5.81	0.01	500	496	130	114	500	94	64	60	60	71
	AL	6.76	0.01	500	498	19	17	497	26	15	18	10	15
	FR _A	5.86	0.01	500	499	98	90	498	89	77	77	67	72
	FR _B	6.73	0.01	500	498	24	21	496	23	16	18	18	15
200	L	5.91	0.00	500	499	119	114	500	88	61	64	41	60
	AL	6.90	0.00	500	499	6	5	500	11	8	7	6	9
	FR _A	5.82	0.00	500	500	84	80	500	86	84	103	67	88
	FR _B	6.83	0.00	500	499	16	11	500	15	14	9	10	10
Scenario 2													
100	L	4.88	0.04	491	492	176	177	497	154	126	140	137	149
	AL	6.38	0.06	492	483	62	46	493	43	39	45	36	39
	FR _A	5.62	0.01	500	499	114	106	497	91	82	96	103	97
	FR _B	6.64	0.02	499	497	32	33	492	20	19	19	26	29
200	L	4.96	0.00	500	500	173	167	500	152	156	118	124	132
	AL	6.76	0.00	500	500	19	13	500	18	20	18	18	13
	FR _A	5.81	0.00	500	500	86	72	500	92	94	73	87	89
	FR _B	6.82	0.00	500	500	17	11	500	12	11	11	16	13
400	L	5.12	0.00	500	500	159	153	500	147	130	92	124	133
	AL	6.84	0.00	500	500	6	11	500	10	17	9	13	16
	FR _A	5.76	0.00	500	500	88	80	500	101	91	78	83	101
	FR _B	6.89	0.00	500	500	6	5	500	10	11	4	7	13

CZ, average number of correct zeros; IZ, average number of incorrect zeros; L, lasso; AL, adaptive lasso; FR_A, forward selection with AIC tuning; FR_B, forward selection with BIC tuning.

in § 3, with the frailty W from the gamma distribution with mean 1 and variance r . For each model, we fit the standard proportional hazards model.

For all scenarios, we consider a ten-dimensional covariate vector $X = (X_1, \dots, X_{10})'$ generated from a multivariate normal distribution with mean 0, variance 1 and correlation between X_j and X_k given by $0.5^{|j-k|}$ for $j \neq k$. The linearity Condition 9 is satisfied for multivariate normal covariates. We set $\beta_0 = (1.0, -0.9, 0, 0, -0.8, 0, 0, 0, 0, 0)'$. In Scenario 3, $H(t) = \log\{0.5(e^{2t} - 1)\}$ and the censoring time C is generated from a uniform distribution on $[0, \tau]$, where τ is chosen to obtain 15% censoring. For Scenario 1, the effect size, defined as the absolute value of the full model estimates divided by the standard errors of the estimates for important variables, ranges from 1.5 to 3.5, while for Scenarios 2 and 3, it ranges from 2 to 5.5. The adaptive lasso estimators with $\gamma = 1$ are computed using the R packages LARS (R Development Core Team, 2012) for Scenario 1 and glmnet for Scenarios 2 and 3. The tuning parameter is chosen by minimizing BIC.

For Scenario 1, we take $n = 50, 100, 200$, while for Scenarios 2 and 3, we take $n = 100, 200, 400$. For each setting, we run 500 simulations. The variable selection performance is summarized by the average number of correct zeros, the average number of incorrect zeros and the frequency that each variable is selected. For comparison, we also include the selection results using three alternative methods, the standard lasso, forward selection with AIC tuning and

