Biometrika (2005), **92,** 2, *pp.* 303–316

Printed in Great Britain

Variable selection for multivariate failure time data

BY JIANWEN CAI

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

cai@bios.unc.edu

JIANQING FAN

Department of Operation Research and Financial Engineering, Princeton University, New Jersey 08544, U.S.A.

jqfan@princeton.edu

RUNZE LI

Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802-2111, U.S.A.

rli@stat.psu.edu

AND HAIBO ZHOU

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

zhou@bios.unc.edu

SUMMARY

In this paper, we propose a penalised pseudo-partial likelihood method for variable selection with multivariate failure time data with a growing number of regression coefficients. Under certain regularity conditions, we show the consistency and asymptotic normality of the penalised likelihood estimators. We further demonstrate that, for certain penalty functions with proper choices of regularisation parameters, the resulting estimator can correctly identify the true model, as if it were known in advance. Based on a simple approximation of the penalty function, the proposed method can be easily carried out with the Newton–Raphson algorithm. We conduct extensive Monte Carlo simulation studies to assess the finite sample performance of the proposed procedures. We illustrate the proposed method by analysing a dataset from the Framingham Heart Study.

Some key words: Cox's model; Marginal hazards model; Penalised likelihood; Smoothly clipped absolute deviation; Variable selection.

1. INTRODUCTION

Deciding which covariates are to be included in the final statistical model has always been a tricky task for investigators, and a valid and unified statistical model selection criterion is desirable. We propose a penalised pseudo-partial likelihood method for variable selection in multivariate failure time analysis. Our research is motivated by the need

to develop a predictive model that relates multiple failure time outcomes, namely time to coronary heart disease and time to cerebrovascular accident, and a vector of risk factors for patients in the Framingham Heart Study (Dawber, 1980). The primary sampling unit is the family, and it is likely that the failure times recorded for subjects within a family are correlated. Extensions of the Cox regression model (Cox, 1972) for the analysis of multivariate failure time data include the frailty model and the marginal model. When the correlation among the observations is not of interest, the marginal proportional hazards models has received considerable attention in the recent literature (Wei et al., 1989; Lee et al., 1992; Liang et al., 1993; Lin, 1994; Cai & Prentice, 1995, 1997; Spiekerman & Lin, 1998; Clegg et al., 1999). Thus, we will focus on the marginal models.

Some of the variable selection criteria and procedures in linear regression analysis have been extended to the Cox model: Tibshirani (1997) extended his LASSO variable selector; Faraggi & Simon (1998) proposed a Bayesian variable selection method for the Cox model; Cai (1999) extended the generalised likelihood ratio method to deal with multivariate failure time data; Fan & Li (2002) extended their nonconcave penalised likelihood approach; Huang & Harrington (2002) proposed penalised partial likelihood with a quadratic penalty to deal with issues of collinearity of covariates; and Bunea & McKeague (2005) extended BIC-type (Schwarz, 1978) variable selection criteria to the Cox model. In general, the variable selection procedure for multivariate failure time data is underdeveloped, and this paper intends to fill that gap.

Fan & Li (2002) studied penalised partial likelihood for variable selection problems and demonstrated that their penalised partial likelihood procedure performs as well as an oracle estimator, namely the estimator constructed with the aid of an oracle who knows the true model, i.e. the subset of variables with nonvanishing coefficients, only in finite-parameter settings. However, they do not address the fundamental issues in model selection. In practice, to reduce possible modelling biases, many variables are introduced at the initial stage of modelling. Huber (1973) noted that, in the context of variable selection, the number of parameters is often large and should be modelled as d_n, which tends to infinity as the sample size n tends to infinity. In this paper, we intend to address the fundamental problems of variable selection for the Cox marginal model with a diverging number of parameters. To this end, we propose a new formulation for variable selection, which differs from that in Fan & Li (2002). We study asymptotic properties of penalised pseudolikelihood in the context of the marginal multivariate failure model, when the number of regression coefficients tends to infinity. This includes of course the use of penalised partial likelihood in the proportional hazards model. We first show the root-n/d_n consistency of the penalised pseudo-partial likelihood estimator and then demonstrate that the newly proposed variable selection procedures still possess the oracle property. As a consequence, our results directly provide the asymptotic behaviours of the maximum partial likelihood estimator (Andersen & Gill, 1982) and the maximum pseudo-partial likelihood estimator (Spiekerman & Lin, 1998; Clegg et al., 1999) when the number of covariates grows with sample size.

2. Model selection with a penalised pseudo-partial likelihood

2·1. *Notation*

To fix notation, suppose that there are n independent clusters and that each cluster has K_i subjects. For each subject, J types of failure may occur. For the failure time in the case of the jth type of failure on subject k in cluster i, the marginal hazards model is taken

either as

$$h_{ijk}\{t|x_{ijk}(t)\} = h_{0j}(t)\exp\{\beta^{\mathrm{T}}x_{ijk}(t)\}, \tag{2·1}$$

or

$$h_{ijk}\{t|x_{ijk}(t)\} = h_{0}(t)\exp\{\beta^{\mathrm{T}}x_{ijk}(t)\}, \tag{2·2}$$

where $\beta = (\beta_1, \ldots, \beta_{d_n})^{\mathrm{T}}$ is a vector of unknown regression coefficients, d_n is the dimension of β, $x_{ijk}(t)$ is a possibly external time-dependent covariate vector, and $h_{0j}(t)$ and $h_0(t)$ are unspecified baseline hazard functions. Model (2·1) is commonly referred to as the mixed baseline hazards model, while (2·2) is the common baseline hazards model.

