Biometrika (2016), **103**, 3, *pp.* 547–562
Printed in Great Britain
doi: 10.1093/biomet/asw027
Advance Access publication 10 August 2016

Variable selection for case-cohort studies with failure time outcome

By AI NI, JIANWEN CAI and DONGLIN ZENG

3101 McGavran-Greenberg Hall, CB 7420, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

aini01@gmail.com cai@bios.unc.edu dzeng@email.unc.edu

Summary

Case-cohort designs are widely used in large cohort studies to reduce the cost associated with covariate measurement. In many such studies the number of covariates is very large, so an efficient variable selection method is necessary. In this paper, we study the properties of a variable selection procedure using the smoothly clipped absolute deviation penalty in a case-cohort design with a diverging number of parameters. We establish the consistency and asymptotic normality of the maximum penalized pseudo-partial-likelihood estimator, and show that the proposed variable selection method is consistent and has an asymptotic oracle property. Simulation studies compare the finite-sample performance of the procedure with tuning parameter selection methods based on the Akaike information criterion and the Bayesian information criterion. We make recommendations for use of the proposed procedures in case-cohort studies, and apply them to the Busselton Health Study.

Some key words: Case-cohort design; Diverging number of parameters; Oracle property; Smoothly clipped absolute deviation; Survival analysis; Variable selection.

1. Introduction

Large-scale epidemiological studies and disease prevention trials often follow thousands of subjects for long periods of time. Measuring covariates for the entire study cohort can be prohibitively expensive, especially when it involves taking biological samples or performing expensive bioassays. Moreover, the rate of occurrence of the event of interest, such as cardiovascular disease, stroke or death, is typically low in such studies. We refer to subjects who develop the event of interest during the study as cases and the other subjects as noncases. If the covariates were to be measured for everyone in the study, most of the cost would be spent on the noncases, who do not contribute as much information as the cases. To reduce the cost and effort in collecting expensive covariates while losing as little efficiency as possible, Prentice (1986) proposed the case-cohort design, where complete covariate information is obtained from only a random subcohort of the sample, as well as from all of the cases.

Various estimation methods have been developed for case-cohort studies under the proportional hazards model (Cox, 1972). Prentice (1986) and Self & Prentice (1988) proposed a pseudo-partial-likelihood method that modifies the risk set to account for subcohort sampling. Barlow (1994) introduced a time-dependent weight to estimate the risk set from the subcohort sample and developed a robust variance estimator for the regression parameters. Kalbfleisch & Lawless (1988) proposed a more efficient weighting that uses the complete covariate history of all cases. Borgan et al. (2000) further studied several types of weights under the stratified case-cohort design. Kulich & Lin (2004) established the asymptotic properties of the efficiently weighted estimator of Kalbfleisch & Lawless (1988); Kang & Cai (2009) extended this estimator to studies with multivariate failure time outcomes, and Kim et al. (2013) further improved the estimator's efficiency in the presence of multivariate failure time outcomes. In this paper, we focus on the efficient weighting proposed by Kalbfleisch & Lawless (1988) in a univariate unstratified case-cohort design.

In large epidemiological studies that use the case-cohort design, many covariates are usually collected, and often one goal of the research is to identify a subset of covariates related to the event of interest. With the inclusion of interactions and polynomial terms, the number of candidate covariates can be very large. As Huber (1973) argued, in the context of variable selection, the number of parameters should be considered as increasing to infinity with the sample size n. In this paper, we consider the scenario where the model size d_n diverges to infinity but at a slower rate than the sample size. Traditional variable selection methods such as stepwise and best subset selection are computationally intensive and unstable. Since the introduction of the lasso by Tibshirani (1996), penalty-based variable selection procedures have achieved great success. Under certain regularity conditions, these methods can simultaneously select variables and estimate their coefficients. Many penalty functions have been proposed, among which the smoothly clipped absolute deviation (Fan & Li, 2001), adaptive lasso (Zou, 2006), adaptive elastic net (Zou & Zhang, 2009) and minimax concave (Zhang, 2010) penalties have been shown to possess the oracle property, namely, as $n \to \infty$ the procedure correctly identifies the true model with probability tending to unity and estimates the standard errors of nonzero parameters as efficiently as if the true model were known. Fan & Li (2002) applied the smoothly clipped absolute deviation penalty to the proportional hazards model and proved its oracle property. Cai et al. (2005) extended the penalized partial likelihood procedure to multivariate models with a diverging number of parameters. However, to the best of our knowledge, the properties of penalized variable selection have not been studied under the case-cohort design where not all covariates are fully observed.

2. Pseudo-partial likelihood for case-cohort designs

Suppose there are n independent subjects in a cohort. Let $Z_i(t)$ be the $d_n \times 1$, possibly time-dependent, covariate vector for subject i at time t. Since d_n goes to infinity with n, all quantities that are functions of the covariates depend on n. For notational simplicity, however, we shall suppress the subscript n. Without loss of generality, we partition the real-valued true parameter vector β_{n0} as $(\beta_{n0,\mathrm{I}}^{\mathrm{T}}, \beta_{n0,\mathrm{II}}^{\mathrm{T}})^{\mathrm{T}}$, where $\beta_{n0,\mathrm{I}}$ and $\beta_{n0,\mathrm{II}}$ are the nonzero and zero components of β_{n0}, respectively. Denote by k_n the dimension of $\beta_{n0,\mathrm{I}}$, which is allowed to diverge with n in such a way that k_n/d_n converges to a constant $c \in [0, 1]$.

Let T and C be, respectively, the time to the outcome of interest and the censoring time. Let $X = \min(T, C)$ be the observed time and let $\Delta = I(T \leqslant C)$ be the censoring indicator, where $I(\cdot)$ denotes the indicator function. We assume that T and C are independent, conditional on Z. Define for subject i the counting process $N_i(t) = I(X_i \leqslant t, \Delta_i = 1)$ and the at-risk process $Y_i(t) = I(X_i \geqslant t)$. Let $\lambda_i(t)$ denote the hazard function for subject i. Cox (1972) proposed the proportional hazards model $\lambda_i\{t \mid Z_i(t)\} = \lambda_0(t)\exp\{\beta^{\mathrm{T}} Z_i(t)\}$, in which $\lambda_0(t)$ is an unspecified baseline hazard function.

Under the case-cohort design, suppose that we randomly select a subcohort of fixed size $\tilde{n}$ from the full cohort. Let ξ_i denote the indicator of the ith subject being selected into the subcohort, and let $\alpha = \tilde{n}/n = \mathrm{pr}(\xi_i = 1) \in (0, 1]$ be the selection probability for the ith subject. Here we consider simple random sampling without replacement. Under this sampling scheme, $(\xi_1, \ldots, \xi_n)$ are correlated. The covariate histories are not observed for censored subjects outside the subcohort.

If complete covariate histories are available for all the cases, one can use the following pseudo-partial likelihood to estimate the regression coefficients β (Kalbfleisch & Lawless, 1988):

$$\tilde{\ell}_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta^{\mathrm{T}} Z_i(t) - \log \sum_{j=1}^{n} \rho_j(t) Y_j(t) \exp\{\beta^{\mathrm{T}} Z_j(t)\} \right] \mathrm{d}N_i(t), \quad (1)$$

where τ is the time at the end of the study and $\rho_i(t) = \Delta_i + (1 - \Delta_i)\xi_i \hat{\alpha}^{-1}(t)$, with $\hat{\alpha}(t) = \sum_{i=1}^{n}(1 - \Delta_i)\xi_i Y_i(t)/\{\sum_{i=1}^{n}(1 - \Delta_i)Y_i(t)\}$ being a time-dependent estimator of the true sampling probability α. The corresponding pseudo-partial score equation is

$$\tilde{\ell}_n'(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \frac{\tilde{S}^{(1)}(\beta, t)}{\tilde{S}^{(0)}(\beta, t)} \right\} \mathrm{d}N_i(t) = 0,$$

where $\tilde{S}^{(k)}(\beta, t) = n^{-1} \sum_{i=1}^{n} \rho_i(t) Y_i(t) Z_i(t)^{\otimes k} \exp\{\beta^{\mathrm{T}} Z_i(t)\}$ for $k = 0, 1, 2$. Here $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^{\mathrm{T}}$ for a vector a.