Table 2. *Variable selection results for normal covariates: scenario 3*

Selection frequency													
n	Method	CZ	IZ	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Scenario 3: Pareto(1) error													
100	L	5.39	0.39	427	418	131	142	461	106	108	101	103	115
	AL	6.41	0.50	422	401	50	50	426	45	47	26	43	36
	FR _A	5.34	0.06	498	491	116	132	481	126	107	114	121	114
	FR _B	6.50	0.25	469	459	43	51	449	38	30	26	30	32
200	L	5.21	0.02	496	492	146	155	500	142	108	109	117	118
	AL	6.51	0.05	496	482	33	36	496	37	33	37	40	31
	FR _A	5.51	0.01	500	497	101	115	499	102	97	103	115	111
	FR _B	6.64	0.02	499	496	30	21	495	31	18	32	27	22
400	L	5.42	0.00	499	499	149	129	500	127	86	97	95	108
	AL	6.74	0.00	500	500	18	19	500	26	18	12	16	20
	FR _A	5.53	0.00	500	500	97	114	500	114	98	95	103	116
	FR _B	6.80	0.00	500	500	10	19	500	16	10	10	15	18
Scenario 3: Pareto(2) error													
100	L	6.23	1.70	210	175	65	76	266	61	39	46	54	45
	AL	6.58	1.65	227	197	35	29	251	39	26	25	35	22
	FR _A	5.37	0.53	449	397	128	133	390	116	104	111	113	111
	FR _B	6.46	1.14	332	275	47	60	324	44	25	34	32	30
200	L	5.77	0.64	385	351	103	115	443	94	59	74	90	80
	AL	6.51	0.65	395	358	39	41	420	42	24	38	33	30
	FR _A	5.48	0.10	496	482	109	115	474	111	89	109	120	105
	FR _B	6.60	0.45	440	401	51	37	435	25	15	26	24	21
400	L	5.50	0.06	490	479	134	140	500	126	81	87	91	92
	AL	6.59	0.09	487	476	28	36	493	41	26	23	31	20
	FR _A	5.57	0.00	499	500	99	103	500	117	95	105	95	100
	FR _B	6.82	0.04	496	491	15	17	493	15	11	11	9	14
Scenario 3: normal error													
100	L	4.77	0.00	500	500	188	161	500	154	146	153	154	158
	AL	6.57	0.00	500	500	35	28	500	30	34	32	23	32
	FR _A	5.37	0.00	500	500	127	111	500	109	120	115	111	122
	FR _B	6.53	0.00	500	500	40	33	500	30	40	33	26	31
200	L	4.80	0.00	500	500	193	183	500	156	139	143	135	150
	AL	6.80	0.00	500	500	16	18	500	15	10	12	11	20
	FR _A	5.48	0.00	500	500	112	111	500	96	103	111	108	117
	FR _B	6.70	0.00	500	500	25	25	500	21	9	20	26	31
400	L	4.90	0.00	500	500	181	170	500	152	144	130	119	153
	AL	6.82	0.00	500	500	14	15	500	13	15	14	11	9
	FR _A	5.39	0.00	500	500	115	123	500	103	120	114	102	130
	FR _B	6.72	0.00	500	500	21	23	500	18	21	21	14	20

CZ, average number of correct zeros; IZ, average number of incorrect zeros; L, lasso; AL, adaptive lasso; FR_A, forward selection with AIC tuning; FR_B, forward selection with BIC tuning.

forward selection with BIC tuning. The results for Scenarios 1 and 2 are reported in Table 1 and for Scenario 3 in Table 2. As the sample size increases, the variable selection performance of the adaptive lasso estimators quickly improves: the average number of correct zeros is close to 7, the average number of incorrect zeros is close to 0 and the selection frequencies of the three important variables approach 500 while those for the unimportant variables decrease quickly. This confirms the theoretical finding that the adaptive lasso may have the variable selection consistency property under model misspecification when Condition 7 holds. In addition, the adaptive lasso and

forward selection with BIC tuning methods have very comparable selection performance under model misspecification, and are much better than the standard lasso and forward selection with AIC tuning, which tend to select more unimportant variables even when sample size increases. This suggests that, under certain conditions, forward selection with BIC tuning may also enjoy the selection consistency property under model misspecification, which needs to be formally investigated in future research.

Next, we conduct simulations when the linearity Condition 7 is violated, with a skewed distribution for the covariates in X . Conditionally, on a positive random variable α , the j th ($j = 1, \dots, 10$) covariate is generated from an exponential distribution with mean α , where α is generated from a gamma distribution with mean 1 and variance σ^2 . We choose $\sigma^2 = 0$ or 1, where $\sigma^2 = 0$ implies that X_j and X_k are mutually independent for $j \neq k$. We consider Scenarios 1 and 2 described above. The simulation results are given in the Supplementary Material. For Scenario 1 with independent covariates, the adaptive lasso estimators tend to miss some important covariates while with correlated covariates, they tend to select more unimportant covariates, with no improvement as the sample size increases. For Scenario 2, the selection performance of the adaptive lasso estimators improves when the sample size increases and is very comparable with that for the normal covariates reported in Table 1. This suggests that even when the linearity Condition 7 is violated, the adaptive lasso may have the variable selection consistency property under certain types of model misspecification. As in the previous simulations, the adaptive lasso and forward selection with BIC tuning methods have very comparable selection performance, and they are better than the standard lasso and forward selection with AIC tuning for Scenario 2 by selecting fewer unimportant variables. However, for Scenario 1, the advantage of the adaptive lasso and forward selection with BIC tuning methods is not obvious since they also tend to miss more important variables.