2·2. *Penalised pseudo-partial likelihood*

The marginal model approach does not specify correlation structure for the failure times within a cluster, and hence inferences are based on a pseudo-partial likelihood approach. For ease of presentation, we drop the subscript and let T, C and $x(t)$ be the survival time, the censoring time and their associated covariates, respectively. Correspondingly, let $Z = \min\{T, C\}$ be the observed time, let $\delta = I(T \leqslant C)$ be the censoring indicator, and let $Y(t) = I(Z \geqslant t)$ be the at-risk indicator. We further assume that T and C are conditionally independent given x and that the censoring mechanism is noninformative. Under a working independence assumption (Wei et al., 1989), i.e. assuming independence among failure times in a cluster, we obtain the logarithm of a pseudo-partial likelihood function for model (2·1) as

$$\ell(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_i}\delta_{ijk}\left(\beta^{\mathrm{T}}x_{ijk}(Z_{ijk}) - \log\left[\sum_{l=1}^{n}\sum_{g=1}^{K}Y_{ljg}(Z_{ijk})\exp\{\beta^{\mathrm{T}}x_{ljg}(Z_{ijk})\}\right]\right). \tag{2·3}$$

To balance modelling biases and estimation variance, many traditional variable selection criteria have resorted to the use of penalised likelihood, including the AIC (Akaike, 1973) and BIC (Schwarz, 1978). We use a penalised pseudo-partial likelihood for model (2·1) which is defined as

$$\mathscr{L}(\beta) = \ell(\beta) - n\sum_{j=1}^{d_n}p_{\lambda_j}(|\beta_j|), \tag{2·4}$$

where $p_{\lambda_j}(|\beta_j|)$ is a given nonnegative function called a penalty function with λ_j as a regularisation or tuning parameter. The tuning parameters can be chosen subjectively by data analysts or objectively by data themselves. In general, large values of λ_j's result in simpler models with fewer selected variables. The penalty term in (2·4) is more general than that in Fan & Li (2001), who considered $\lambda_j \equiv \lambda$. Allowing covariate-specific tuning parameters enables different regression coefficients to have different penalty functions, and thus the penalised pseudo-partial likelihood may directly incorporate hierarchical prior information about the unknown coefficients. For instance, we may wish to keep the main effects of some important confounding variables in the model by not penalising their corresponding coefficients.

Many classical variable selection criteria are special cases of (2·4). For instance, consider the L_0 penalty $p_\lambda(|\theta|) = \frac{1}{2}\lambda^2 I\{|\theta| \neq 0\}$, also called the entropy penalty, where $I(.)$ is an indicator function. Also, AIC (Akaike, 1973), BIC (Schwarz, 1978), the ϕ-criterion (Shibata, 1984) and RIC (Foster & George, 1994) correspond to $\lambda = (2/n)^{\frac{1}{2}}$, $\{\log(n)/n\}^{\frac{1}{2}}$,

$[\log\{\log(n)\}]^{\frac{1}{2}}$ and $\{\log(d_n)/n\}^{\frac{1}{2}}$, respectively, where n is the sample size, although these criteria were motivated from different principles. Since the entropy penalty function is discontinuous, one requires to search over all possible subsets to maximise (2·4). Hence it is very expensive computationally. Furthermore, as analysed by Breiman (1996), best-subset variable selection suffers from several drawbacks, including its lack of stability.

Recently, authors have considered continuous penalty functions. The L_1 penalty, defined by $p_\lambda(|\theta|) = \lambda|\theta|$, results in the LASSO variable selector (Tibshirani, 1996). Fan & Li (2001) advocated the smoothly clipped absolute deviation penalty, whose first derivative is defined by

$$p'_\lambda(\theta) = \lambda I(\theta \leqslant \lambda) + \frac{(a\lambda - \theta)_+}{a-1} I(\theta > \lambda), \tag{2·5}$$

for some $a > 2$ and $\theta > 0$, with $p_\lambda(0) = 0$. This penalty improves the entropy penalty function by saving computational cost and resulting in a continuous solution to avoid unnecessary modelling variation. Furthermore, it improves the L_1 penalty by avoiding excessive estimation bias.

2·3. *Oracle properties*

The penalised pseudo-partial likelihood estimator, denoted by $\hat{\beta}$, maximises (2·4). For certain penalty functions, such as the L_1 penalty and the smoothly clipped absolute deviation penalty, maximising $\mathscr{L}(\beta)$ will result in some vanishing estimates of coefficients and their associated variables are deleted. Hence, by maximising $\mathscr{L}(\beta)$, we select a model and estimate its parameters simultaneously. We now present the asymptotic properties for $\hat{\beta}$ and show that it could perform as well as an oracle estimator.

Denote by β_0 the true value of β. Furthermore, let β_{10} and β_{20} denote the nonzero and zero components of β_0, respectively. Denote by s_n the dimension of β_{10} and let

$$a_n = \max_{1 \leqslant j \leqslant s_n} \{|p'_{\lambda_{jn}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \quad b_n = \max_{1 \leqslant j \leqslant s_n} \{|p''_{\lambda_{jn}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \tag{2·6}$$

In this section, we use λ_{jn} rather than λ_j to emphasise its dependence on n. We first show that there exists a penalised pseudo-partial likelihood estimator that converges at rate $O_P\{\sqrt{(d_n)}(n^{-\frac{1}{2}} + a_n)\}$, and then we establish the oracle property for the resulting estimator. We only state the main results here and relegate the regularity conditions and proofs to the Appendix.

THEOREM 1. *Under Conditions* A1–A4 *in the Appendix, if* $a_n \to 0$, $b_n \to 0$ *and* $d_n^4/n \to 0$, *as* $n \to \infty$, *then, with probability tending to one, there exists a local maximiser* $\hat{\beta}$ *of* $\mathscr{L}(\beta)$, *defined in* (2·4), *such that* $\|\hat{\beta} - \beta_0\| = O_P\{\sqrt{(d_n)}(n^{-\frac{1}{2}} + a_n)\}$.

From Theorem 1, provided that $a_n = O(n^{-\frac{1}{2}})$, there exists a $\sqrt{(n/d_n)}$-consistent penalised pseudo-partial likelihood estimator. This consistency rate is the same as that of the maximum likelihood estimator for the exponential family (Portnoy, 1988).

We now establish an oracle property. Let

$$\Sigma = \operatorname{diag}\{p''_{\lambda_{1n}}(|\beta_{10}|), \ldots, p''_{\lambda_{s_n n}}(|\beta_{s_n 0}|)\},$$

$$b = (p'_{\lambda_{1n}}(|\beta_{10}|)\operatorname{sgn}(\beta_{10}), \ldots, p'_{\lambda_{s_n n}}(|\beta_{s_n 0}|)\operatorname{sgn}(\beta_{s_n 0}))^{\mathrm{T}}.$$

Since the length of $\hat{\beta}_1$ depends on n, we will follow the formulation in Huber (1973) and Portnoy (1988), and consider any linear combination $c_n^{\mathrm{T}}\hat{\beta}_1$ in the following theorem.

