Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(8) 1833–1849

DOI: 10.1177/09622802211009259
journals.sagepub.com/home/smm
SAGE

A unified approach to variable selection for Cox's proportional hazards model with interval-censored failure time data

Mingyue Du[1], Hui Zhao[2] and Jianguo Sun[3]

Abstract

Cox's proportional hazards model is the most commonly used model for regression analysis of failure time data and some methods have been developed for its variable selection under different situations. In this paper, we consider a general type of failure time data, case *K* interval-censored data, that include all of other types discussed as special cases, and propose a unified penalized variable selection procedure. In addition to its generality, another significant feature of the proposed approach is that unlike all of the existing variable selection methods for failure time data, the proposed approach allows dependent censoring, which can occur quite often and could lead to biased or misleading conclusions if not taken into account. For the implementation, a coordinate descent algorithm is developed and the oracle property of the proposed method is established. The numerical studies indicate that the proposed approach works well for practical situations and it is applied to a set of real data arising from Alzheimer's Disease Neuroimaging Initiative study that motivated this study.

Keywords

Bernstein polynomials, case K interval-censored data, informative censoring, penalized procedure, variable selection

1 Introduction

It is well known that Cox's proportional hazards model is the most commonly used model for regression analysis of failure time data and many methods have been developed for its inference under various situations.[1–5] In particular, Cox[1] proposed a partial likelihood approach for the situation of right-censored data. In this paper, we discuss variable selection for the model when one faces a general type of failure time data, case *K* interval-censored data, which include right-censored and many other types of data as special cases.[6] By interval-censored data, we mean that the failure time of interest is known or observed only to belong to a window or an interval instead of being observed exactly or right-censored, and a general type of such data is the mixed or case *K* interval-censored data where there exists a sequence of observation times for each subject. It is easy to see that many medical studies such as clinical trials and medical follow-up studies can produce such data as well as many others, such as studies in demography, economics and reliability.[4,7]

A great amount of literature has been developed for covariate or variable selection and this is especially the case under the context of linear regression. In particular, many penalized procedures, which optimize an objective function with a penalty function, has recently been developed, including the least absolute shrinkage and selection operator (LASSO) procedure,[8] the smoothly clipped absolute deviation (SCAD) procedure,[9] the adaptive LASSO

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[2]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China
[3]Department of Statistics, University of Missouri, Columbia, MO, USA

Corresponding author:
Hui Zhao, School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China.
Email: hzhao@zuel.edu.cn

(ALASSO) procedure,[10] the smooth integration of counting and absolute deviation (SICA) procedure,[11] the seamless-L_0 (SELO) procedure,[12] and the broken adaptive ridge (BAR) regression.[13] A number of authors have also discussed the variable selection for right-censored failure time data, and especially, Tibshirani,[14] Fan and Li,[15] and Zhang and Lu[16] generalized the LASSO, SCAD, and ALASSO penalty-based procedures, respectively, to the Cox's proportional hazards model situation.

For the variable selection based on interval-censored failure time data, two parametric procedures were developed in Scolas et al.[17] and Wu and Cook,[18] and in particular, the latter assumed that the baseline hazard function is a piecewise function. Also Zhao et al.[19,20] considered a special case of interval-censored data, case II interval-censored data, and proposed a semiparametric procedure. Note that in terms of estimation and variable selection on the proportional hazards model, one significant difference between right-censored data and interval-censored data is that for the former, a simple partial likelihood function is available and commonly used as the objective function in a penalized procedure, while there does not exist such function for the latter and one has to work with a much more complicated objective function. Furthermore, a significant limitation for all of the variable selection procedures mentioned above for failure time data is that they apply only to the independent censoring situation. It is well known that sometimes the censoring, either right or interval censoring, may be informative and for this case, the use of the methods that assume the independent censoring can lead to biased results or even misleading conclusions.[3,4] In addition, it is also much more complicated to deal with informative interval censoring than informative right censoring as the former is a process and the latter can be simply characterized by a random variable.