Additional simulations were also conducted to compare the selection performance of the adaptive lasso estimates under the true and misspecified models, to consider models with varying effect sizes and to check the asymptotic distributions established in Theorem 1. The results are given in the Supplementary Material. The results show that if Condition 7 is satisfied, the selection performance of adaptive lasso is very similar under the true and misspecified models. In addition, for moderate sample sizes, the sampling distribution of the penalized estimators agrees reasonably well with the limiting distributions.

4.2. Application to diabetic retinopathy data

To further illustrate the robustness of the adaptive lasso method under model misspecification, we analyse data from the Wisconsin Epidemiological Study of Diabetic Retinopathy. In this study, the baseline examination of patients with retinopathy occurred in 1980–1982, with additional data collected at 4-, 10-, 14- and 20-year follow-ups. Study details may be found in Klein et al. (1984, 1989, 1998). The current analysis employs 648 binary response variables coding the status of 4-year progression of retinopathy. As in previous analyses (Wahba et al., 1995; Zhang et al., 2004), we consider 14 potential risk factors, which include 9 continuous covariates, namely the duration of diabetes at baseline examination, glycosylated haemoglobin, body mass index, systolic blood pressure, retinopathy level, pulse rate in 30 s, insulin dose per day, years of school completed, intraocular pressure, and five categorical covariates such as smoking status, sex, use of at least one aspirin for at least 3 months while diabetic, family history of diabetes and marital status.

We fit binary regression models with different link functions g in the model for the probability of retinopathy at 4 years, $E(Y | X) = g(\beta'X)$. Five link functions are examined: probit, log-log and $g(u) = 1 - (1 + re^u)^{-1/r}$ with $r = 0.5, 1$ and 2, where $r = 1$ corresponds to the logistic link. The adaptive lasso method with $\gamma = 1$ is used for variable selection with the tuning parameter

Table 3. *Estimation and selection results for diabetic retinopathy data*

Covariates	Link functions				
	$r = 0.5$	$r = 1$	$r = 2$	Probit	Log-Log
Intercept	-5.176 (0.679)	-6.011 (0.738)	-7.567 (0.933)	-3.637 (0.607)	-4.335 (0.649)
Gly	0.391 (0.044)	0.468 (0.050)	0.617 (0.068)	0.284 (0.038)	0.311 (0.040)
BMI	0.025 (0.019)	0.032 (0.020)	0.046 (0.023)	0.019 (0.018)	0.019 (0.018)

Gly, glycosylated haemoglobin; BMI, body mass index.

chosen using BIC. The adaptive lasso estimates for the regression parameters under different link functions are obtained using the least squares approximation method of Wang & Leng (2007) and the nonzero components of the adaptive lasso estimates are summarized in Table 3. In addition, the standard errors for the nonzero estimates, obtained based on the asymptotic distribution established in Theorem 1, are given in the parentheses. On the basis of the results, the adaptive lasso estimators all select the same two covariates: glycosylated haemoglobin and body mass index, under different link functions, although the estimated covariate effects may be different. The adaptive lasso method exhibits quite robust selection performance across models, where Condition 7 would seem to be violated, owing to the categorical covariates. This is further exemplified by comparing the ratios of the three nonzero coefficients between two models, which are rather constant, as predicted by Corollary 1 assuming Condition 7 is satisfied. Increased glycosylated haemoglobin and increased body mass index increase the risk of disease progression, with the relative effects being comparable across models.

5. REMARKS

While the theoretical findings in the current paper were established for the adaptive lasso under model misspecification, we anticipate that other shrinkage methods will have similar robustness properties as long as the sparsity of the least false parameter is equivalent to that in the true parameter. The objective function is critical in determining whether the necessary conditions are met, not the penalty function. If Corollary 1 holds for the generalized linear model or Corollary 2 holds for proportional hazards model, any suitably penalized likelihood or partial likelihood estimator will achieve sparsity, selection consistency, and the oracle property under misspecification. Further work is needed to rigorously establish such results for other variable selection techniques, like smoothly clipped absolute deviation.