THEOREM 2. *Under Conditions* A1–A5 *in the Appendix, if* $b_n \to 0$, $d_n^4/n \to 0$, $\lambda_{jn} \to 0$, $\lambda_{jn}\sqrt{(n/d_n)} \to \infty$ *and* $a_n = O(n^{-\frac{1}{2}})$, *then, under the conditions of Theorem* 1, *with probability tending to* 1, *the* $\sqrt{(n/d_n)}$*-consistent local maximiser* $\hat{\beta} = (\hat{\beta}_1^{\mathrm{T}}, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$ *in Theorem* 1 *must be such that* (i) $\hat{\beta}_2 = 0$ *and* (ii) , *for any nonzero constant* $s_n \times 1$ *vector* c_n *with* $c_n^{\mathrm{T}} c_n = 1$,

$$\sqrt{n} c_n^{\mathrm{T}} \Gamma_{11}^{-\frac{1}{2}} (A_{11} + \Sigma)\{\hat{\beta}_1 - \beta_{10} + (A_{11} + \Sigma)^{-1} b\} \to N(0, 1), \tag{2·7}$$

in distribution, where A_{11} *and* Γ_{11} *consist of the first* s_n *columns and rows of* $A(\beta_{10}, 0)$ *and* $\Gamma(\beta_{10}, 0)$, *defined in the Appendix, respectively.*

Theorem 2 provides a foundation for choosing estimators that will have the oracle property. For example, with the smoothly clipped absolute deviation penalty, we have $a_n = 0$, $b = 0$ and $\Sigma = 0$ for sufficiently large n. Hence, according to Theorem 2, we have that

$$\sqrt{n} c_n^{\mathrm{T}} \Gamma_{11}^{-\frac{1}{2}} A_{11} (\hat{\beta}_1 - \beta_{10}) \to N(0, 1),$$

in distribution. The estimator $\hat{\beta}_1$ shares the same sampling property as the oracle estimator. Furthermore, $\hat{\beta}_2 = 0$ is the same as the oracle estimator that knows in advance that $\beta_2 = 0$. In other words, the penalised pseudo-partial likelihood estimator possesses the oracle property. In contrast, it can easily be shown from Theorems 1 and 2 that the procedure based on the L_1 penalty does not possess the oracle property, because of the excessive biases.

2·4. *Issues in practical implementation*

Since penalty functions such as the smoothly clipped absolute deviation and the L_1 are singular at the origin, it is challenging to maximise $\mathscr{L}(\beta)$. Following Fan & Li (2001), we will use a local quadratic approximation to the penalty function in our implementation. Suppose that we are given an initial value $\beta^{(0)}$ that is close to the true value of β. If $\beta_j^{(0)}$ is not close to 0, then the penalty function is locally approximated by a quadratic function as

$$p_{\lambda_j}(|\beta_j|) \doteqdot q_{\lambda_j}(|\beta_j|) \equiv p_{\lambda_j}(|\beta_j^{(0)}|) + \tfrac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}).$$

Otherwise, set $\hat{\beta}_j = 0$. With the aid of the local quadratic approximation, the Newton–Raphson algorithm can be applied to maximise the penalised pseudo-partial likelihood function. We set the maximum pseudo-partial likelihood estimate $\hat{\beta}^u$, the maximiser of $\ell(\beta)$ in (2·3), as the initial value of β since it is $(n/d_n)^{\frac{1}{2}}$-consistent by Theorem 1 with $\lambda_j = 0$.

The modified Newton–Raphson algorithm also allows us to estimate the variance-covariance matrix for $\hat{\beta}$ by using the sandwich formula:

$$\hat{\mathrm{cov}}(\hat{\beta}) = \{\mathscr{L}''_a(\hat{\beta})\}^{-1} \,\hat{\mathrm{cov}}\{\mathscr{L}'_a(\hat{\beta})\}\{\mathscr{L}''_a(\hat{\beta})\}^{-1},$$

where $\mathscr{L}_a(\beta) = \ell(\beta) - n\sum_{j=1}^{d_n} q_{\lambda_j}(|\beta_j|)$. Therefore, $\mathscr{L}''_a(\hat{\beta}) = \ell''(\hat{\beta}) - n\Sigma_\lambda(\hat{\beta})$, where

$$\Sigma_\lambda(\hat{\beta}) = \mathrm{diag}\{p'_{\lambda_1}(|\hat{\beta}_1|)/|\hat{\beta}_1|, \ldots, p'_{\lambda_{d_n}}(|\hat{\beta}_{d_n}|)/|\hat{\beta}_{d_n}|\},$$

and $\mathrm{cov}\{\mathscr{L}'_a(\hat{\beta})\}$ is estimated by $\hat{\mathrm{cov}}\{\ell'(\hat{\beta})\}$. The sandwich formula applies only to non-zero estimated coefficients. The performance of this estimator will be examined in our simulation studies.

Similarly to Fan & Li (2002), we will employ generalised crossvalidation to select the λ_j's. In the last step of the Newton–Raphson iteration, we may compute the effective number of parameters, given by

$$e(\lambda_1, \ldots, \lambda_{d_n}) = \mathrm{tr}[\{\mathscr{L}''_a(\hat{\beta})\}^{-1} \ell''(\hat{\beta})].$$

The generalised crossvalidation statistic is defined by

$$\text{GCV}(\lambda_1, \ldots, \lambda_{d_n}) = \frac{-\ell(\hat{\beta})}{n\{1 - e(\lambda_1, \ldots, \lambda_{d_n})/n\}^2}.$$

The minimisation problem over a d_n-dimensional space is difficult. However, it is expected that the magnitude of λ_j should be proportional to the standard error of the unpenalised maximum pseudo-partial likelihood estimator of β_j. In practice, we suggest taking $\lambda_j = \lambda \text{SE}(\hat{\beta}_j^u)$, where $\text{SE}(\hat{\beta}_j^u)$ is the estimated standard error of $\hat{\beta}_j^u$. Such a choice of λ_j works well from our simulation experience. Thus, the minimisation problem will reduce to a one-dimensional problem, and the tuning parameter can be estimated by a grid search.