3. Variable selection with a penalized pseudo-partial likelihood

3·1. *Penalized pseudo-partial likelihood*

We define a penalized pseudo-partial likelihood as

$$\tilde{Q}_n(\beta) = \tilde{\ell}_n(\beta) - n \sum_{j=1}^{d_n} P_{\lambda_{nj}}(|\beta_j|), \quad (2)$$

where $P_{\lambda_{nj}}(|\beta_j|)$ is a nonnegative penalty function with $P_{\lambda_{nj}}(0) = 0$. The nonnegative tuning parameter λ_{nj} controls the model complexity. We use the smoothly clipped absolute deviation penalty (Fan & Li, 2001) with covariate-specific tuning parameters λ_{nj}, which allows different regression coefficients to have different penalty functions. The smoothly clipped absolute deviation penalty is

$$P_{\lambda_{nj}}(\theta) = \begin{cases} \lambda_{nj}\theta, & \theta \leqslant \lambda_{nj}, \\ -\dfrac{\theta^2 - 2a\lambda_{nj}\theta + \lambda_{nj}^2}{2(a-1)}, & \lambda_{nj} < \theta \leqslant a\lambda_{nj}, \\ \dfrac{(a+1)\lambda_{nj}^2}{2}, & \theta > a\lambda_{nj}, \end{cases}$$

for some $a > 2$ and $\theta > 0$. The first derivative of the penalty is

$$P'_{\lambda_{nj}}(\theta) = \lambda_{nj} I(\theta \leqslant \lambda_{nj}) + \frac{(a\lambda_{nj} - \theta)_+}{a-1} I(\theta > \lambda_{nj}).$$

3·2. *Regularity conditions*

For each n, we define

$$S_n^{(k)}(\beta_n, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) Z_i(t)^{\otimes k} \exp\{\beta_n^{\mathrm{T}} Z_i(t)\} \quad (k = 0, 1, 2),$$

$$s_n^{(k)}(\beta_n, t) = E\{S_n^{(k)}(\beta_n, t)\} \quad (k = 0, 1, 2),$$

$$e_n(\beta_n, t) = s_n^{(1)}(\beta_n, t)/s_n^{(0)}(\beta_n, t),$$

$$V_n(\beta_n, t) = \frac{S_n^{(2)}(\beta_n, t)S_n^{(0)}(\beta_n, t) - S_n^{(1)}(\beta_n, t)^{\otimes 2}}{S_n^{(0)}(\beta_n, t)^2},$$

$$\tilde{V}_n(\beta_n, t) = \frac{\tilde{S}_n^{(2)}(\beta_n, t)\tilde{S}_n^{(0)}(\beta_n, t) - \tilde{S}_n^{(1)}(\beta_n, t)^{\otimes 2}}{\tilde{S}_n^{(0)}(\beta_n, t)^2},$$

$$I_n(\beta_n) = E\left\{\int_0^\tau V_n(\beta_n, t)S_n^{(0)}(\beta_n, t)\,\mathrm{d}\Lambda_0(t)\right\},$$

$$\Gamma_n(\beta_n) = \mathrm{var}\left\{n^{-1/2}\tilde{\ell}_n'(\beta_n)\right\}.$$

We require the following regularity conditions:

Condition 1. $\int_0^\tau \lambda_0(t)\,\mathrm{d}t < \infty$ and $E\{Y(\tau)\} > 0$;

Condition 2. $|Z_{ij}(0)| + \int_0^\tau |\mathrm{d}Z_{ij}(t)| < C_1 < \infty$ almost surely for some constant C_1, for $i = 1, \ldots, n$ and $j = 1, \ldots, d_n$;

Condition 3. there exists a neighbourhood $\mathcal{B}_n$ of β_{n0} such that for all $\beta_n \in \mathcal{B}_n$ and $t \in [0, \tau]$, $\partial s_n^{(0)}(\beta_n, t)/\partial\beta_n = s_n^{(1)}(\beta_n, t)$ and $\partial^2 s_n^{(0)}(\beta_n, t)/(\partial\beta_n\partial\beta_n^{\mathrm{T}}) = s_n^{(2)}(\beta_n, t)$. The functions $s_n^{(k)}(\beta_n, t)$ $(k = 0, 1, 2)$ are continuous and bounded, and $s_n^{(0)}(\beta_n, t)$ is bounded away from zero on $\mathcal{B}_n \times [0, \tau]$;

Condition 4. there exist positive constants C_2, C_3, C_4 and C_5 such that

$$0 < C_2 < \lambda_{\min}\{I_n(\beta_{n0})\} \leqslant \lambda_{\max}\{I_n(\beta_{n0})\} < C_3 < \infty,$$
$$0 < C_4 < \lambda_{\min}\{\Gamma_n(\beta_{n0})\} \leqslant \lambda_{\max}\{\Gamma_n(\beta_{n0})\} < C_5 < \infty,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues of a matrix;

Condition 5. $\min_{1\leqslant j\leqslant k_n} |\beta_{nj0}|/\lambda_{nj} \to \infty$ as $n \to \infty$;

Condition 6. $\liminf_{n\to\infty} \liminf_{\theta\to 0+} P'_{\lambda_{nj}}(\theta)/\lambda_{nj} > 0$ for $j = 1, \ldots, d_n$.

Condition 1 guarantees a finite baseline cumulative hazard and a nonempty risk set at the end of the study. Condition 2 requires the stochastic process of each time-dependent covariate to have bounded variation almost surely. Condition 3 essentially requires $\exp\{\beta_n^{\mathrm{T}} Z_i(t)\}$ to be integrable under a diverging dimension so that integration and differentiation with respect to $S_n^{(k)}(\beta_n, t)$ $(k = 0, 1)$ can be interchanged. Condition 4 ensures that the covariance matrices of the score function under both regular and case-cohort designs are positive definite and have uniformly bounded eigenvalues for all n; it assumes a nonsingular Hessian matrix of the objective function used for variable selection. The same condition has been assumed in other works on variable selection (Peng & Fan, 2004; Cai et al., 2005; Cho & Qu, 2013). Condition 5 specifies the rate at which the proposed procedure can distinguish nonzero parameters from zero ones. As $n \to \infty$, the size of nonzero parameters detectable by the procedure can approach zero, but at a slower rate than the tuning parameter. This condition is required for the derivation of the asymptotic properties of the proposed procedure, and has been assumed by many authors (e.g., Peng & Fan, 2004; Wang et al., 2009; Cho & Qu, 2013; Fan & Tang, 2013). In real-world biomedical research, there usually exists a fixed minimum clinically important effect size. Any effect smaller than this size can effectively be treated as zero. Thus, Condition 5 is a reasonable requirement. Condition 6 implies that those zero parameters whose finite-sample estimates are of about the same scale as the λ_{nj} will automatically be shrunk to zero; this helps to establish the oracle property of variable selection.

3·3. *Asymptotic properties*

Throughout this paper we use $O_{\rm p}(\cdot)$ and $o_{\rm p}(\cdot)$ to denote probability order relations and $O(\cdot)$ and $o(\cdot)$ to denote almost-sure order relations. Let $a_n = \max_{1\leqslant j\leqslant k_n}\{|P'_{\lambda_{nj}}(|\beta_{nj0}|)|\}$ and $b_n = \max_{1\leqslant j\leqslant k_n}\{|P''_{\lambda_{nj}}(|\beta_{nj0}|)|\}$. We first prove the existence of a penalized pseudo-partial-likelihood estimator that converges at rate $O_{\rm p}\{d_n^{1/2}(n^{-1/2}+a_n)\}$ and then establish its oracle property. The proofs of Theorems 1 and 2 are provided in the Appendix.

THEOREM 1. *Under Conditions 1–5, if $b_n \to 0$ and $d_n^4/n \to 0$ as $n\to\infty$, then with probability tending to 1 there exists a local maximizer $\hat\beta_n$ of $\tilde Q_n(\beta_n) = \tilde\ell_n(\beta_n) - n\sum_{j=1}^{d_n} P_{\lambda_{nj}}(|\beta_{nj}|)$ such that $\|\hat\beta_n - \beta_{n0}\| = O_{\rm p}\{d_n^{1/2}(n^{-1/2}+a_n)\}$.*

From Theorem 1 one can obtain a $(n/d_n)^{1/2}$-consistent penalized pseudo-partial-likelihood estimator, provided that $a_n = O(n^{-1/2})$, which is the case for the smoothly clipped absolute deviation penalty under Condition 5. This consistency rate is the same as that of the maximum likelihood estimator for the exponential family (Portnoy, 1988). For the next theorem, we define

$$\Sigma_n = {\rm diag}\{P''_{\lambda_{1n}}(|\beta_{n01}|),\ldots,P''_{\lambda_{k_n n}}(|\beta_{n0k_n}|)\}, \tag{3}$$

$$B_n = \{P'_{\lambda_{1n}}(|\beta_{n01}|)\,{\rm sgn}(\beta_{n01}),\ldots,P'_{\lambda_{k_n n}}(|\beta_{n0k_n}|)\,{\rm sgn}(\beta_{n0k_n})\}^{\rm T}. \tag{4}$$

THEOREM 2. *Under Conditions 1–6, if $b_n\to 0$, $d_n^5/n\to 0$, $\lambda_{nj}\to 0$, $\lambda_{nj}(n/d_n)^{1/2}\to\infty$ and $a_n = O(n^{-1/2})$ as $n\to\infty$, the $(n/d_n)^{1/2}$-consistent local maximizer $\hat\beta_n = (\hat\beta_{n,\rm I}^{\rm T}, \hat\beta_{n,\rm II}^{\rm T})^{\rm T}$ must be such that $\hat\beta_{n,\rm II}=0$ with probability tending to unity and, for any nonzero $k_n\times 1$ constant vector u_n with $\|u_n\|=1$,*

$$n^{1/2}u_n^{\rm T}\Gamma_{n11}^{-1/2}(I_{n11}+\Sigma_n)\{\hat\beta_{n,\rm I}-\beta_{n0,\rm I}+(I_{n11}+\Sigma_n)^{-1}B_n\}\to N(0,1)$$

in distribution, where Σ_n and B_n are defined in (3) and (4), respectively, I_{n11} consists of the first $k_n\times k_n$ components of $I_n(\beta_{n0})$, and Γ_{n11} consists of the first $k_n\times k_n$ components of $\Gamma_n(\beta_{n0})$.