In an as yet unpublished 2009 report from Cornell University G. V. Rocha, X. Wang and B. Yu studied penalized M-estimation using an L_1 penalty, such as the lasso, and established selection consistency for the true model under Gaussian covariates. They do not consider whether their estimators are oracle, in the sense of having the same limiting distribution as that for an estimator for which the zero parameters are known a priori. Our results go further in several respects. First, we study penalized estimation with right censoring under misspecified proportional hazards models, which does not fit into the M-estimation framework of the report by Rocha et al., and obtain selection consistency for the true model under certain conditions on the covariate and censoring distributions. Secondly, we establish selection consistency for M-estimators under weaker conditions than in the report by Rocha et al.. Thirdly, we obtain a general oracle result for M-estimators and misspecified proportional hazards models that does not require selection consistency.

The setting considered in the current paper is that where the number of regression parameters p is fixed and small relative to sample size n , which is assumed to grow. There is recent interest in the case that p grows at some rate of n . Here, one must first establish that the least false parameter

from the penalized estimation procedure is well defined under the model being estimated. This will presumably require sparseness, in the sense that the number of nonzero parameters in the least false parameter must grow at a certain rate. Such sparseness may hold under the fitted model even when it does not hold under the true model. Conversely, it is possible that sparseness holds under the true model but that the penalized least false parameter is not well defined owing to a lack of sparseness under the misspecified model.

With large p , once the existence of the least false parameter has been established, one may study the properties of the corresponding penalized estimator. One may derive conditions such that this estimator is oracle, in the sense of having the same large sample properties as an estimator in which the zero components of the least false parameter are known a priori. Finally, one can determine whether the nonzero components of the least false parameter match those of the regression parameter from the true model. If so, then consistent variable selection is achievable. The primary technical difficulty is showing that there is sufficient information to permit the definition of the least false parameter from the penalized estimation procedure, which is defined implicitly as the limiting value of the resulting estimator. In situations where p grows too fast, this limiting value may not exist with correctly specified models, as discussed in [Fan & Peng \(2004\)](#), [Lam & Fan \(2008\)](#) and [Zou & Zhang \(2009\)](#). One can adapt such proofs, being careful to state the necessary regularity conditions under a potentially misspecified model. This warrants further study.

ACKNOWLEDGEMENT

We would like to thank Michael Kosorok for helpful discussions regarding misspecification of frailty models. We also thank the editor, an associate editor and three referees for very insightful comments. Lu and Fine's research was supported by the National Cancer Institute.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional simulation study results.

APPENDIX

Proof of Theorem 1. The proof is constructed from three steps. The first step studies the behaviour of $\hat{\beta}^{(1)} = \beta^* + n^{-1/2}u$ for a fixed u . The second step proves the asymptotic normality of $\hat{\beta}^{(1)}$. The third step shows the sparsity property, i.e., that $\text{pr}(\hat{\beta}_2^{(1)} = 0) \rightarrow 1$.

Define

$$H_n(u) = Q_n^{(1)}(\beta^* + n^{-1/2}u) - Q_n^{(1)}(\beta^*). \quad (\text{A1})$$

Let $\hat{u}_n = \text{argmax}_u H_n(u)$; then $\hat{u}_n = n^{1/2}(\hat{\beta}^{(1)} - \beta^*)$.

In the following, we study the asymptotic behaviour of (A1) under the generalized linear model. By Taylor expansion, we obtain

$$\begin{aligned} H_n(u) &= n^{-1/2} \sum_{i=1}^n \dot{\ell}(X_i, Y_i; \beta^*)' u + (2n)^{-1} u' \left\{ \sum_{i=1}^n \ddot{\ell}(X_i, Y_i; \beta^*) \right\} u \\ &\quad + (n^{-3/2}/6) \sum_{i=1}^n \sum_{j,k,l=1}^p \frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \beta_l} \ell(X_i, Y_i; \beta^*) u_j u_k u_l \end{aligned}$$

$$\begin{aligned}
& + \lambda_n n^{-1/2} \sum_{j=1}^p \hat{w}_{nj} n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) \\
& \equiv T_1^{(n)} + T_2^{(n)} + T_3^{(n)} + T_4^{(n)}
\end{aligned} \tag{A2}$$

for some $\check{\beta}$ between $\beta^* + n^{-1/2}u$ and β^* .