2·5. *Extensions*

The rate of convergence and the oracle property for the marginal model (2·1) can be easily extended to other marginal hazards models, such as (2·2), with a slightly different pseudo-partial likelihood function. For example, for the common baseline hazards model (2·2), we can use the following pseudo-partial likelihood:

$$\ell_c(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K_i} \delta_{ijk} \left(\beta^{\mathrm{T}} x_{ijk}(Z_{ijk}) - \log \left[\sum_{l=1}^{n} \sum_{m=1}^{J} \sum_{g=1}^{K_i} Y_{lmg}(Z_{ijk}) \exp\{\beta^{\mathrm{T}} x_{lmg}(Z_{ijk})\} \right] \right).$$

The corresponding asymptotic results in Theorems 1 and 2 for the estimator based on the penalised pseudo-partial likelihood $\ell_c(\beta) - n \sum_{j=1}^{d_n} p_{\lambda_j}(|\beta_j|)$ can be established using similar arguments to those in the Appendix.

3. Numerical study and application

3·1. *Simulation study*

In our simulations, we take $J = K = 2$, and the failure times T_{i11}, T_{i12}, T_{i21} and T_{i22} for the ith cluster are generated from the multivariate Clayton–Oakes distribution (Clayton & Cuzick, 1985; Oakes, 1989) with a marginal exponential distribution for the two types of failure and for the two subjects in a cluster

$$\begin{aligned} &\text{pr}(T_{i11} > t_{i11}, T_{i12} > t_{i12}, T_{i21} > t_{i21}, T_{i22} > t_{i22} | x_{ijk}, j = 1, 2, k = 1, 2) \\ &\qquad = \left[\sum_{j=1}^{2} \sum_{k=1}^{2} \exp\{t_{ijk} \lambda_{0j} \theta^{-1} \exp(\beta^{\mathrm{T}} x_{ijk})\} - 3 \right]^{-\theta}, \end{aligned}$$

where $\beta = (0{\cdot}6, 0, 0, -0{\cdot}8, 0, 0, 0{\cdot}35, 0)^{\mathrm{T}}$, which is an eight-dimensional vector consisting of three nonzero components and five zero components. In our simulation, $\lambda_{01} = 1$ and $\lambda_{02} = 5$. The covariate vector x_{ijk} has a normal distribution with standard normal marginals and the correlation between x_{ijkl} and $x_{ijkl'}$ being $\rho^{|l-l'|}$ with $\rho = 0{\cdot}5$. Censoring times C_{ijk} are generated from the $\text{Un}(0, c)$ distribution. We took $c = 5$ or 1, corresponding to censoring rates of approximately 15% and 40%, respectively. For the multivariate Clayton–Oakes distribution, $\theta \to 0$ gives the maximal positive correlation of 1 between failure times and $\theta \to \infty$ corresponds to independence. In our simulation, θ was chosen to be 0·25, 1·5 or 5, which corresponds to high, moderate or low positive dependence, respectively. The number of clusters was taken as $n = 100$. The smoothly clipped absolute deviation penalty function involves two tuning parameters λ and a. Following Fan & Li (2001), we set $a = 3{\cdot}7$ throughout § 3. We use generalised crossvalidation to select the

tuning parameter for the smoothly clipped absolute deviation method and the L_1 method. We compare the proposed penalised pseudo-partial likelihood procedures with the best subset variable selection and the oracle procedure in terms of model error, model complexity and rate of correctly identifying the true model.

We examine the performance of the proposed penalised likelihood procedures with various penalties using model error. The model error is defined as $\text{ME}(\hat{\mu}) = E\{E(Y|x) - \hat{\mu}(x)\}^2$ for a general regression model with $E(Y|x) = \mu(x)$. In our simulations, the baseline hazard function $h_{0j}(t)$ is taken to be a constant h_j. By some straightforward calculations, the model error for $\mu_j(x) = E(T_{.j}|x)$ in our simulation settings may be approximated by

$$\text{ME}(\hat{\mu}_j) \doteq h_j^{-2}(\hat{\beta} - \beta)^{\mathrm{T}}\{Exx^{\mathrm{T}} \exp(-2\beta^{\mathrm{T}}x)\}(\hat{\beta} - \beta),$$

which will be referred to as the approximate model error.

We define the relative approximate model error of a procedure to be the ratio of its approximate model error to that of the maximum pseudo-partial likelihood estimates from the full model. Table 1 gives the median and median absolute deviation of ratios of approximate model error of the proposed procedures over 500 simulations. The average number of zero coefficients demonstrates how the proposed procedure reduces model

Table 1: *Simulation study. Relative approximate model errors, where c is the range of censoring time,* C *is the average number of coefficients correctly estimated as* 0, *and* I *is the average number of coefficients erroneously estimated as* 0

	$c = 5$				$c = 1$			
	RAME	Zero coef.		RITM	RAME	Zero coef.		RITM
Method	median (MAD)	C	I	(%)	median (MAD)	C	I	(%)
				$\theta = 0{\cdot}25$				
SCAD	0·685 (0·201)	4·948	0·018	93·2	0·632 (0·234)	4·910	0·064	86·6
L_1	0·916 (0·067)	3·600	0·000	24·0	0·909 (0·080)	3·502	0·002	19·6
AIC	0·849 (0·104)	4·252	0·000	45·6	0·826 (0·120)	4·162	0·004	40·6
BIC	0·724 (0·181)	4·852	0·002	85·6	0·670 (0·192)	4·828	0·010	83·4
RIC	0·821 (0·122)	4·406	0·000	50·8	0·776 (0·138)	4·352	0·004	47·6
ϕ	0·891 (0·083)	3·900	0·000	32·2	0·871 (0·099)	3·810	0·002	26·0
Oracle	0·651 (0·219)	5·000	0·000	100	0·577 (0·218)	5·000	0·000	100
				$\theta = 1{\cdot}5$				
SCAD	0·591 (0·223)	4·940	0·008	94·4	0·613 (0·231)	4·922	0·060	88·0
L_1	0·900 (0·076)	3·610	0·000	22·2	0·899 (0·081)	3·530	0·000	21·0
AIC	0·821 (0·131)	4·198	0·000	43·4	0·818 (0·120)	4·182	0·002	41·6
BIC	0·629 (0·211)	4·848	0·000	86·2	0·650 (0·209)	4·838	0·002	86·2
RIC	0·780 (0·148)	4·376	0·000	51·0	0·784 (0·134)	4·382	0·000	50·8
ϕ	0·866 (0·102)	3·882	0·000	29·2	0·862 (0·092)	3·854	0·000	28·0
Oracle	0·561 (0·226)	5·000	0·000	100	0·541 (0·231)	5·000	0·000	100
				$\theta = 5$				
SCAD	0·604 (0·214)	4·926	0·016	92·2	0·585 (0·229)	4·932	0·074	86·8
L_1	0·904 (0·083)	3·474	0·002	20·0	0·895 (0·085)	3·622	0·002	21·8
AIC	0·794 (0·128)	4·150	0·002	40·4	0·808 (0·135)	4·228	0·002	41·8
BIC	0·644 (0·201)	4·848	0·002	86·6	0·613 (0·209)	4·840	0·004	85·0
RIC	0·749 (0·134)	4·370	0·002	48·2	0·772 (0·147)	4·386	0·004	47·8
ϕ	0·870 (0·108)	3·800	0·002	27·4	0·856 (0·103)	3·948	0·000	30·0
Oracle	0·565 (0·205)	5·000	0·000	100	0·511 (0·209)	5·000	0·000	100