In the following, we will consider case K interval-censored failure time data and develop a general or unified penalized variable selection procedure. In the proposed method, the sieve approach based on Bernstein polynomials will be employed to approximate the unknown baseline cumulative hazard function and a two-step estimation procedure will be developed. Also the latent variable will be used to describe the relationship between the failure time of interest and the observation process, and thus the proposed approach has the advantage that it allows dependent or informative censoring. In addition, it is flexible in that no distribution assumption is needed for the latent variables. Of course, instead of Bernstein polynomials and the latent variable approach, one may employ other smooth functions such as B-splines and the copula model approach, respectively, in the proposed method and more comments on these are given below.

The remainder of the paper is organized as follows. We will begin in Section 2 with introducing some notation and assumptions that will be used throughout the paper and then briefly discuss the method that would be used if only estimation is of interest. In particular, we will assume that the failure time of interest follows the proportional hazards model and a counting process is employed to describe the underlying censoring or observation process. In Section 3, the proposed sieve penalized variable selection procedure will be presented and in the method, the number of covariates is allowed to diverge with the sample size. For the implementation of the method, a coordinate descent algorithm is developed, and the oracle property of the proposed procedure is established in Section 4. The proposed method can be applied with various penalty functions, although we will focus on the BAR penalty function. Section 5 presents some results obtained from a simulation study conducted for the assessment of the proposed method and they suggest that the method works well for practical situations. An application to a motivated real study is provided in Section 6, and Section 7 contains some discussion and concluding remarks.

2 Notation, assumptions and estimation

Consider a failure time study that consists of n independent subjects and let T_i denote the failure time of interest associated with subject i. Also for subject i, suppose that there exist a p-dimensional vector of covariate denoted by z_i and a sequence of observation time points denoted by $U_{i0} = 0 < U_{i1} < U_{i2} < \ldots < U_{iK_i}$, where K_i is a random integer, $i = 1, \ldots, n$. Define $\tilde{N}_i(t) = \sum_{j=1}^{K_i} I(U_{ij} \le t)$ and $\delta_{ij} = I(U_{i,j-1} < T_i \le U_{i,j})$, $i = 1, \ldots, n, j = 1, \ldots, K_i$. Then, $\tilde{N}_i(t)$ is a point process characterizing the observation process on subject i and jumps only at the observation times. In the following, we will assume that the observed data have the form

$$\{O_i = (\tau_i, U_{ij}, \delta_{ij}, z_i, j = 1, \ldots, K_i), i = 1, \ldots, n\}$$

where τ_i denotes a follow-up time for the ith subject that is assumed to be independent of T_i. That is, we only have case K interval-censored data.

To describe the covariate effect on T_i, suppose that for subject i, there exists a latent variable u_i and given z_i and u_i, T_i follows the proportional hazards frailty model

$$\lambda_i(t|z_i, u_i) = \lambda_0(t)\exp(x_i^T\boldsymbol{\beta}) \tag{1}$$

In the above, $\lambda_0(t)$ denotes an unknown baseline hazard function, $x_i = (u_i, z_i^T)^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ are unknown regression parameters. Furthermore, we will assume that given z_i and u_i, T_i and $\tilde{N}_i(t)$ are independent and $\tilde{N}_i(t)$ is a nonhomogeneous Poisson process with the intensity function

$$\lambda_{ih}(t|z_i, u_i) = \lambda_{0h}(t)\exp(z_i^T\boldsymbol{\alpha} + u_i) \tag{2}$$

Here $\lambda_{0h}(t)$ denotes a completely unknown continuous baseline intensity function and $\boldsymbol{\alpha}$ is a vector of regression parameters as $\boldsymbol{\beta}$. It is apparent that under models (1) and (2), the parameter β_0 represents the extent of the association between the failure time and the observation process. The two will be independent if $\beta_0 = 0$. In addition, the positive value of β_0 means that the failure time and the observations are positively correlated, while the negative value of β_0 means the negative association. Note that instead of one latent variable in the models above, one could replace it by two correlated latent variables and the development below would be still valid.