By White (1982, Theorem 3.2), $T_1^{(n)} \rightarrow N\{0, u' B^{(1)}(\beta^*) u\}$ in distribution. By the law of large numbers, $T_2^{(n)} \rightarrow u' A^{(1)}(\beta^*) u$ in probability. By Condition 5, we obtain that $T_3^{(n)} \rightarrow 0$ in probability. It follows from White (1982, Theorem 2.2) that $\hat{w}_{nj} \rightarrow |\beta_j^*|^{-\gamma}$ in probability. Thus, if $\beta_j^* \neq 0$,

$$n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) \rightarrow u_j \text{sign}(\beta_j^*)$$

in probability, and thus

$$\lambda_n n^{-1/2} \hat{w}_{nj} n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) \rightarrow 0$$

in probability. If $\beta_j^* = 0$, then $n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) = -|u_j|$ and $\lambda_n n^{-1/2} \hat{w}_{nj} = \lambda_n n^{(\gamma-1)/2} |n^{1/2} \tilde{\beta}_j^{(1)}|^{-\gamma}$, where $n^{1/2} \tilde{\beta}_j^{(1)} = O_p(1)$ by White (1982, Theorem 3.2). Thus, similar to Zou (2006), we obtain that

$$\lambda_n n^{-1/2} \hat{w}_{nj} n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) \rightarrow \begin{cases} 0, & \beta_j^* \neq 0, \\ 0, & \beta_j^* = 0, u_j = 0, \\ -\infty, & \beta_j^* = 0, u_j \neq 0 \end{cases} \tag{A3}$$

in probability. Summarizing, for every u , $H_n(u) \rightarrow H(u)$ in probability, where

$$H(u) = \begin{cases} \frac{1}{2} u_1' A_{11}^{(1)} u_1 + u_1' W, & u_j = 0, j \notin \mathcal{A}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{A4}$$

Here $\mathcal{A}$ is the index set of nonzero components in the least false parameters β^* , u_1 is the p_1 -length beginning part of the vector u and $W \sim N(0, B_{11})$.

Next, we show that $\hat{u}_n \rightarrow \hat{u}$ in probability, where $\hat{u}$ is the maximizer of $H(u)$. Note that $H_n(u)$ and $H(u)$ are stochastic processes with sample paths that are upper semicontinuous and $H(u)$ possesses a unique maximum at $\hat{u} = \{-W'(A_{11}^{(1)})^{-1}, 0'\}'$. The inverse of A_{11} is well defined since A_{11} is symmetric and negative definite. To establish that $\hat{u}_n = O_p(1)$, we will show with probability tending to one, there is a local maximizer $\hat{\beta}^{(1)}$ of $\mathcal{Q}_n^{(1)}(\beta)$ such that $n^{-1/2} \hat{u}_n = \|\hat{\beta}^{(1)} - \beta^*\| = O_p(n^{-1/2})$. In particular, for any $\varepsilon > 0$, there exists a constant C such that

$$\text{pr} \left\{ \sup_{\|u\|=C} \mathcal{Q}_n^{(1)}(\beta^* + n^{-1/2} u) < \mathcal{Q}_n^{(1)}(\beta^*) \right\} \geq 1 - \varepsilon. \tag{A5}$$

By Taylor expansion of (A2), we obtain

$$\mathcal{Q}_n^{(1)}(\beta^* + n^{-1/2} u) - \mathcal{Q}_n^{(1)}(\beta^*) = T_1^{(n)} + T_2^{(n)} + T_3^{(n)} + T_4^{(n)}.$$

Note that $T_1^{(n)} = O_p(1)$, $T_2^{(n)} = u' A(\beta^*) u \{1 + o_p(1)\}$, $T_3^{(n)} = o_p(1)$,

$$T_4^{(n)} \leq \lambda_n n^{-1/2} \sum_{j=1}^{p_1} \hat{w}_{nj} n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|),$$

and that $n^{-1/2} \lambda_n \rightarrow 0$, $\hat{w}_{nj} \rightarrow |\beta_j^*|^{-\gamma}$ and

$$n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) \rightarrow -\text{sign}(\beta_j^*) u_j$$