Because of the diverging dimension of $\beta_{n0,\rm I}$, Theorem 2 establishes the asymptotic normality of some linear combination of standardized estimators. However, by choosing a particular u_n, it can give the asymptotic distribution of each individual estimator. Thus, it provides a theoretical basis for inference on individual coefficients. The matrix $I_n(\beta_{n0})$ can be consistently estimated by $\hat I_n(\hat\beta_n) = n^{-1}\sum_{i=1}^n\int_0^\tau \tilde V_n(\hat\beta_n,t)\,{\rm d}N_i(t)$. The estimator of the matrix $\Gamma_n(\beta_{n0})$ is given in the Supplementary Material. For the smoothly clipped absolute deviation penalty, $a_n=0$, $\Sigma_n=0$ and $B_n=0$ for large n under Condition 5. Therefore, the result of Theorem 2 reduces to

$$n^{1/2}u_n^{\rm T}\Gamma_{n11}^{-1/2}I_{n11}(\hat\beta_{n,\rm I}-\beta_{n0,\rm I})\to N(0,1)$$

in distribution as $n\to\infty$. The conditions $d_n^4/n\to 0$ and $d_n^5/n\to 0$ in the above theorems describe the divergence rate of d_n relative to the sample size; they do not impose any one-to-one relationship between finite d_n and n.

4. Practical implementation considerations

4·1. *Local quadratic approximation and variance estimation*

Since the smoothly clipped absolute deviation penalty function is not differentiable at the origin, in practical implementations the Newton–Raphson algorithm cannot be applied directly to maximize (2). Instead, we use a modified Newton–Raphson algorithm with a local quadratic approximation to the penalty function. The unpenalized pseudo-partial likelihood (1) can be seen as a special case of the penalized pseudo-partial likelihood (2) with $P_{\lambda_{nj}}(|\beta_{nj}|) = 0$ for all $j = 1, \ldots, d_n$. Applying Theorem 1 with $\lambda_{nj} = 0$ for all $j = 1, \ldots, d_n$, we know there exists a $(n/d_n)^{1/2}$-consistent maximizer of (1). The concavity of (1) ensures that the maximizer is unique. We use this maximizer as the initial value $\beta_n^{(0)}$ for the modified Newton–Raphson algorithm. If $|\beta_{nj}^{(0)}|$ is less than a prespecified small positive constant c_j, then we set $\hat{\beta}_{nj} = 0$. Otherwise, the penalty function is locally approximated by a quadratic function, $P_{\lambda_{nj}}(|\beta_{nj}|) \approx P_{\lambda_{nj}}\{|\beta_{nj}^{(0)}|\} + P'_{\lambda_{nj}}\{|\beta_{nj}^{(0)}|\}\{2|\beta_{nj}^{(0)}|\}^{-1}[\beta_{nj}^2 - \{\beta_{nj}^{(0)}\}^2]$, which has the same value and first derivative as the original penalty at $\beta_{nj}^{(0)}$. It follows that $P'_{\lambda_{nj}}(|\beta_{nj}|) \approx [P'_{\lambda_{nj}}\{|\beta_{nj}^{(0)}|\}/|\beta_{nj}^{(0)}|]\beta_{nj}$. This approximation is local in the sense that it is valid only in a neighbourhood of $\beta_{nj}^{(0)}$. With the approximated penalty function, one Newton–Raphson step is performed and the updated nonzero estimate is used as the new initial value. The process is iterated until convergence or until all parameters are estimated as zero. Hunter & Li (2005) showed that the local quadratic approximation is an extension of the expectation-maximization algorithm and has the same properties.

The sandwich estimate of the covariance matrix for $\hat{\beta}_n$ can be obtained directly from the last iteration of the above algorithm as $\widehat{\text{cov}}(\hat{\beta}_n) = \{\tilde{\ell}''_n(\hat{\beta}_n) - n\Sigma_\lambda(\hat{\beta}_n)\}^{-1} n\hat{\Gamma}_n(\hat{\beta}_n)\{\tilde{\ell}''_n(\hat{\beta}_n) - n\Sigma_\lambda(\hat{\beta}_n)\}^{-1}$, where $\Sigma_\lambda(\beta_n) = \text{diag}\{P'_{\lambda_{1n}}\{|\beta_{n1}^{(0)}|\}/|\beta_{n1}^{(0)}|, \ldots, P'_{\lambda_{d_n n}}\{|\beta_{nd_n}^{(0)}|\}/|\beta_{nd_n}^{(0)}|\}$. The sandwich estimate of the covariance matrix is applicable only to the nonzero parameter estimates.

4·2. *Selection of tuning parameters*

The tuning parameter λ in the smoothly clipped absolute deviation penalty function $P_\lambda(\cdot)$ controls the magnitude of the penalty on each regression coefficient and thereby controls the complexity of the selected model. Typical methods of selecting tuning parameters include data-driven procedures such as K-fold crossvalidation and generalized crossvalidation (Craven & Wahba, 1979). We follow Fan & Li (2002) and Cai et al. (2005) and use generalized crossvalidation. The effective number of parameters measures the degrees of freedom in a regularized regression model (Hastie et al., 2009). For the proportional hazards model, the effective number of parameters is defined as $e(\lambda_{1n}, \ldots, \lambda_{d_n n}) = \text{tr}\,[\{\tilde{\ell}''_n(\hat{\beta}_n) - n\Sigma_\lambda(\hat{\beta}_n)\}^{-1}\tilde{\ell}''_n(\hat{\beta}_n)]$ (Fan & Li, 2002). The generalized crossvalidation statistic is defined as

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

which is guaranteed to be positive since the log-pseudo-partial likelihood in the numerator is negative. The optimal tuning parameters are chosen as $\arg\min_{(\lambda_{1n},\ldots,\lambda_{d_n n})} \text{GCV}\,(\lambda_{1n}, \ldots, \lambda_{d_n n})$. This d_n-dimensional optimization problem is difficult to solve in practice. We follow Cai et al. (2005) and take $\lambda_{nj} = \lambda_n\, \widehat{\text{SE}}\{\beta_{nj}^{(0)}\}$, where $\widehat{\text{SE}}\{\beta_{nj}^{(0)}\}$ is the estimated standard error of the unpenalized pseudo-partial-likelihood estimator used in § 4·1. Then the optimization problem reduces to a one-dimensional search for the optimal λ_n.

When $e(\lambda_n)/n$ is small, as is the case under the conditions for Theorems 1 and 2, we can write $\log \text{GCV}\,(\lambda_n) = \log\{-\tilde{\ell}_n(\hat{\beta}_n)/n\} - 2\log\{1 - e(\lambda_n)/n\} \approx \log\{-\tilde{\ell}_n(\hat{\beta}_n)/n\} + 2e(\lambda_n)/n$.

This expression is analogous to the Akaike information criterion (Akaike, 1973), so we write $\log \textsc{gcv}(\lambda_n)$ as $\textsc{aic}(\lambda_n)$ and define $\lambda_n^{\textsc{aic}} = \arg\min_{\lambda_n} \textsc{aic}(\lambda_n)$. Following the idea of the Bayesian information criterion (Schwarz, 1978), we define another tuning parameter selection criterion, where the optimal tuning parameter, denoted by $\lambda_n^{\textsc{bic}}$, minimizes $\textsc{bic}(\lambda_n) = \log\{-\tilde{\ell}_n(\hat{\beta}_n)/n\} + \log(n)e(\lambda_n)/n$. Wang et al. (2007) and Zhang et al. (2010) showed that, in linear and generalized linear models with a finite number of parameters, $\lambda_n^{\textsc{aic}}$ overfits the model with a positive probability whereas $\lambda_n^{\textsc{bic}}$ consistently identifies the true model. Such a result has not been established in the Cox proportional hazards model so far as we know. In the simulation section that follows, we investigate the performance of $\lambda_n^{\textsc{aic}}$ and $\lambda_n^{\textsc{bic}}$. Following Fan & Li (2001), we set the second tuning parameter a in the penalty function to 3·7 in our simulations.