If one is only interested in estimation of regression parameters, under the assumptions above, it would be natural to employ the conditional likelihood function

$$L(\boldsymbol{\beta}, \Lambda_0|u_i's) = \prod_{i=1}^{n}\left\{\prod_{j=1}^{K_i}(\exp(-\Lambda_0(U_{ij-1})\exp(x_i^T\boldsymbol{\beta})) - \exp(-\Lambda_0(U_{ij})\exp(x_i^T\boldsymbol{\beta})))^{\delta_{ij}} \times (\exp(-\Lambda_0(U_{iK_i})\exp(x_i^T\boldsymbol{\beta})))^{1-\sum_{j=1}^{K_i}\delta_{ij}}\right\}$$

given the $u_i's$, where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, the baseline cumulative hazard function. Of course, in general, the u_i's are unknown and for this, Wang et al.[6] proposed a two-step estimation procedure that employs the strength-borrowing method discussed in Huang and Wang[21] and replaces them by their estimators.

More specifically, let $\Lambda_{0h}(t) = \int_0^t \lambda_{0h}(s)ds$ and assume that $\Lambda_{0h}(\tau_0) = 1$, where τ_0 denotes the longest follow-up time. Then Wang et al.[6] suggested to estimate $\boldsymbol{\alpha}$ by using the estimating equations

$$U(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\tilde{z}_i\left(K_i\hat{\Lambda}_{0h}^{-1}(\tau_i) - E(e^{u_i})\exp(z_i^T\boldsymbol{\alpha})\right) = 0$$

where $\tilde{z}_i^T = (1, z_i^T)$, and

$$\hat{\Lambda}_{0h}(t) = \prod_{s_{(l)}>t}\left(1 - \frac{d_{(l)}}{R_{(l)}}\right)$$

the estimation of Λ_{0h}. Let $\hat{\boldsymbol{\alpha}}$ denote the estimator of $\boldsymbol{\alpha}$ given by the estimating equation above. Then one can naturally estimate u_i by

$$\hat{u}_i = \log\left\{\frac{K_i}{\hat{\Lambda}_{0h}(\tau_i)\exp(z_i^T\hat{\boldsymbol{\alpha}})}\right\}$$

and estimate $\boldsymbol{\beta}$ and Λ_0 by maximizing $L(\boldsymbol{\beta}, \Lambda_0|\hat{u}_i's)$.

3 Sieve penalized variable selection procedure

Now we consider both estimation and covariate selection with the focus on model (1). For this, motivated by the discussion in the previous section, we propose to maximize the penalized estimated log-likelihood

function

$$l_p(\boldsymbol{\beta}, \Lambda_0|\hat{u}_i's) = \log L(\boldsymbol{\beta}, \Lambda_0|\hat{u}_i's) - \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|)$$

where p_{λ_n} denotes a penalty function that depends on a tuning parameter $\lambda_n > 0$. For the penalty function, in the following, we will consider several commonly used ones, including the LASSO,[14] ALASSO,[16] SCAD,[9] SICA,[11] SELO[12] and BAR[22] penalty functions.

Note that for the maximization of $l_p(\boldsymbol{\beta}, \Lambda_0|\hat{u}_i's)$, one challenge will be that $l_p(\boldsymbol{\beta}, \Lambda_0|\hat{u}_i's)$ involves the infinite-dimensional function Λ_0. To deal with this, by following others,[5] we propose first to employ the sieve approach to approximate Λ_0 by using Bernstein polynomials. More specifically, let

$$\Theta = \{\nu = (\boldsymbol{\beta}, \Lambda_0) \in \mathbb{B} \otimes \mathbb{M}\}$$