in probability. Since $A(\beta^*)$ is negative definite by Condition 4, taking C large enough, $T_2^{(n)}$ dominates the other three terms, and thus (A5) holds and $\hat{u}_n$ is uniformly tight. Hence, all the conditions of the argmax theorem (Kosorok, 2008, Theorem 14.1) hold, and consequently $\hat{u}_n \rightarrow \hat{u}$ in probability. It follows that $\hat{u}_{n1} \rightarrow -A_{11}^{-1}W$ and $\hat{u}_{n2} \rightarrow 0$ in probability, and hence $n^{1/2}(\hat{\beta}_1^{(1)} - \beta_1^*) \rightarrow N\{0, (A_{11}^{(1)})^{-1}B_{11}^{(1)}(A_{11}^{(1)})^{-1}\}$ in distribution.

The sparsity proof generalizes the proof of Theorem 4 in Zou (2006). For all $j \in \mathcal{A}$, the asymptotic normality above gives that $\text{pr}(j \in \mathcal{A}_n) \rightarrow 1$, where $\mathcal{A}_n$ is the index set of nonzero components in the adaptive lasso estimates $\hat{\beta}^{(1)}$. Thus, it suffices to show that for every $j' \notin \mathcal{A}$, $\text{pr}(j' \in \mathcal{A}_n) \rightarrow 0$. Consider the event $j' \in \mathcal{A}_n$. By the Karush–Kuhn–Tucker optimality conditions, we have

$$\sum_{i=1}^n \dot{\ell}_{j'}(X_i, Y_i, \hat{\beta}^{(1)}) = \lambda_n \hat{w}_{nj'} \text{sign}(\hat{\beta}_{j'}^{(1)}).$$

Thus, $\text{pr}(j' \in \mathcal{A}_n) \leq \text{pr}\{\sum_{i=1}^n \dot{\ell}_{j'}(X_i, Y_i, \hat{\beta}^{(1)}) = \lambda_n \hat{w}_{nj'} \text{sign}(\hat{\beta}_{j'}^{(1)})\}$. By Taylor expansion,

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n \dot{\ell}_{j'}(X_i, Y_i, \hat{\beta}^{(1)}) &= n^{-1/2} \sum_{i=1}^n \dot{\ell}_{j'}(X_i, Y_i, \beta^*) + n^{-1} \sum_{i=1}^n \ddot{\ell}_{kj'}(X_i, Y_i, \beta^*) n^{1/2}(\hat{\beta}^{(1)} - \beta^*) \\ &\quad + (n^{-3/2}/2) \sum_{i=1}^n \sum_{k,l=1}^p \frac{\partial^3}{\partial \beta_l \partial \beta_k \partial \beta_j} \ell(X_i, Y_i; \check{\beta}) \{n^{1/2}(\hat{\beta}_k^{(1)} - \beta_k^*)\} \{n^{1/2}(\hat{\beta}_l^{(1)} - \beta_l^*)\} \\ &\equiv U_1^{(n)} + U_2^{(n)} + U_3^{(n)}, \end{aligned}$$

where $\check{\beta}$ is between β^* and $\hat{\beta}^{(1)}$. By Condition 3, and the fact that $n^{1/2}(\hat{\beta}^{(1)} - \beta^*)$ converges in distribution to a normal random vector, both $U_1^{(n)}$ and $U_2^{(n)}$ converge in distribution to normal random vectors. By Condition 5, $U_3^{(n)} = O_p(n^{-1/2})$. On the other hand, $n^{-1/2} \lambda_n \hat{w}_{nj'} = \lambda_n n^{(\gamma-1)/2} |n^{1/2} \check{\beta}_{j'}^{(1)}|^{-\gamma} \rightarrow \infty$ in probability. Thus, $\text{pr}(j' \in \mathcal{A}_n) \rightarrow 0$ and the proof is complete. $\square$