In practice, researchers can perform a grid search to identify $\lambda_n^{\textsc{aic}}$ and $\lambda_n^{\textsc{bic}}$. The lower limit of the search range is zero and the upper limit is the smallest λ_n that gives an empty model. From our simulation experience, the upper limit rarely exceeds 2. Moreover, the model selection results are fairly robust with respect to the fineness of the search grid.

5. Numerical study and data application

5·1. *Simulation study*

Independent failure times are generated from the proportional hazards model. We let the baseline hazard be $\lambda_0(t) = 2$ and set the model dimension to $d_n = [5n_c^{1/5-1/500}]$ to reflect its dependence on sample size, where n_c is the expected number of cases for a given censoring rate and $[x]$ denotes x rounded to the nearest integer. We relate the model dimension to the number of cases rather than to the sample size directly, because the former better represents the amount of information in the dataset. We follow Tibshirani (1997) and consider two scenarios for the true parameter: a few large effects and many small effects. In the first scenario, $\beta_{n0} = (0{\cdot}35, 0, 0, 0{\cdot}6, 0, 0, -0{\cdot}8, 0, 0, 0{\cdot}6, 0, 0, -0{\cdot}8, 0, 0, \ldots)$; so a third of the components of β_{n0} are nonzero and the smallest nonzero effect in absolute value is 0·35, which corresponds to a hazard ratio of 1·4. In the second scenario, all components of β_{n0} equal 0·1, which corresponds to a hazard ratio of 1·1. In both scenarios, we generate the design matrix Z as a mixture of correlated binary and continuous variables. First, a d_n-dimensional multivariate standard normal variable Z^* is generated with $\mathrm{corr}(Z_i^*, Z_j^*) = 0{\cdot}5^{|i-j|}$. Then the first three components of Z^* are kept continuous while the next three components are dichotomized at zero, and this pattern is repeated for the rest of Z^*. Thus, half of the covariates become binary with parameter 0·5. The censoring times C_i are generated from a uniform distribution $\mathrm{Un}(0, c)$, with c adjusted to achieve the desired censoring percentage.

Various sample sizes, censoring rates, and noncase-to-case ratios are considered for both scenarios. Performance of the penalized variable selection with tuning parameter $\lambda_n^{\textsc{aic}}$ or $\lambda_n^{\textsc{bic}}$ is assessed. As a benchmark, we use the hard-threshold variable selection procedure, where the unpenalized full model is fitted and the components of the unpenalized estimates that give a significant Wald test at level 0·05 are included in the final model. We also consider the oracle procedure where the correct subset of covariates is used to fit the model. As the censoring rate is typically high in case-cohort studies, we set it to 80% or 90%, with 1000 replications for each setting.

We define the model error for a given model to be $\textsc{me}(\hat{\mu}) = E\{E(T \mid z) - \hat{\mu}(z)\}^2$. Under the proportional hazards model with constant baseline hazard λ_0, $\textsc{me}(\hat{\mu}) = \lambda_0^{-2} E\{\exp(-\hat{\beta}_n^{\mathrm{T}} z) - \exp(-\beta_{n0}^{\mathrm{T}} z)\}^2$. The relative model error of a given model is defined as the ratio of its model error to that of the unpenalized full model. We use the median and median absolute deviation of the relative model error to evaluate the prediction performance of different procedures. As

Table 1. *Model selection performance in the scenario of a few large effects*

		Noncase : Case = 1 : 1					Noncase : Case = 2 : 1			
		RME	Zeros		RITM		RME	Zeros		RITM
Method	α	Median (MAD)	C	I	(%)	α	Median (MAD)	C	I	(%)
		$n = 3000$, 80% censored, $d_n = 18$								
HT	0·25	0·67 (0·21)	11·2	0·0	45·4	0·50	0·65 (0·21)	11·3	0·0	52·1
SCAD(AIC)		0·63 (0·20)	10·7	0·0	30·3		0·49 (0·22)	11·5	0·0	61·6
SCAD(BIC)		0·39 (0·20)	12·0	0·2	83·7		0·37 (0·18)	12·0	0·0	95·2
Oracle		0·34 (0·16)	12·0	0·0	100·0		0·36 (0·17)	12·0	0·0	100·0
		$n = 3000$, 90% censored, $d_n = 15$								
HT	0·11	0·88 (0·30)	9·2	0·5	25·1	0·22	0·75 (0·29)	9·3	0·2	42·7
SCAD(AIC)		0·92 (0·14)	6·4	0·1	1·2		0·82 (0·20)	7·6	0·0	8·3
SCAD(BIC)		0·74 (0·38)	9·3	0·5	33·3		0·49 (0·30)	9·8	0·3	63·9
Oracle		0·32 (0·18)	10·0	0·0	100·0		0·33 (0·17)	10·0	0·0	100·0
		$n = 6000$, 90% censored, $d_n = 18$								
HT	0·11	0·71 (0·24)	11·1	0·1	39·6	0·22	0·64 (0·21)	11·3	0·0	48·4
SCAD(AIC)		0·89 (0·12)	7·9	0·0	1·2		0·80 (0·16)	9·5	0·0	9·4
SCAD(BIC)		0·49 (0·24)	11·5	0·1	58·6		0·38 (0·18)	11·9	0·0	87·8
Oracle		0·36 (0·17)	12·0	0·0	100·0		0·33 (0·15)	12·0	0·0	100·0
		$n = 10\,000$, 90% censored, $d_n = 20$								
HT	0·11	0·69 (0·20)	12·1	0·0	36·4	0·22	0·65 (0·20)	12·2	0·0	48·0
SCAD(AIC)		0·88 (0·14)	8·9	0·0	1·2		0·80 (0·18)	10·2	0·0	8·0
SCAD(BIC)		0·47 (0·21)	12·5	0·0	60·8		0·39 (0·18)	12·9	0·0	92·8
Oracle		0·34 (0·15)	13·0	0·0	100·0		0·35 (0·17)	13·0	0·0	100·0

α, subcohort sampling probability; RME, relative model error; MAD, median absolute deviation; C, average number of zero parameters correctly identified as zero; I, average number of nonzero parameters incorrectly identified as zero; RITM, rate of identifying true model; HT, hard threshold; SCAD(AIC), smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC), smoothly clipped absolute deviation with λ_n^{BIC}.

measures of variable selection performance, we also calculate the average number of parameters correctly estimated as zero, the average number of parameters erroneously estimated as zero, and the overall rate of identifying the true model. Point estimates, empirical and model-based standard errors, and empirical 95% confidence interval coverages are calculated for $\beta_{n01} = 0{\cdot}35$ in the first scenario.

Table 1 summarizes the simulation results in the scenario of a few large effects. The penalized method with tuning parameter λ_n^{BIC} has by far the best performance in all settings in terms of the relative model error and the rate of identifying the true model. The inferior performance of λ_n^{AIC} is apparently due to overfitting, as reflected by the low average number of correctly identified zero parameters; this is consistent with the theoretical findings of Wang et al. (2007) and Zhang et al. (2010). For both λ_n^{AIC} and λ_n^{BIC}, more noncases in the case-cohort design and lower censoring rates are associated with better prediction and variable selection performance. Table 2 summarizes the parameter estimation results of $\beta_{n01} = 0{\cdot}35$ under the same settings as for Table 1, but using only simulation replications where β_{n01} is correctly identified as nonzero. Conditional on $\hat{\beta}_{n1} \neq 0$, all procedures produce approximately unbiased point and standard error estimates, with coverage close to the nominal level. The normality of the sampling distributions of $\hat{\beta}_{n1}$ was assessed by Q-Q plots, shown in the Supplementary Material. The sampling distribution of $\hat{\beta}_{n1}$ is a mixture of a point mass at zero and a left-truncated distribution that is well approximated by a truncated normal distribution. As the rate of identifying the true model increases, the point mass at zero vanishes and the sampling distribution of $\hat{\beta}_{n1}$ becomes normal.