RAME, relative approximate model error; RITM, rate of identifying the true model; MAD, median absolute deviation; SCAD, smoothly clipped absolute deviation.

complexity and is reported in Table 1, in which the column labelled 'C' stands for the average number of coefficients correctly estimated as 0, while the column labelled 'I' depicts the average number of coefficients erroneously estimated as 0. The rate of correctly identifying the true model is also reported in Table 1, in which SCAD, L_1, AIC, BIC, RIC and ϕ stand for the penalised likelihood procedure with the smoothly clipped absolute deviation, L_1, AIC, BIC, RIC and ϕ penalties, as defined in § 2, respectively, and 'Oracle' for the oracle procedure. Since the entropy penalty is discontinuous, the solutions for AIC, BIC, RIC and ϕ are obtained by exhaustively searching over all possible subsets. Thus, the resulting subsets are indeed the best subsets for the corresponding criterion, and the computational cost for these procedures is much more expensive than that for the smoothly clipped absolute deviation and L_1 methods. Table 1 shows that the smoothly clipped absolute deviation method outperforms the other variable selection procedures in terms of model error, model complexity and rate of correctly identifying the true model. Furthermore, its ratio of approximate model error is very close to that of the oracle estimator, which is consistent with the result in Theorem 2, and the method reduces the model complexity almost as effectively as the oracle procedure.

We have also tested the accuracy of the standard error formula using the sandwich formula. To save space, we do not present the results here; see the authors' technical report for thorough discussion. In general, the sandwich formula gives us accurate estimates of standard errors and coverage probabilities which are close to the nominal level.

3·2. *Analysis of the Framingham study dataset*

We illustrate the proposed variable selection procedures by an analysis of a dataset collected in the Framingham Heart Study. The study was initiated in 1948, with 2336 men and 2873 women aged between 30 and 62 years at their baseline examination (Dawber, 1980). Multiple failure outcomes, such as times to coronary heart disease and cerebrovascular accident, were observed from the same individual. In addition, as the primary sampling unit was the family, failure times are likely to be dependent for the individuals within a family.

For simplicity, we consider only time taken to obtain first evidence of coronary heart disease and of a cerebrovascular accident, and analyse only data for participants who had an examination at age 44 or 45 and were disease-free at that examination. By disease-free we mean that there exists no history of hypertension or glucose intolerance and no previous experience of coronary heart disease or a cerebrovascular accident. The time origin is the time of the examination at which an individual participated in the study and the follow-up information is up to the year 1980. The risk factors of interest are as follows: body mass index, denoted by x_1; cholesterol level, x_2; systolic blood pressure, x_3; smoking status, x_4, coded as 1 if this individual is a smoker, and 0 otherwise; gender, x_5, coded as 1 for female and 0 for male. The values of risk factors were taken from the biennial examination at which an individual was included in the sample. Since some individuals were in the study for several years prior to inclusion into the dataset, the waiting time, denoted by x_6, from entering the study to reaching 44 or 45 years of age was used as a covariate to account for the cohort effect. Since x_1, x_2, x_3 and x_6 are continuous covariates, they are standardised individually in our analysis.

To explore possible nonlinear effects and interaction effects of the risk factors, we include all main effects, quadratic effects and interaction effects of the risk factors and covariates

in the full model. This results in a mixed baseline hazard model with 50 covariates:

$$h_{ijk}(t, x_{ijk}) = h_{0j}(t) \exp(\beta_j^{\mathrm{T}} x_{ijk}), \tag{3·1}$$

where x_{ijk} consists of all possible linear, quadratic and interaction terms of the risk factors and covariates x_1 to x_6. Model (3·1) allows different baseline hazards and different regression coefficients for coronary heart disease and cerebrovascular accident, but an identical baseline hazards for siblings.

The maximum pseudo-partial likelihood estimate for β is computed. The logarithm of the pseudo-partial likelihood for the full model of 50 covariates is $-2017{\cdot}9590$. Next we apply the smoothly clipped absolute deviation procedure to model (3·1) to select significant variables. In the implementation of the procedure, since all covariates are important confounding variables or risk factors, we include them in the model by not penalising the linear main effects of x_1 to x_6. Thus, all linear effects are included in the selected model. The generalised crossvalidation method is used to select the regularisation parameter, giving $\lambda = 0{\cdot}9053$. The logarithm of the pseudo-partial likelihood for the model selected by the smoothly clipped absolute deviation method with the selected λ is $-2022{\cdot}6635$. This represents a decrease of 4·7045 over that of the full model, which is

Table 2. *Estimated coefficients and standard errors for the Framingham Heart Study data*