Proof of Theorem 2. Define $H_n(u)$ as in (A2) for $\mathcal{Q}_n^{(2)}$. Then

$$\begin{aligned} H_n(u) &= n^{-1/2} \sum_{i=1}^n \Delta_i \left\{ X_i - \frac{S^{(1)}(\beta^*, X_i)}{S^{(0)}(\beta^*, X_i)} \right\}' u \\ &\quad - (2n)^{-1} u' \left(\sum_{i=1}^n \Delta_i \left[\frac{S^{(2)}(\check{\beta}, X_i)}{S^{(0)}(\check{\beta}, X_i)} - \left\{ \frac{S^{(1)}(\check{\beta}, X_i)}{S^{(0)}(\check{\beta}, X_i)} \right\}^{\otimes 2} \right] \right) u \\ &\quad + \lambda_n n^{-1/2} \sum_{j=1}^p \hat{w}_{nj} n^{1/2} (|\beta_j^*| - |\beta_j^* + n^{-1/2} u_j|) \\ &\equiv D_1^{(n)} + D_2^{(n)} + D_3^{(n)}, \end{aligned} \tag{A6}$$

where $\check{\beta}$ is on a line segment between β and β^* . It follows from Lin & Wei (1989) that $D_1^{(n)} \rightarrow N(0, u' B^{(2)} u)$ in distribution and that $D_2^{(n)} \rightarrow -u' A(\beta^*) u$ in probability. The asymptotic behaviour of $D_3^{(n)}$ is given by (A3), where the facts that $\hat{w}_{nj} \rightarrow |\beta_j^*|^{-\gamma}$ in probability when $\beta_j^* \neq 0$ and that $n^{1/2} \check{\beta}_j^{(2)} = O_p(1)$ when $\beta_j^* = 0$ follow from Sasiemi (1993, Corollary 4.1). Therefore, for every u , $H_n(u) \rightarrow H(u)$ in probability, where $H(u)$ is given by (A4).

We would like to show that $\hat{u}_n \rightarrow \hat{u}$ in probability, where $\hat{u}$ is the maximizer of $H(u)$. If $\hat{u} = O_p(1)$, then the argmax theorem (Kosorok, 2008, Theorem 14.1) can be applied, similar to the proof for the generalized linear model. We need that, for any given $\varepsilon > 0$, there exists a constant C such that

$$\text{pr} \left\{ \sup_{\|u\|=C} \mathcal{Q}_n^{(2)}(\beta^* + n^{-1/2} u) < \mathcal{Q}_n^{(2)}(\beta^*) \right\} \geq 1 - \varepsilon. \tag{A7}$$

By Taylor expansion of (A6), we obtain

$$\mathcal{Q}_n^{(2)}(\beta^* + n^{-1/2}u) - \mathcal{Q}_n^{(2)}(\beta^*) = D_1^{(n)} + D_2^{(n)} + D_3^{(n)}.$$

Note that $D_1^{(n)} = O_p(1)$, $D_2^{(n)} = -u' A(\beta^*)u\{1 + o_p(1)\}$ and $D_3^{(n)} = o_p(1)$. Since $A(\beta^*)$ is positive definite (see Sasieni, 1993, Lemma 7.3), taking C large enough, $D_2^{(n)}$ dominates the other two terms, and thus (A7) holds and $\hat{u}_n$ is uniformly tight.

For sparsity, it suffices to show that for every $j' \notin \mathcal{A}$, $\text{pr}(j' \in \mathcal{A}_n) \rightarrow 0$. Consider the event $j' \in \mathcal{A}_n$. By the Karush–Kuhn–Tucker optimality conditions, we have

$$\sum_{i=1}^n \Delta_i \left\{ X_i - \frac{S^{(1)}(\hat{\beta}^{(2)}, X_i)}{S^{(0)}(\hat{\beta}^{(2)}, X_i)} \right\} = \lambda_n \hat{w}_{nj'} \text{sign}(\hat{\beta}_{j'}^{(2)}).$$

Thus,

$$\text{pr}(j' \in \mathcal{A}_n) \leq \text{pr} \left[\sum_{i=1}^n \Delta_i \left\{ X_i - \frac{S^{(1)}(\hat{\beta}^{(2)}, X_i)}{S^{(0)}(\hat{\beta}^{(2)}, X_i)} \right\} = \lambda_n \hat{w}_{nj'} \text{sign}(\hat{\beta}_{j'}^{(2)}) \right].$$