Table 2. *Estimation performance for $\beta_{n01} = 0{\cdot}35$ in the scenario of a few large effects; results are based on replications where $\hat{\beta}_{n1} \neq 0$*

Method	Noncase : Case = 1 : 1 n_c	$\hat{\beta}_{n1}$	SE$_e$ ($\times 10^{-2}$)	SE$_m$ ($\times 10^{-2}$)	95% CI$_e$	Noncase : Case = 2 : 1 n_c	$\hat{\beta}_{n1}$	SE$_e$ ($\times 10^{-2}$)	SE$_m$ ($\times 10^{-2}$)	95% CI$_e$
				$n = 3000$, 80% censored, $d_n = 18$						
HT	998	0·36	7·00	6·66	92·6	1000	0·35	5·85	5·55	92·7
SCAD(AIC)	1000	0·35	6·68	5·95	92·0	1000	0·35	5·28	4·87	92·7
SCAD(BIC)	991	0·35	5·96	5·88	94·8	1000	0·35	5·12	4·84	93·3
Oracle	1000	0·35	6·06	5·89	94·5	1000	0·35	5·08	4·84	93·5
				$n = 3000$, 90% censored, $d_n = 15$						
HT	888	0·40	10·9	11·0	92·8	971	0·37	9·26	9·20	94·4
SCAD(AIC)	981	0·38	11·9	10·2	89·8	997	0·36	9·24	8·29	92·2
SCAD(BIC)	916	0·38	10·3	9·83	92·5	964	0·36	8·19	8·04	94·7
Oracle	1000	0·36	10·8	9·87	92·1	1000	0·35	8·37	8·05	93·8
				$n = 6000$, 90% censored, $d_n = 18$						
HT	992	0·37	8·27	7·95	92·5	1000	0·36	7·01	6·53	92·2
SCAD(AIC)	1000	0·36	8·40	7·32	91·2	1000	0·36	6·73	5·92	91·0
SCAD(BIC)	992	0·36	7·68	7·09	92·5	996	0·35	6·06	5·74	93·8
Oracle	1000	0·35	7·64	7·10	93·0	1000	0·35	6·03	5·74	94·0
				$n = 10\,000$, 90% censored, $d_n = 20$						
HT	1000	0·36	6·51	6·29	93·2	1000	0·35	5·27	5·10	94·4
SCAD(AIC)	1000	0·36	6·31	5·83	91·6	1000	0·35	5·11	4·63	94·0
SCAD(BIC)	1000	0·36	5·93	5·67	94·0	1000	0·35	4·55	4·50	94·8
Oracle	1000	0·36	5·74	5·67	95·0	1000	0·35	4·53	4·50	94·8

n_c, number of simulation replications where $\hat{\beta}_{n1} \neq 0$; SE$_e$, empirical standard error; SE$_m$, model-based standard error; 95% CI$_e$, empirical 95% confidence interval coverage; HT, hard threshold; SCAD(AIC), smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC), smoothly clipped absolute deviation with λ_n^{BIC}.

Table 3 summarizes the simulation results in the scenario of many small effects, where all $\beta_{n0} = 0{\cdot}1$. In this scenario the oracle model is just the unpenalized full model with the relative model error being unity by definition, which is not very informative and hence not included in the table. With many small but nonzero effects, none of the three methods can identify all the effects with a high probability, as reflected by the near-zero rate of identifying the true model in all settings, which is not shown in the table. The inference results are not satisfactory either; they are not shown due to space limitations. Nevertheless, λ_n^{AIC} produces the smallest relative model error, suggesting that it has the best prediction performance among the three methods. Moreover, λ_n^{AIC} correctly identifies the largest number of small effects as nonzero. The Bayesian information criterion tends to select sparse models, so it may not perform as well as the Akaike information criterion when there are many small nonzero parameters. The relative model error is not comparable across different settings because it depends on the model error of the full model, which shows large variation in this scenario.

5·2. *Analysis of the Busselton Health Study*

We use the proposed variable selection procedures to analyse the Busselton Health Study data (Cullen, 1972; Knuiman et al., 2003). The study comprises a series of cross-sectional health surveys conducted in the town of Busselton in Western Australia. Every three years from 1966 to 1981, general health information was collected from adult participants by questionnaire and through clinical visits. In this analysis we are interested in identifying risk factors for stroke. In particular, the main risk factor of interest is serum ferritin level. We also consider several other

Table 3. *Model selection performance in the scenario of many small effects with all $\beta_{n0} = 0{\cdot}1$*

Method	α	Noncase : Case = 1 : 1 RME Median (MAD)	Nonzero estimates	α	Noncase : Case = 2 : 1 RME Median (MAD)	Nonzero estimates
		$n = 3000$, 80% censored, $d_n = 18$				
HT	0·25	2·90 (1·50)	4·0	0·50	3·59 (1·82)	5·2
SCAD(AIC)		1·79 (0·88)	6·0		3·15 (1·59)	5·5
SCAD(BIC)		5·62 (2·39)	1·3		8·94 (3·46)	1·1
		$n = 3000$, 90% censored, $d_n = 15$				
HT	0·11	1·89 (1·00)	2·6	0·22	2·91 (1·63)	3·5
SCAD(AIC)		0·99 (0·29)	6·0		1·67 (0·78)	5·4
SCAD(BIC)		2·48 (1·23)	1·8		4·92 (2·08)	1·5
		$n = 6000$, 90% censored, $d_n = 18$				
HT	0·11	2·82 (1·45)	3·4	0·22	3·48 (1·69)	4·5
SCAD(AIC)		1·08 (0·28)	8·6		1·41 (0·54)	8·3
SCAD(BIC)		3·17 (1·52)	3·0		5·36 (2·47)	2·6
		$n = 10\,000$, 90% censored, $d_n = 20$				
HT	0·11	3·85 (2·02)	6·0	0·22	4·49 (2·37)	7·7
SCAD(AIC)		1·26 (0·39)	11·6		1·84 (0·81)	11·4
SCAD(BIC)		4·91 (2·49)	4·7		8·38 (3·75)	4·2

α, subcohort sampling probability; RME, relative model error; MAD, median absolute deviation; Nonzero estimates, average number of parameters not estimated as zero; HT, hard threshold; SCAD(AIC), smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC), smoothly clipped absolute deviation with λ_n^{BIC}.

risk factors in the variable selection process: age, body mass index, blood pressure treatment, systolic blood pressure, cholesterol, triglycerides, haemoglobin and smoking status. All variables were measured at baseline. The full cohort of this analysis consists of 1401 subjects aged 40 to 89 years who participated in the Busselton Health Survey in 1981 and had no history of diagnosed coronary heart disease or stroke at that time. Subjects were followed until 31 December 1998, and their time to stroke, if one took place, was recorded. Subjects were treated as censored if they left Western Australia during the follow-up period. There were 118 incidences of stroke in the full cohort during the follow-up period. To reduce costs and preserve stored serum, a case-cohort design was used where the serum ferritin level was measured for only a randomly selected subcohort plus all stroke cases. The size of the random subcohort was 450, and the case-cohort size was 513.

Table 4 summarizes the baseline characteristics of the full cohort and subcohort. The average ferritin level is not available for the full cohort due to the case-cohort design. The summary statistics for the baseline characteristics of the full cohort and the subcohort are similar, suggesting that the subcohort is representative of the full cohort.

We apply the hard-threshold method and the penalized variable selection procedures with tuning parameters λ_n^{AIC} and λ_n^{BIC} to the Busselton Health Study. In order to avoid missing any potentially important effects, we also include in the initial model the quadratic terms of all continuous covariates as well as interactions between ferritin and all covariates. The total number of parameters is 28. All continuous covariates are standardized using the means and standard deviations from the subcohort, shown in Table 4. To decrease their skewness, we log-transformed the values of ferritin and triglycerides before standardization. The tuning parameter selector identifies $\lambda_n^{\text{AIC}} = 0{\cdot}244$ and $\lambda_n^{\text{BIC}} = 0{\cdot}305$. Table 5 shows the models identified by the three methods. Due to space limitations, only terms that are selected by at least one method are shown. The use of λ_n^{AIC} results in seven terms being selected, and the use of λ_n^{BIC} results in four terms being selected. Both methods select age, sex, blood pressure treatment, and squared systolic blood pressure as

Table 4. *Baseline characteristics of the Busselton Health Study*

Variables	Full cohort ($n = 1401$) Mean (SD) or %	Subcohort ($\tilde{n} = 450$) Mean (SD) or %
Age (years)	58·0 (10·8)	58.9 (10·9)
Body mass index	25·9 (3·9)	25.9 (4·0)
Blood pressure treatment (%)	17·2	18.4
Systolic blood pressure (mmHg)	132·2 (20·0)	132.9 (20·2)
Cholesterol (mmol/L)	6·14 (1·14)	6.24 (1·17)
Triglycerides (mmol/L)	1·52 (0·97)	1.55 (0·97)
Haemoglobin (g/100 ml)	141·9 (12·0)	142.0 (11·5)
Smoking (%)		
Never	49·5	51.6
Former	32·4	32.0
Current	18·1	16.4
Ferritin (μg/L)	–	148.1 (140.8)
log(ferritin)	–	4.57 (1.01)

SD, standard deviation.

Table 5. *Estimated coefficients and standard errors for the Busselton Health Study data; all continuous covariates were standardized using the means and standard deviations based on the random subcohort before applying the variable selection procedure*

Variable	Hard threshold $\hat{\beta}$ ($\hat{\text{SE}}$)	SCAD(AIC) $\hat{\beta}$ ($\hat{\text{SE}}$)	SCAD(BIC) $\hat{\beta}$ ($\hat{\text{SE}}$)
Age (years)	0.92 (0·27)	0·87 (0·15)	0.85 (0·14)
Sex (1 = female)	0 (–)	−0·61 (0·26)	−0.65 (0·25)
Blood pressure treatment	0.83 (0.34)	0·83 (0.29)	0.89 (0·25)
Systolic blood pressure	0 (–)	0·21 (0·15)	0 (–)
Systolic blood pressure^2	0 (–)	0·092 (0.067)	0.16 (0·044)
log(triglycerides)	0 (–)	−0·24 (0·18)	0 (–)
log^2(triglycerides)	0 (–)	0·18 (0·093)	0 (–)

SCAD(AIC), smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC), smoothly clipped absolute deviation with λ_n^{BIC}.

important risk factors for stroke. The procedure using λ_n^{AIC} additionally selects the linear term of systolic blood pressure and the linear and squared terms of triglycerides. The hard-threshold method selects only age and blood pressure treatment.