Effect	CHD $\hat{\beta}$ (SE($\hat{\beta}$))	CVA $\hat{\beta}$ (SE($\hat{\beta}$))
x_1	0·0810 (0·1288)	0·4773 (0·2423)
x_2	0·0576 (0·1200)	−0·2409 (0·2655)
x_3	0·4129 (0·1570)	0·2917 (0·1477)
x_4	0·4754 (0·2361)	0·7077 (0·3587)
x_5	−0·3666 (0·2543)	−0·1016 (0·2890)
x_6	0·0249 (0·0802)	−0·1395 (0·1916)
x_1^2	−0·0743 (0·0512)	0 (−)
x_2^2	0 (−)	−0·0768 (0·1052)
x_3^2	0 (−)	0 (−)
x_6^2	0 (−)	0·2062 (0·1229)
$x_1 * x_2$	0 (−)	0 (−)
$x_1 * x_3$	0 (−)	−0·2224 (0·1435)
$x_1 * x_4$	0·1409 (0·1495)	−0·2207 (0·2628)
$x_1 * x_5$	0 (−)	0 (−)
$x_1 * x_6$	−0·1060 (0·0808)	0 (−)
$x_2 * x_3$	0 (−)	0 (−)
$x_2 * x_4$	0·1550 (0·1425)	0·5702 (0·3766)
$x_2 * x_5$	0 (−)	0 (−)
$x_2 * x_6$	0 (−)	0 (−)
$x_3 * x_4$	−0·1952 (0·1489)	0 (−)
$x_3 * x_5$	−0·2054 (0·1378)	0 (−)
$x_3 * x_6$	0 (−)	0 (−)
$x_4 * x_5$	−0·3071 (0·3106)	0 (−)
$x_4 * x_6$	0 (−)	0 (−)
$x_5 * x_6$	0 (−)	0·5753 (0·2545)

CHD, coronary heart disease; GVA, cerebrovascular accident; SE, standard error.

much less than 25/2, half of the number of covariates excluded from the full model; see Table 2. From an extension of Theorem 3 of Cai (1999), the limiting distribution of the pseudo-partial likelihood ratio statistic is a weighted sum of χ^2_1 distributions. Based on 100 000 Monte Carlo simulations, we computed the p-value, which equals 0·9926 and which supports the selected model.

In another confirmation of the selected model, we compare it with the linear main effects model which includes only the linear main effects of x_1 to x_6. The relevant pseudo-partial likelihood ratio statistic is 23·9783. Based on 100 000 Monte Carlo simulations, the corresponding p-value equals 0·0353, indicating that the selected model fits the data better than the model with only the linear main effects.

The resulting estimates and standard errors for β in the selected model are given in Table 2. For all terms associated with x_1, x_2, x_3 and x_6, the results in Table 2 are based on the standardised variables rather than the original ones. Table 2 clearly indicates that there are a few possible quadratic effects and many interactions among the risk factors on coronary heart disease and cerebrovascular accident. It shows that subjects with higher cholesterol level have higher risk of developing coronary heart disease. There is interaction between cholesterol level and smoking status, and the hazard ratio is $\exp(0{\cdot}0576 + 0{\cdot}1550) = 1{\cdot}24$ for smokers and $\exp(0{\cdot}0576) = 1{\cdot}059$ for nonsmokers, for an increase of 3·6 mg/dL, that is one standard deviation, in cholesterol level. For a given cholesterol level x_2, the hazard ratio for smokers relative to nonsmokers is $\exp(0{\cdot}4754 + 0{\cdot}1550x_2)$.

ACKNOWLEDGEMENT

This research was supported by grants from the U.S. National Institutes of Health and National Sciences Foundation. Fan's research is also supported by a Research Grants Council grant at the Chinese University of Hong Kong. The authors thank the editor and referees for their helpful comments which greatly improved the paper.

APPENDIX

Theory

Let $N_{ijk}(t) = I(Z_{ijk} \leqslant t, \delta_{ijk} = 1)$ be the counting process, and let $h_{ijk}(t)$ and

$$M_{ijk}(t) = N_{ijk}(t) - \int_0^t Y_{ijk}(u) h_{ijk}(u)\, du$$

be their corresponding marginal hazards function and marginal martingale, respectively, with respect to the filtration $\mathscr{F}_{jk}(t^-)$, where $\mathscr{F}_{jk}(t)$ is the σ-field generated by

$$\{N_{ijk}(u),\, Y_{i11}(u), \ldots,\, Y_{iJK}(u),\, x_{i11}(u), \ldots,\, x_{iJK}(u);\, 0 \leqslant u \leqslant t,\, i = 1, \ldots, n\}.$$

Here, without loss of generality, we take $K_i = K$. Define

$$S^{(d)}_{jk}(\beta; t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ijk}(t) x_{ijk}(t)^{\otimes d} \exp\{\beta^{\mathrm{T}} x_{ijk}(t)\} \quad (d = 0, 1, 2),$$

$$S^{(d)}_{j\cdot}(\beta; t) = \sum_{k=1}^{K} S^{(d)}_{jk}(\beta; t) \quad (d = 0, 1, 2),$$

$$E_j(\beta; t) = S^{(1)}_{j\cdot}(\beta; t)/S^{(0)}_{j\cdot}(\beta; t), \quad V_j(\beta; t) = S^{(2)}_{j\cdot}(\beta, t)/S^{(0)}_{j\cdot}(\beta, t) - E_j(\beta; t)^{\otimes 2},$$

where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^{\mathrm{T}}$ for a vector a.

We require the following regularity conditions.

Condition A1. For simplicity, assume that T_{ijk} takes values on a finite interval $[0, \tau]$, and that $\int_0^\tau h_{0j}(t)\,dt < \infty$ for $j = 1, \ldots, J$.

Condition A2. There exists a neighbourhood $\mathscr{B}$ of the true value β_0 that satisfies each of the following conditions: (i) there exists a scalar, a vector and a matrix function $s_{jk}^{(d)}(\beta, t)\,(d = 0, 1, 2)$ defined on $\mathscr{B} \times [0, \tau]$ such that $\sup_{t \in [0,\tau], \beta \in \mathscr{B}} \|S_{jk}^{(d)}(\beta, t) - s_{jk}^{(d)}(\beta, t)\| \to 0$ in probability; (ii) there exists a matrix $\Gamma = \Gamma(\beta)$ such that $\|(1/n)\sum_{i=1}^n \mathrm{var}(D_i) - \Gamma\| \to 0$, where

$$D_i = \sum_{j=1}^{J} \sum_{k=1}^{K} \int_0^\tau \{x_{ijk}(t) - e_j(\beta_0; t)\} dM_{ijk}(t),$$

and $e_j(\beta; t) = \{\sum_{k=1}^K s_{jk}^{(1)}(\beta; t)\}/\{\sum_{k=1}^K s_{jk}^{(0)}(\beta; t)\}$. Assume further that there exist constants C_1 and C_2, such that

$$0 < C_1 < \lambda_{\min}(\Gamma) \leqslant \lambda_{\max}(\Gamma) < C_2 < \infty \tag{A·1}$$

for all n, where $\lambda_{\min}(\Gamma)$ and $\lambda_{\max}(\Gamma)$ stand for the minimal and maximal eigenvalues of Γ, respectively.