denote the parameter space of ν, where $\mathbb{B} = \{\boldsymbol{\beta}|\boldsymbol{\beta} \in R^p, ||\boldsymbol{\beta}|| \leq M\}$ with M being a positive constant and $\mathbb{M}$ is the collection of all bounded and continuous nondecreasing, non-negative functions over the interval $[c, u]$ with $0 \leq c < u < \infty$. In practice, $[c, u]$ is usually taken as the range of observed data. Furthermore, define the sieve space

$$\Theta_n = \{\nu_n = (\boldsymbol{\beta}, \Lambda_{0n}) \in \mathbb{B} \otimes \mathbb{M}_n\}$$

where

$$\mathbb{M}_n = \left\{ \Lambda_{0n}(t) = \sum_{k=0}^{m} \phi_k^* B_k(t, m, c, u) : \sum_{0 \leq k \leq m} |\phi_k^*| \leq M_n, 0 \leq \phi_0^* \leq \phi_1^* \leq \ldots \leq \phi_m^* \right\}$$

with M_n being a constant and

$$B_k(t, m, c, u) = C_m^k \left(\frac{t-c}{u-c}\right)^k \left(1 - \frac{t-c}{u-c}\right)^{m-k}, \quad k = 0, \ldots, m$$

which are Bernstein basis polynomials of degree $m = o(n^s)$ for some $s \in (0, 1)$. Note that M_n controls the size of the sieve space and is usually chosen as $M_n = o(n^a)$ with $a \in (0, 1)$.[23] The value of m can be chosen by the cross-validation method, and in practice, one may perform grid search over some possible range of m or fix m to be the closest integer to $n^{0.25}$.[19] Note that here Bernstein polynomials are chosen simply because of their natural monotone property and simplicity. Instead it is apparent that one could alternatively use other smooth functions such as B-spline functions and the proposed method could be similarly developed.

By focusing on the sieve space Θ_n, one can rewrite the penalized estimated log-likelihood function $l_p(\boldsymbol{\beta}, \Lambda_0|\hat{u}_i's)$ as

$$l_p(\boldsymbol{\beta}, \phi^*|\hat{u}_i's) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K_i} \delta_{ij} \log(\exp(-\Lambda_{0n}(U_{i,j-1})\exp(\hat{x}_i^T\boldsymbol{\beta})) - \exp(-\Lambda_{0n}(U_{ij})\exp(\hat{x}_i^T\boldsymbol{\beta}))) - (1 - \sum_{j=1}^{K_i} \delta_{ij})\Lambda_{0n}(U_{iK_i})\exp(\hat{x}_i^T\boldsymbol{\beta}) \right\} - \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|)$$

where $\hat{x}_i = (\hat{u}_i, z_i^T)^T$. Note that due to the non-negative and non-decreasing constraint of the cumulative baseline hazard function Λ_0, in the maximization above, the constraint $0 \leq \phi_0^* \leq \phi_1^* \leq \ldots \leq \phi_m^*$ is required but it can be easily removed by the reparameterization $\phi_0^* = e^{\phi_0}, \phi_k^* = \sum_{i=0}^{k} e^{\phi_i}, \forall 1 \leq k \leq m$. To maximize $l_p(\boldsymbol{\beta}, \phi|\hat{u}_i's)$, we will employ an alternative algorithm given below that estimates $\boldsymbol{\beta}$ and ϕ alternately. In particular, we will use the Nelder-Mead simplex algorithm to update the estimator of ϕ given the current estimator of $\boldsymbol{\beta}$ and then update the estimator of $\boldsymbol{\beta}$ by employing the coordinate descent algorithm while fixing the ϕ. Specifically,

Step 1. Choose the initial values $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\phi}^{(0)}$ for both $\boldsymbol{\beta}$ and ϕ.

Step 2. At the kth iteration, given the current $\hat{\boldsymbol{\beta}}^{(k-1)}$, obtain $\hat{\phi}^{(k)}$ by using the Nelder-Mead simplex algorithm.