To shed some light on which model provides the best fit to the data, we performed five-fold crossvalidation. The average log-pseudo-partial likelihood from the test datasets is used as the validation statistic. The hard-threshold method and penalized variable selection with λ_n^{AIC} and λ_n^{BIC} give validation statistics of −621·5, −627·7 and −614·0, respectively. Therefore, we consider the model with λ_n^{BIC} to be the best fit to the Busselton data. According to this model, increased age, maleness, blood pressure treatment, and increased systolic blood pressure are associated with a higher risk of stroke. There is no evidence that serum ferritin level is associated with stroke.

6. Discussion

One potential limitation of the theorems presented in this paper is that they only establish the consistency and oracle property for a local maximizer of the penalized objective function. Because of the nonconcavity of the penalized objective function, there may be multiple maximizers. However, based on Fan & Li (2001, § 3.5) and judging from the small bias in the estimates in Table 2, it is reasonable to assume that the maximizer identified by using the unpenalized estimator as the initial value is the $(n/d_n)^{1/2}$-consistent local maximizer described in Theorems 1 and 2.

In this paper the quantity $\hat{\alpha}(t)$ used in the weight function $\rho(t)$ is calculated at each failure time-point and so is time-dependent. When cases are rare, $\hat{\alpha}(t)$ is almost constant across t. However, using time-dependent $\hat{\alpha}(t)$ is more general and allows the sampling probability to vary with time. Therefore, we use $\hat{\alpha}(t)$ in this paper. A potential practical issue is that $\hat{\alpha}(t)$ may not be reliable if the number of noncases in the random subcohort is very small, although this is highly unlikely due to the use of case-cohort design for studies of rare disease. In the unlikely situation where there is no noncase left in the subcohort, $\hat{\alpha}(t)$ is not well-defined. To avoid computational difficulties, one can define $(1-\Delta)\xi/\hat{\alpha}(t) = 0$ if $\hat{\alpha}(t) = 0$. In fact, when $\hat{\alpha}(t) = 0$, $1-\Delta$ is necessarily zero for all subjects remaining in the subcohort.

There is a strong line of research on the convergence of and post-selection inference for penalized estimators (Leeb & Pötscher, 2005; Leeb & Pötscher, 2006; Pötscher & Leeb, 2009). In particular, Pötscher & Leeb (2009) showed that the penalized estimators are not uniformly consistent, and that their asymptotic distributions are nonnormal if the true parameter lies within a shrinking neighbourhood of zero with rate $(d_n/n)^{1/2}$. The lack of local regularity is a theoretical limitation of penalized variable selection methods. However, in this paper Condition 5, together with the requirement that $\lambda_{nj}(n/d_n)^{1/2} \to \infty$ for all j, ensures that the nonzero parameters are uniformly larger than $O\{(d_n/n)^{1/2}\}$, hence avoiding the aforementioned irregularity. Our simulation study suggests that the performance of the proposed variable selection method depends on the true effect size. In practice, since this size is unknown, we suggest conducting penalized variable selection with both Akaike and Bayesian information criteria-based tuning parameter selection, and then using crossvalidation to choose the best model, as done in § 5·2. Theoretical justification of these model selection approaches will be investigated further. Moreover, as the regularity conditions required for our asymptotic results may not be testable in finite samples, it will be important to replicate findings from one particular finite data analysis. One possible way to examine the consistency of findings is to use bootstrap data or to apply a resampling-based variable selection approach such as stability selection (Meinshausen & Bühlmann, 2010).

In the Busselton data analysis we standardized all continuous covariates, for several reasons. First, this makes the regression coefficients comparable. Second, it reduces the correlation between the linear and quadratic terms and between the main effect and interaction terms, which generally results in more robust and precise parameter estimates. More importantly, penalized regression procedures are not invariant with respect to covariate scaling, and standardization makes the penalization fair for all covariates (Tibshirani, 1997). For these reasons, we recommend standardizing continuous covariates before carrying out penalized regression.

Acknowledgement

We thank Professor Matthew Knuiman and the Busselton Population Medical Research Foundation for permission to use the data in the analysis of § 5·2. This work was partially supported by the U.S. National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of the lemmas in the Appendix, the estimation of the covariance matrix $\Gamma_n(\beta_{n0})$, and Q-Q plots of the estimate $\hat{\beta}_{n1}$ in the simulation scenario of a few large effects.

Appendix

Proofs of the theorems

Throughout the proofs, we write $\tilde{\ell}'_n(\beta_{n0})_j = \partial\tilde{\ell}_n(\beta_{n0})/\partial\beta_{nj}$, $\tilde{\ell}''_n(\beta_{n0})_{jk} = \partial^2\tilde{\ell}_n(\beta_{n0})/(\partial\beta_{nj}\partial\beta_{nk})$ and $\tilde{\ell}'''_n(\beta_{n0})_{jkl} = \partial^3\tilde{\ell}_n(\beta_{n0})/(\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl})$. We also let $\tilde{V}_{njk}(\beta_{n0}, t)$, $V_{njk}(\beta_{n0}, t)$, $\tilde{S}^{(2)}_{njk}(\beta_{n0}, t)$ and $S^{(2)}_{njk}(\beta_{n0}, t)$ be the (j, k)th components of the corresponding matrices. For a matrix $A = \{a_{ij}\}$ $(i, j = 1, \ldots, n)$, the norm is defined as $\|A\| = (\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2)^{1/2}$. The following lemma will be used repeatedly.

Lemma A1. *Let $W_n(t)$ and $G_n(t)$ be two sequences of processes with bounded variation almost surely, and suppose that $G_n(t)$ is progressively measurable and cadlag. For some constant τ, assume that $\sup_{0\leqslant t\leqslant\tau} \|W_n(t) - W(t)\| \to 0$ in probability for some bounded process $W(t)$, that $W_n(t)$ is monotone on $[0, \tau]$, and that $G_n(t)$ converges to a zero-mean process with continuous sample paths in the metric space* BV$[0, \tau]$*, the bounded variation function space on $[0, \tau]$. Then both $\sup_{0\leqslant t\leqslant\tau} \| \int_0^t \{W_n(s) - W(s)\}\, \mathrm{d}G_n(s)\|$ and $\sup_{0\leqslant t\leqslant\tau} \| \int_0^t G_n(s)\, \mathrm{d}\{W_n(s) - W(s)\}\|$ converge to zero in probability as $n \to \infty$.*

The proof of this lemma follows straightforwardly from that of Lemma 1 in Lin (2000), upon noting that a process with bounded variation can be decomposed into two monotone processes.

We also need the following lemmas, proofs of which are provided in the Supplementary Material.

Lemma A2. *Let $\xi = (\xi_1, \ldots, \xi_n)$ be a random vector containing $\tilde{n}$ ones and $n - \tilde{n}$ zeros, with each permutation equally likely. Let $X_{ni}(t)$ $(i = 1, \ldots, n)$ be a triangular array of real-valued random processes on $[0, \tau]$, with $E\{X_{ni}(t)\} = \mu_n(t)$, $\mathrm{var}\{X_{ni}(0)\} < \infty$ and $\mathrm{var}\{X_{ni}(\tau)\} < \infty$ for all i and n. Let $X_n(t) = \{X_{n1}(t), \ldots, X_{nn}(t)\}$ and ξ be independent. Suppose that almost all paths of $X_{ni}(t)$ have finite variation. Then $n^{-1/2}\sum_{i=1}^n \xi_i\{X_{ni}(t) - \mu_n(t)\}$ converges weakly to a tight zero-mean Gaussian process and hence $n^{-1}\sum_{i=1}^n \xi_i\{X_{ni}(t) - \mu_n(t)\}$ converges in probability to zero uniformly in t.*

Lemma A3. *Given that ξ is independent of Δ and $Y(t)$, $n^{1/2}\{\hat{\alpha}^{-1}(t) - \alpha^{-1}\}$ converges weakly to a zero-mean Gaussian process.*

Lemma A4. *Under Conditions 1–3, for any nonzero $d_n \times 1$ constant vector u_n with $\|u_n\| = C < \infty$ and $\|u_n\|_0 = c_n > 0$, where $\|\cdot\|_0$ denotes the number of nonzero components of a vector, $n^{1/2}\{\tilde{S}^{(0)}_n(\beta_{n0}, t) - S^{(0)}_n(\beta_{n0}, t)\}$, $(n/c_n)^{1/2}u_n^{\mathrm{T}}\{\tilde{S}^{(1)}_n(\beta_{n0}, t) - S^{(1)}_n(\beta_{n0}, t)\}$ and $n^{1/2}c_n^{-1}u_n^{\mathrm{T}}\{\tilde{S}^{(2)}_n(\beta_{n0}, t) - S^{(2)}_n(\beta_{n0}, t)\}u_n$ all converge weakly to tight zero-mean Gaussian processes.*