Condition A3. Using the notation in Condition A2, define

$$v_j = \frac{\sum_{k=1}^K s_{jk}^{(2)}(\beta, t)}{\sum_{k=1}^K s_{jk}^{(0)}(\beta, t)} - e_j(\beta; t)^{\otimes 2}.$$

Then, for all $\beta \in \mathscr{B}$, $t \in [0, \tau]$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$, define $s_{jk}^{(1)}(\beta, t) = \partial s_{jk}^{(0)}(\beta; t)/\partial\beta$ and $s_{jk}^{(2)}(\beta; t) = \partial s_{jk}^{(1)}(\beta; t)/\partial\beta$. Assume that $s_{jk}^{(0)}(\beta; t)$ is bounded away from 0 on $\mathscr{B} \times [0, \tau]$. Let $A(\beta) = \sum_j \int_0^\tau v_j(\beta; t) \sum_k s_{jk}^{(0)}(\beta; t) h_{0j}(t)\,dt$, and assume that there exist positive constants C_3 and C_4 such that

$$0 < C_3 < \lambda_{\min}\{A(\beta_0)\} \leqslant \lambda_{\max}\{A(\beta_0)\} < C_4 < \infty, \tag{A·2}$$

for all n.

Condition A4. There exists a constant C_5 such that $\sup_{1 \leqslant i \leqslant n} E D_{ik}^2 D_{il}^2 \leqslant C_5 < \infty$ for all $1 \leqslant k, l \leqslant d_n$.

Condition A5. Assume that $p_{\lambda_j}(|\beta_j|)$ satisfies $\liminf_{n\to\infty} \liminf_{\beta_j \to 0+} p'_{\lambda_j}(\beta_j)/\lambda_j > 0$, for all $j = 1, \ldots, d_n$. Assume further that there exists a constant C_6 such that, for nonzero θ_1 and θ_2, $|p''_{\lambda_j}(\theta_1) - p''_{\lambda_j}(\theta_2)| \leqslant C_6|\theta_1 - \theta_2|$, for all $j = 1, \ldots, d_n$.

Proof of Theorem 1. *Let* $\alpha_n = \sqrt{(d_n)}(n^{-1/2} + a_n)$. To prove Theorem 1, it is sufficient to show that, for any given $\varepsilon > 0$, there exists a large constant C such that

$$\mathrm{pr}\left\{ \sup_{\|u\| = C} \mathscr{L}(\beta_0 + \alpha_n u) < \mathscr{L}(\beta_0) \right\} \geqslant 1 - \varepsilon. \tag{A·3}$$

This implies that there exists a local maximiser such that $\|\hat\beta - \beta_0\| = O_p(\alpha_n)$.

Note that $p_\lambda(0) = 0$ and $p_\lambda(.) \geqslant 0$. It follows by Taylor expansion that

$$\begin{aligned} \mathscr{L}(\beta_0 + \alpha_n u) - \mathscr{L}(\beta_0) &\leqslant \{\ell(\beta_0 + \alpha_n u) - \ell(\beta_0)\} - n \sum_{j=1}^{s_n} \{p_{\lambda_{jn}}(|\beta_{j0} + \alpha_n u|) - p_{\lambda_{jn}}(|\beta_{j0}|)\} \\ &= I_1 + I_2, \end{aligned}$$

say. We first consider I_1. It follows by Taylor expansion that

$$I_1 = \alpha_n u^{\mathrm{T}} \ell'(\beta_0) + \frac{\alpha_n^2}{2} u^{\mathrm{T}} \ell''(\beta_n^*) u = I_{11} + I_{12},$$

say, where β_n^* lies between β_0 and $\beta_0 + \alpha_n u$. By the Cauchy–Schwartz inequality, it follows that

$$I_{11} = \alpha_n u^{\mathrm{T}} \ell'(\beta_0) \leqslant \alpha_n \|\ell'(\beta_0)\|\, \|u\| = O_P\{\alpha_n \sqrt{(n d_n)}\}\, \|u\| = O_P(n\alpha_n^2)\|u\|.$$

We next deal with I_{12}. By the Chebyshev inequality, we can show that

$$\mathrm{pr}\{\|n^{-1}\ell''(\beta)+A(\beta)\| \geqslant \varepsilon d_n^{-1}\} \leqslant \frac{d_n^4}{n\varepsilon^2} = o(1),$$

as $d_n^4/n \to 0$ by assumption. Thus,

$$\|n^{-1}\ell''(\beta)+A(\beta)\| = o_P(d_n^{-1}), \tag{A·4}$$

in probability, uniformly in $\beta \in \mathscr{B}$. Hence $I_{12} = -\frac{1}{2}n\alpha_n^2 u^{\mathrm{T}} A(\beta_0)u\{1+o_P(1)\}$. By the assumption that $\lambda_{\min}\{A(\beta_0)\} \geqslant C_1 > 0$, I_{12} dominates I_{11} uniformly in $\|u\| = C$ for a sufficiently large C. The proof is completed by showing that I_{12} also dominates I_2 uniformly in $\|c\| = C$ for a sufficiently large constant C. To this end, from the Taylor expansion of I_2, it can be shown by the Cauchy–Schwarz inequality that I_2 is dominated by $\alpha_n^2\|u\| + 2b_n n\alpha_n^2\|u\|^2$. Since $b_n \to 0$, I_{12} dominates I_2 if we choose a sufficiently large C. Thus, (A·3) holds. □

The following lemma shows that the penalised pseudo-partial likelihood estimator must possess the sparsity property $\hat{\beta}_2 = 0$. Its proof is given in the authors' technical report.