Step 3. Given the current estimate of $\hat{\phi}^{(k)}$, update the estimate of $\boldsymbol{\beta}$ by using the coordinate descent algorithm or update each element of $\boldsymbol{\beta}$ by maximizing $l_p(\boldsymbol{\beta}, \hat{\phi}^{(k)}|\hat{u}_i's)$ while holding the other elements of $\boldsymbol{\beta}$ fixed.

Step 4. Repeat steps 2 and 3 until convergence.

Note that the algorithm above can apply to any penalty function. On the other hand, one may employ an alternative for the BAR penalty $p_{\lambda_n}(|\beta_j|) = \lambda_n \beta_j^2/\tilde{\beta}_j^2$ with $\tilde{\beta}_j$ denoting a nonzero "good" estimate of β_j. More specifically, one can use the ridge regression estimator

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\max_{\boldsymbol{\beta}} \left\{ l_p(\boldsymbol{\beta}|\hat{u}) - \xi_n \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3}$$

as the initial value $\hat{\boldsymbol{\beta}}^{(0)}$ and then update $\hat{\boldsymbol{\beta}}^{(k-1)}$ iteratively by the following reweighed L_2-penalized estimator

$$\hat{\boldsymbol{\beta}}^{(k)} = \arg\max_{\boldsymbol{\beta}} \left\{ l_p(\boldsymbol{\beta}|\hat{u}) - \lambda_n \sum_{j=1}^{p} \frac{\beta_j^2}{\left(\hat{\beta}_j^{(k-1)}\right)^2} \right\}$$

In equation (3), ξ_n denotes another nonnegative tuning parameter to be discussed below.

Note that the coordinate algorithm described above is essentially to conduct univariate maximization for each element of $\boldsymbol{\beta}$ vector repeatedly, and for each univariate maximization, one can use the golden-section search algorithm[24] or Newton-Raphson algorithm. In the numerical studies reported below, the algorithm seems to work well and we did not have any convergence issues. On the covariate selection, at the convergence, we will set the estimates of the components of $\boldsymbol{\beta}$ whose values are less than a pre-specified threshold of zero. For the numerical studies reported below, we used the threshold of 10^{-6} by following Wang et al.[25] and 0 as the initial values for both $\boldsymbol{\beta}$ and ϕ by following Fan and Lv[26] and Lin and Lv.[27] Also for numerical study below, we implement the Nelder-Mead algorithm by using the R function *optim* and employ the R function *optimize* for the implementation of the golden-section search algorithm.

To implement the variable selection procedure described above, it is apparent that one needs to choose the two tuning parameters λ_n and ξ_n. For this, by following others, we suggest to use the C-fold cross-validation, which the numerical study below indicated works well. Also as pointed out by others and shown in the numerical study, the BAR-based approach is not sensitive to ξ_n and thus it can be taken to be a constant. Specifically, let C be an integer and suppose that the observed data can be divided into C non-overlapping parts with approximately the same size. Also let l^c denote the observed log-likelihood function based on the cth part of the whole data set and $\hat{\boldsymbol{\beta}}^{-c}$ and $\hat{\phi}^{-c}$ the proposed sieve penalized estimates of $\boldsymbol{\beta}$ and ϕ, respectively, obtained based on the whole data without the cth part. For given λ_n, the cross-validation statistics can be defined as

$$CV(\lambda_n) = \sum_{c=1}^{C} l^c\left(\hat{\boldsymbol{\beta}}^{-c}, \hat{\phi}^{-c}\right)$$

and one can choose the value of λ_n that maximizes $CV(\lambda_n)$.