Lemma A5. *Under Conditions 1–4, for any nonzero $d_n \times 1$ constant vector u_n with $\|u_n\| = 1$, $n^{-1/2}u_n^{\mathrm{T}}\Gamma_n^{-1/2}(\beta_{n0})\tilde{\ell}'_n(\beta_{n0})$ converges to a standard normal distribution, where $\Gamma_n(\beta_{n0})$ is the covariance matrix of $n^{-1/2}\tilde{\ell}'_n(\beta_{n0})$.*

Lemma A6. *Under Conditions 1–4, $n^{-1/2}\{\tilde{\ell}''_n(\beta_{n0})_{jk} + nI_n(\beta_{n0})_{jk}\}$ is $O_{\mathrm{p}}(1)$ for $j, k = 1, \ldots, d_n$, where $I_n(\beta_{n0})_{jk}$ is the (j, k)th component of $I_n(\beta_{n0})$ as defined in § 3·2.*

Lemma A7. *Under Conditions 1–6, if $d_n^4/n \to 0$, $\lambda_{nj} \to 0$ and $\lambda_{nj}n^{1/2}d_n^{-1/2} \to \infty$ with probability tending to 1, for any given $\beta_{n,\mathrm{I}}$ satisfying $\|\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\| = O(d_n^{1/2}n^{-1/2})$ and any constant C, we have that $\tilde{Q}_n\{(\beta_{n,\mathrm{I}}^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}\} = \max_{\|\beta_{n,\mathrm{II}}\|\leqslant Cd_n^{1/2}n^{-1/2}} \tilde{Q}_n\{(\beta_{n,\mathrm{I}}^{\mathrm{T}}, \beta_{n,\mathrm{II}}^{\mathrm{T}})^{\mathrm{T}}\}$.*

Proof of Theorem 1. Let β_{n0} be the true parameters, and let $\alpha_n = d_n^{1/2}(n^{-1/2} + a_n)$. It suffices to show that for any $\varepsilon > 0$ and any constant vector u_n with $\|u_n\| = C$, there exists a large enough C such that $\mathrm{pr}\{\sup_{\|u_n\|=C} \tilde{Q}_n(\beta_{n0} + \alpha_n u_n) < \tilde{Q}_n(\beta_{n0})\} \geqslant 1 - \varepsilon$. This implies the existence of a local maximizer $\hat{\beta}_n$

such that $\|\hat{\beta}_n - \beta_{n0}\| = O_{\mathrm{p}}(\alpha_n)$. Since $P_{\lambda_{nj}}(0) = 0$ and $P_{\lambda_{nj}}(\cdot) \geqslant 0$, we have

$$\tilde{Q}_n(\beta_{n0} + \alpha_n u_n) - \tilde{Q}_n(\beta_{n0})$$
$$\leqslant \{\tilde{\ell}_n(\beta_{n0} + \alpha_n u_n) - \tilde{\ell}_n(\beta_{n0})\} - n\sum_{j=1}^{k_n}\{P_{\lambda_{nj}}(|\beta_{n0j} + \alpha_n u_{nj}|) - P_{\lambda_{nj}}(|\beta_{n0j}|)\} = I_1 - I_2.$$

We first consider I_1. By Taylor expansion,

$$I_1 = \alpha_n u_n^{\mathrm{T}}\tilde{\ell}_n'(\beta_{n0}) + \frac{1}{2}\alpha_n^2 u_n^{\mathrm{T}}\tilde{\ell}_n''(\beta_{n0})u_n + \frac{1}{6}\alpha_n^3\sum_{i=1}^{n}\sum_{j,k,l=1}^{d_n}\tilde{\ell}_i'''(\beta_n^*)_{jkl}u_{nj}u_{nk}u_{nl} = I_{11} + I_{12} + I_{13},$$

where β_n^* lies between β_{n0} and $\beta_{n0} + \alpha_n u_n$. From Lemma A5 we have $\tilde{\ell}_n'(\beta_{n0})_j = O_{\mathrm{p}}(n^{1/2})$ for $j = 1, \ldots, d_n$. Therefore,

$$|I_{11}| = |\alpha_n u_n^{\mathrm{T}}\tilde{\ell}_n'(\beta_{n0})| \leqslant \alpha_n\|u_n\|\|\tilde{\ell}_n'(\beta_{n0})\| = \alpha_n\|u_n\|O_{\mathrm{p}}\{(d_n n)^{1/2}\} = \|u_n\|O_{\mathrm{p}}(\alpha_n^2 n).$$

The term I_{12} can be written as $\alpha_n^2 u_n^{\mathrm{T}}\{\tilde{\ell}_n''(\beta_{n0}) + nI_n(\beta_{n0})\}u_n/2 - \alpha_n^2 u_n^{\mathrm{T}} nI_n(\beta_{n0})u_n/2 = J_1 - J_2$. By the Cauchy–Schwarz inequality, the fact that $\tilde{\ell}_n''(\beta_{n0})_{jk} + nI_n(\beta_{n0})_{jk} = O_{\mathrm{p}}(n^{1/2})$ for $j, k = 1, \ldots, d_n$ and Lemma A6, we have $|J_1| \leqslant \alpha_n^2\|u_n\|^2\|\tilde{\ell}_n''(\beta_{n0}) + nI_n(\beta_{n0})\|/2 = \|u_n\|^2 O_{\mathrm{p}}(\alpha_n^2 n^{1/2} d_n) = \|u_n\|^2 o_{\mathrm{p}}(\alpha_n^2 n)$. By spectral decomposition of $I_n(\beta_{n0})$ and Condition 4, $|J_2| \geqslant \alpha_n^2\|u_n\|^2 n\lambda_{\min}\{I_n(\beta_{n0})\}/2 \geqslant \|u_n\|^2(\alpha_n^2 n)C_2/2$. Under Conditions 1–3, $\partial\tilde{V}_{njk}(\beta_n^*, t)/\partial\beta_{nl}$ has bounded variation in t for $i = 1, \ldots, n$ and $j, k, l = 1, \ldots, d_n$. Therefore $\tilde{\ell}_i'''(\beta_n^*)_{jkl} = -\int_0^\tau \partial\tilde{V}_{njk}(\beta_n^*, t)/\partial\beta_{nl}\,\mathrm{d}N_i(t)$ is $O_{\mathrm{p}}(1)$. Combining this with $\alpha_n = d_n^{1/2}(n^{-1/2} + a_n)$, $d_n^4/n \to 0$ and $d_n^2 a_n \to 0$, we obtain $|I_{13}| = O_{\mathrm{p}}(d_n^{3/2})n\alpha_n^3\|u_n\|^3 = O_{\mathrm{p}}\{d_n^2(n^{-1/2} + a_n)\}n\alpha_n^2\|u_n\|^3 = \|u_n\|^3 o_{\mathrm{p}}(\alpha_n^2 n)$. Therefore, for large enough $\|u_n\|$, $|J_2|$ dominates $|I_{11}|$, $|J_1|$ and $|I_{13}|$.

Now consider I_2. By Taylor expansion and the Cauchy–Schwarz inequality,

$$|I_2| = \left|n\sum_{j=1}^{k_n}P_{\lambda_{nj}}'(|\beta_{n0j}|)\mathrm{sgn}(\beta_{n0j})\alpha_n u_{nj} + \frac{1}{2}n\sum_{j=1}^{k_n}P_{\lambda_{nj}}''(|\beta_{n0j}|)\alpha_n^2 u_{nj}^2\{1 + o(1)\}\right|$$
$$\leqslant n\left|\sum_{j=1}^{k_n}P_{\lambda_{nj}}'(|\beta_{n0j}|)\alpha_n u_{nj}\right| + \frac{1}{2}n\left|\sum_{j=1}^{k_n}P_{\lambda_{nj}}''(|\beta_{n0j}|)\alpha_n^2 u_{nj}^2\{1 + o(1)\}\right|$$
$$\leqslant n\alpha_n a_n k_n^{1/2}\|u_n\| + \frac{1}{2}n\alpha_n^2 b_n\|u_n\|^2\{1 + o(1)\}$$
$$= \|u_n\|O_{\mathrm{p}}(\alpha_n^2 n).$$

The last equality holds because $a_n = O_{\mathrm{p}}(\alpha_n d_n^{-1/2})$ and $b_n \to 0$ under Condition 5. Therefore, $|J_2|$ dominates $|I_2|$ for large enough C. Since J_2 is negative, it follows that for large enough C, $\tilde{Q}_n(\beta_{n0} + \alpha_n u_n) - \tilde{Q}_n(\beta_{n0})$ is negative with probability tending to 1 as $n \to \infty$. This completes the proof of Theorem 1. □