By Taylor expansion, we have

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n \Delta_i \left\{ X_i - \frac{S^{(1)}(\hat{\beta}^{(2)}, X_i)}{S^{(0)}(\hat{\beta}^{(2)}, X_i)} \right\} &= n^{-1/2} \sum_{i=1}^n \Delta_i \left\{ X_i - \frac{S^{(1)}(\beta^*, X_i)}{S^{(0)}(\beta^*, X_i)} \right\} \\ &\quad - n^{-1} \left(\sum_{i=1}^n \Delta_i \left[\frac{S^{(2)}(\check{\beta}, X_i)}{S^{(0)}(\check{\beta}, X_i)} - \left\{ \frac{S^{(1)}(\check{\beta}, X_i)}{S^{(0)}(\check{\beta}, X_i)} \right\}^{\otimes 2} \right] \right) n^{1/2}(\hat{\beta}^{(2)} - \beta^*) \\ &\equiv J_1^{(n)} + J_2^{(n)}, \end{aligned}$$

where $\check{\beta}$ is between β^* and $\hat{\beta}^{(2)}$. Combining earlier approximations yields

$$n^{-1/2} \sum_{i=1}^n \Delta_i \left\{ X_i - \frac{S^{(1)}(\hat{\beta}^{(2)}, X_i)}{S^{(0)}(\hat{\beta}^{(2)}, X_i)} \right\} \rightarrow N\{0, B^{(2)}(\beta^*)\}$$

in distribution, and

$$n^{-1} \sum_{i=1}^n \Delta_i \left[\frac{S^{(2)}(\check{\beta}, X_i)}{S^{(0)}(\check{\beta}, X_i)} - \left\{ \frac{S^{(1)}(\check{\beta}, X_i)}{S^{(0)}(\check{\beta}, X_i)} \right\}^{\otimes 2} \right] \rightarrow A^{(2)}(\beta^*)$$

in probability. By Sasieni (1993, Corollary 4.1), $n^{1/2}(\hat{\beta}^{(2)} - \beta^*)$ converges in distribution to a normal random vector. By applying Slutsky's theorem, we obtain that both $J_1^{(n)}$ and $J_2^{(n)}$ converge in distribution to normal random vectors. On the other hand, $n^{-1/2} \lambda_n \hat{w}_{nj'} \rightarrow \infty$ in probability. Thus, $\text{pr}(j' \in \mathcal{A}_n) \rightarrow 0$ and the proof is complete. $\square$

REFERENCES

- AKAIKE, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–65.
- ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–20.
- COX, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34**, 187–220.
- 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.

- FAN, J. & PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–61.
- KLEIN, R., KLEIN, B., MOSS, S. E., DAVIS, M. D. & DEMETS, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopathy II: prevalence and risk of diabetes when age at diagnosis is less than 30 years. *Arch. Ophthalmol.* **102**, 520–6.
- KLEIN, R., KLEIN, B., MOSS, S. E., DAVIS, M. D. & DEMETS, D. L. (1989). The Wisconsin epidemiologic study of diabetic retinopathy IX: four-year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.* **107**, 237–43.
- KLEIN, R., KLEIN, B., MOSS, S. & CRUICKSHANKS, K. (1998). The Wisconsin epidemiologic study of diabetic retinopathy XVII: the 14-year incidence and progression of diabetic retinopathy and associated risk factors in type 1 diabetes. *Ophthalmology* **105**, 1801–15.
- KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.
- KOSOROK, M. R., LEE, B. L. & FINE, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.* **32**, 1448–91.
- LAM, C. & FAN, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36**, 2232–60.
- LI, K.-C. & DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009–52.
- LIN, D. Y. & WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074–8.
- MALLOWS, C. L. (1973). Some comments on C_p . *Technometrics* **15**, 661–75.
- MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. New York: Chapman and Hall.
- R DEVELOPMENT CORE TEAM. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- SASIENI, P. (1993). Some new estimators for Cox regression. *Ann. Statist.* **21**, 1721–59.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
- TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–88.
- TIBSHIRANI, R. J. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385–95.
- WAHBA, G., WANG, Y., GU, C., KLEIN, R. & KLEIN, B. (1995). Smoothing spline anova for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23**, 1865–95.
- WANG, H. & LENG, C. (2007). Unified lasso estimation with least squares approximation. *J. Am. Statist. Assoc.* **102**, 1039–48.
- WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
- ZHANG, H. H. & LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.
- ZHANG, H. H., WAHBA, G., LIN, Y., VOELKER, M., FERRIS, M., KLEIN, R. & KLEIN, B. (2004). Variable selection and model building via likelihood basis pursuit. *J. Am. Statist. Assoc.* **99**, 659–72.
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.
- ZOU, H. & ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733–51.

[Received April 2011. Revised February 2012]