LEMMA A1. *Under the conditions of Theorem* 2, *with probability tending to* 1, *for any given* β_1 *satisfying* $\|\beta_1 - \beta_{10}\| = O_P\{\sqrt{(d_n/n)}\}$ *and any constant* C, *it holds that*

$$\mathscr{L}\{(\beta_1^{\mathrm{T}}, 0)^{\mathrm{T}}\} = \max_{\|\beta_2\| \leqslant C\sqrt{(d_n/n)}} \mathscr{L}\{(\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}\}. \tag{A·5}$$

Proof of Theorem 2. Part (i) immediately follows by Lemma A1. We next prove the asymptotic normality of $\hat{\beta}_1$. As shown in our technical report, it holds that

$$\ell_1'(\beta_0) - n(A_{11}+\Sigma)\{\hat{\beta}_1 - \beta_{10} + (A_{11}+\Sigma)^{-1}b\} + o_p(\sqrt{n}) = 0, \tag{A·6}$$

where $\ell_1'(\beta_0)$ consists of the first s_n components of $\ell'(\beta_0)$, and A_{11} is the first $s_n \times s_n$ upper-left submatrix of $A(\beta_0)$. Therefore,

$$\sqrt{n}c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}(A_{11}+\Sigma)\{\hat{\beta}_1 - \beta_{10} + (A_{11}+\Sigma)^{-1}b\} = n^{-1/2}c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}\ell_1'(\beta_0) + o_p(1).$$

We now show the asymptotic normality of $n^{-1/2}c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}\ell_1'(\beta_0)$, because then the asymptotic normality of β_1 in (2·7) can be established by Slutsky's theorem.

It can be shown by similar arguments to those in Andersen & Gill (1982) and Clegg et al. (1999) that

$$n^{-1/2}c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}\ell_1'(\beta_0) = n^{-1/2}\sum_{i=1}^{n} c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}D_{i1} + o_P(1),$$

where D_{i1} consists of the first s_n components of D_i. Let

$$Y_{ni} = n^{-1/2}c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}D_{i1}.$$

Next we verify the Lindeberg condition for Y_{ni}. By Condition A2 and the Chebyshev inequality, we obtain

$$\begin{aligned}\sum_{i=1}^{n}\mathrm{pr}(|Y_{ni}| > \varepsilon) &\geqslant (n\varepsilon^2)^{-1}\sum_{i=1}^{n} Ec_n^{\mathrm{T}}\Gamma_{11}^{-1/2}D_{i1}D_{i1}^{\mathrm{T}}\Gamma_{11}^{-1/2}c_n \\ &= \varepsilon^{-2}c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(D_{i1})\right\}\Gamma_{11}^{-1/2}c_n = O(1).\end{aligned}$$

By the Cauchy–Schwarz inequality and because $c_n^{\mathrm{T}}c_n = 1$,

$$\sum_{i=1}^{n} EY_{ni}^4 = \frac{1}{n^2}\sum_{i=1}^{n} E(c_n^{\mathrm{T}}\Gamma_{11}^{-1/2}D_{i1})^4 \leqslant \frac{1}{n^2}\lambda_{\max}^2(\Gamma_{11}^{-1})\sum_{i=1}^{n} E\|D_{i1}\|^4.$$

By Condition A4, $E\|D_{i1}\|^4 = O(d_n^2)$ and, by (A·1), $\lambda_{\max}(\Gamma_{11}^{-1}) = \lambda_{\min}^{-1}(\Gamma_{11}) \leqslant \lambda_{\min}^{-1}(\Gamma) \leqslant C_1^{-1}$. It follows from the Cauchy–Schwartz inequality that, for any ε, $\sum_{i=1}^n EY_{ni}^2 I\{|Y_{ni}| > \varepsilon\} = O(d_n/\sqrt{n}) = o(1)$. On the other hand, as $c_n^{\mathrm{T}} c_n = 1$,

$$\sum_{i=1}^n \operatorname{cov}(Y_{ni}) = \frac{1}{n} \sum_{i=1}^n c_n^{\mathrm{T}} D_{11}^{-1} \operatorname{cov}(D_{i1}) D_{11}^{-1} c_n \to 1.$$

Thus, Y_{ni} satisfies the conditions of the Lindeberg–Feller central limit theorem. This also means that $n^{-1/2} \sum_{i=1}^n c_n^{\mathrm{T}} \Gamma_{11}^{-1/2} D_{i1}$ is asymptotically $N(0, 1)$ since D_{i1} has zero mean. By Slutsky's Theorem,

$$n^{-1/2} c_n^{\mathrm{T}} D_{11}^{-1/2} \ell_1'(\beta_0) \to N(0, 1)$$

in distribution as $n \to \infty$. This completes the proof. □

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.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–83.

Bunea, F. & McKeague, I. W. (2005). Covariate selection for semiparametric hazard function regression models. *J. Mult. Anal.* **92**, 186–204.

Cai, J. (1999). Hypothesis testing of hazard ratio parameters in marginal models for multivariate failure time data. *Lifetime Data Anal.* **5**, 39–53.

Cai, J. & Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151–64.

Cai, J. & Prentice, R. L. (1997). Regression estimation using multivariate time data and a common baseline hazard function model. *Lifetime Data Anal.* **3**, 197–213.

Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with Discussion). *J. R. Statist. Soc.* A **148**, 82–117.

Clegg, L. X., Cai, J. & Sen, P. K. (1999). A marginal mixed baseline hazards model for multivariate failure time data. *Biometrics* **55**, 805–12.

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

Dawber, T. R. (1980). *The Framingham Study, The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalised 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.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–75.

Huang, J. & Harrington, D. (2002). Penalised partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. *Biometrics* **58**, 781–91.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Lee, E. W., Wei, L. J. & Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, Ed. J. P. Klein and P. Goel, pp. 237–48. Boston: Kluwer Academic Publishers.

Liang, K.-Y., Self, S. G. & Chang, Y.-C. (1993). Modelling marginal hazards in multivariate failure time data. *J. R. Statist. Soc.* B **55**, 441–53.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statist. Med.* **13**, 2233–47.

Oakes, D. (1989). Bivariate survival models induced by frailty. *J. Am. Statist. Assoc.* **84**, 487–92.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356–66.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
SHIBATA, R. (1984). Approximation efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43–9.
SPIEKERMAN, C. F. & LIN, D. Y. (1998). Marginal regression models for multivariate failure time data. *J. Am. Statist. Assoc.* **93**, 1164–75.
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.
WEI, L. J., LIN, D. Y. & WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *J. Am. Statist. Assoc.* **84**, 1065–73.

[*Received March* 2004. *Revised October* 2004]