Proof of Theorem 2. The assertion that $\hat{\beta}_{n,\mathrm{II}} = 0$ with probability tending to 1 as $n \to \infty$ follows directly from Lemma A7. To prove the second assertion, we first show that

$$n^{1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\left[(I_{n11} + \Sigma_n)(\hat{\beta}_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})\{1 + o_{\mathrm{p}}(1)\} + B_n\right] = n^{-1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\tilde{\ell}_{n1}'(\beta_{n0}) + o_{\mathrm{p}}(1), \quad \text{(A1)}$$

where $\tilde{\ell}_{n1}'(\beta_{n0})$ consists of the first k_n components of $\tilde{\ell}_n'(\beta_{n0})$. Since $\hat{\beta}_{n,\mathrm{I}}$ is the maximum penalized pseudo-partial-likelihood estimator, $\partial\tilde{Q}_n(\hat{\beta}_n)/\partial\beta_{n,\mathrm{I}} = 0$. By Taylor expansion of $\partial\tilde{Q}_n(\hat{\beta}_n)/\partial\beta_{n,\mathrm{I}}$ at $\beta_{n0,\mathrm{I}}$ and the fact that $\hat\beta_{n,\mathrm{II}} - \beta_{n0,\mathrm{II}} = 0$ with probability tending to 1, we have

$$\tilde\ell'_{n1}(\beta_{n0}) + \tilde\ell''_{n1}(\beta_{n0})(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) + (\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})^{\mathrm{T}}\tilde\ell'''_{n1}(\beta_n^*)(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})/2 \\ - nB_n - n\Sigma_n^{**}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) = 0 \qquad (\mathrm{A2})$$

with probability tending to 1, where $\tilde\ell''_{n1}(\beta_{n0})$ consists of the first $k_n \times k_n$ components of $\tilde\ell''_n(\beta_{n0})$, $\tilde\ell'''_{n1}(\beta_n^*)$ consists of the first $k_n \times k_n \times k_n$ components of $\tilde\ell'''_n(\beta_n^*)$, β_n^* lies between $\hat\beta_n$ and β_{n0}, and $\Sigma_n^{**} = \Sigma_n(\beta_n^{**})$ with β_n^{**} between $\hat\beta_n$ and β_{n0}. Upon rearranging (A2), we get

$$\{\tilde\ell''_{n1}(\beta_{n0}) - n\Sigma_n^{**}\}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) - nB_n = -\tilde\ell'_{n1}(\beta_{n0}) - \frac{1}{2}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})^{\mathrm{T}}\tilde\ell'''_{n1}(\beta_n^*)(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}). \qquad (\mathrm{A3})$$

Write $\nu_n = (\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})^{\mathrm{T}}\tilde\ell'''_{n1}(\beta_n^*)(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})$. Multiplying both sides of (A3) by $n^{-1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}$ gives

$$n^{1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde\ell''_{n1}(\beta_{n0}) - \Sigma_n^{**}\right\}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) - n^{1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}B_n \\ = -n^{-1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\tilde\ell'_{n1}(\beta_{n0}) - n^{-1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\nu_n/2. \qquad (\mathrm{A4})$$

By the Cauchy–Schwarz inequality, $\|\nu_n\| \leqslant \|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\|^2 \sum_{i=1}^n\{\sum_{j,k,l=1}^{k_n}\tilde\ell'''_{i1}(\beta^*)^2_{jkl}\}^{1/2}$. As shown in the proof of Theorem 1, $\tilde\ell'''_{i1}(\beta^*)_{jkl} = O_{\mathrm{p}}(1)$, so $\|\nu_n\| = O_{\mathrm{p}}\{(d_n/n)nk_n^{3/2}\} = O_{\mathrm{p}}(d_n^{5/2})$. By spectral decomposition of $\Gamma_{n11}^{-1/2}$, $d_n^5/n \to 0$ and Condition 4, we have

$$\frac{1}{2}n^{-1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\nu_n \leqslant \frac{\|u_n\|\|\nu_n\|}{2}n^{-1/2}\lambda_{\max}(\Gamma_n^{-1/2}) = O_{\mathrm{p}}(d_n^{5/2}n^{-1/2}) = o_{\mathrm{p}}(1). \qquad (\mathrm{A5})$$

The inequality in (A5) holds by the Cauchy–Schwarz inequality and the Cauchy interlacing inequality for symmetric matrices. Moreover, $u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}n^{-1}\tilde\ell''_{n1}(\beta_{n0})(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) = u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\{n^{-1}\tilde\ell''_{n1}(\beta_{n0}) + I_{n11}(\beta_{n0})\}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) - u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}I_{n11}(\beta_{n0})(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) = J_1 - J_2$. By the Cauchy–Schwarz inequality and Lemma A6, $|J_1| \leqslant \|u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\|\|n^{-1}\tilde\ell''_{n1}(\beta_{n0}) + I_{n11}(\beta_{n0})\|\|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\| = \|u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\|\|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\|O_{\mathrm{p}}(d_n n^{-1/2})$. Also, we have $|J_2| \geqslant \|u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\|\|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\|\lambda_{\min}(I_{n11}) \geqslant \|u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\|\|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\|\lambda_{\min}(I_n)$. Then, by Condition 4,

$$\left|\frac{J_1}{J_2}\right| \leqslant \frac{\|u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\|\|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\|O_{\mathrm{p}}(d_n n^{-1/2})}{\|u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\|\|\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}\|\lambda_{\min}(I_n)} = O_{\mathrm{p}}(d_n n^{-1/2}) = o_{\mathrm{p}}(1).$$

Therefore $J_1 = o_{\mathrm{p}}(J_2)$ and $u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}n^{-1}\tilde\ell''_{n1}(\beta_{n0})(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) = -u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}I_{n11}(\beta_{n0})(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})\{1 + o_{\mathrm{p}}(1)\}$. Since $\hat\beta_n$ converges to β_{n0} in probability, it follows that

$$u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde\ell''_{n1}(\beta_{n0}) - \Sigma_n^{**}\right\}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}}) = -u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\left\{I_{n11}(\beta_{n0}) + \Sigma_n\right\}(\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}})\{1 + o_{\mathrm{p}}(1)\}. \qquad (\mathrm{A6})$$

By (A4), (A5) and (A6), we have that (A1) holds. By Lemma A5, $n^{-1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}\tilde\ell'_{n1}(\beta_{n0})$ converges to the standard normal distribution. Therefore, $n^{1/2}u_n^{\mathrm{T}}\Gamma_{n11}^{-1/2}(I_{n11} + \Sigma_n)\{\hat\beta_{n,\mathrm{I}} - \beta_{n0,\mathrm{I}} + (I_{n11} + \Sigma_n)^{-1}B_n\} \to N(0, 1)$ in distribution. This proves the second assertion of Theorem 2. □

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–65.

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064–72.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* **6**, 39–58.

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

CHO, H. & QU, A. (2013). Model selection for correlated data with diverging number of parameters. *Statist. Sinica* **23**, 901–27.
COX, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.
CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.
CULLEN, K. J. (1972). Mass health examinations in the Busselton population, 1966 to 1970. *Austr. J. Med.* **2**, 714–8.
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, Y. & TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B **75**, 531–52.
HASTIE, T., TIBSHIRANI, R. J. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Berlin: Springer, 2nd ed.
HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1**, 799–821.
HUNTER, D. & LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617–42.
KALBFLEISCH, J. D. & LAWLESS, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.* **7**, 149–60.
KANG, S. & CAI, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika* **96**, 887–901.
KIM, S., CAI, J. & LU, W. (2013). More efficient estimators for case-cohort studies. *Biometrika* **100**, 695–708.
KNUIMAN, M. W., DIVITINI, M. L., OLYNYK, J. K., CULLEN, D. J. & BARTHOLOMEW, H. C. (2003). Serum ferritin and cardiovascular disease: A 17-year follow-up study in Busselton, Western Australia. *Am. J. Epidemiol.* **158**, 144–9.
KULICH, M. & LIN, D. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Am. Statist. Assoc.* **99**, 832–44.
LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Economet. Theory* **21**, 21–59.
LEEB, H. & PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34**, 2554–91.
LIN, D. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87**, 37–47.
MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection (with Discussion). *J. R. Statist. Soc.* B **72**, 417–73.
PENG, H. & FAN, J. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–61.
PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356–66.
PÖTSCHER, B. M. & LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Mult. Anal.* **100**, 2065–82.
PRENTICE, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
SELF, S. G. & PRENTICE, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16**, 64–81.
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.
WANG, H., LI, B. & LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc.* B **71**, 671–83.
WANG, H., LI, R. & TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–68.
ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
ZHANG, Y., LI, R. & TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Am. Statist. Assoc.* **105**, 312–23.
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 June* 2014. *Revised May* 2016]