4 Asymptotic properties

Now we discuss the oracle property of the variable selection procedure described in the previous sections with the focus on the use of the BAR penalty function. Let $\hat{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k)}$ denote the BAR estimator given by the procedure above and $\boldsymbol{\beta}_0 = (\beta_{0,0}, \beta_{0,1}, \ldots, \beta_{0,p})^T$ the true value of $\boldsymbol{\beta}$. Without loss of generality, assume that we can write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \boldsymbol{\beta}_{02}^T)^T$, where $\boldsymbol{\beta}_{01}$ is a $q+1$ vector consisting of $\beta_{0,0}$ and all q ($q \ll p$) nonzero components and $\boldsymbol{\beta}_{02}$ the remaining zero components. Correspondingly, we denote the BAR estimator of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$. In the following, we assume that $p < n$ but p and q can diverge or increase with the sample size n.

To establish the oracle property of the BAR estimator, we need the following regularity conditions.

(C1). (i) The set $\mathcal{B}$ is a compact subset of $\mathcal{R}^{p+1}$ and $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$. (ii) The matrix $E(\hat{x}\hat{x}^T)$ is non-singular with $\hat{x}$ being bounded. That is, there exists $\hat{x}_0 > 0$ such that $P(||\hat{x}|| \le \hat{x}_0) = 1$.

(C2). The function $\Lambda_0(\cdot)$ is continuously differentiable up to order r in $[u, v]$ and satisfies $a^{-1} < \Lambda_0(u) < \Lambda_0(v) < a$ for some positive constant a.

(C3). There exists a compact neighborhood $\mathcal{B}_0$ of the true value $\boldsymbol{\beta}_0$ and a positive definite $(p+1) \times (p+1)$ matrix $I(\boldsymbol{\beta}_0)$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}_0} || - n^{-1} \ddot{l}_n(\boldsymbol{\beta}) - I(\boldsymbol{\beta}_0)|| \xrightarrow{a.s.} 0$$

where $\ddot{l}_n(\boldsymbol{\beta})$ is the second derivative of $l_p(\boldsymbol{\beta}|\hat{u})$.

(C4). There exists some constant $C > 1$ such that $C^{-1} < \lambda_{min}(I(\boldsymbol{\beta}_0)) \leq \lambda_{max}(I(\boldsymbol{\beta}_0)) < C$ for sufficiently large n, where $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ denote the smallest and largest eigenvalues of the matrix, respectively.

(C5). There exist positive constants a_0 and a_1 such that $a_0 \leq |\beta_{0,j}| \leq a_1$, $1 \leq j \leq q$.

(C6). As $n \to \infty$, $p^2 q/\sqrt{n} \to 0$, $\lambda_n/\sqrt{n} \to 0$, $\xi_n/\sqrt{n} \to 0$, $\lambda_n \sqrt{q/n} \to 0$ and $\lambda_n^2/(p\sqrt{n}) \to \infty$.

Conditions (C1) to (C3) are necessary for the existence and consistence of the sieve estimator of $\Lambda_0(t)$ and usually satisfied in practice. Condition (C4) assumes that $I(\boldsymbol{\beta}_0)$ is positive definite almost surely and its eigenvalues are bounded away from zero and infinity, and Condition (C5) assumes that the nonzero coefficients are uniformly bounded away from zero and infinity. Condition (C6) gives some sufficient but not necessary conditions needed to prove the numerical convergence and asymptotic properties of the BAR estimator $\hat{\boldsymbol{\beta}}$. To establish the oracle property, for a vector of $\boldsymbol{\theta}_1$ and given $\boldsymbol{\beta}_1$, define $Q_{n1}(\boldsymbol{\theta}_1) = Q_{n1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) = l_{p1}(\boldsymbol{\theta}_1) - \lambda_n \boldsymbol{\theta}_1^T D_1(\boldsymbol{\beta}_1)\boldsymbol{\theta}_1$, where $l_{p1}(\boldsymbol{\theta}_1) = l_p(\boldsymbol{\theta}_1, 0|\hat{u}_i's)$ and $D_1(\boldsymbol{\beta}_1) = \text{diag}\{0, \beta_1^{-2}, \ldots, \beta_q^{-2}\}$. Then the oracle property can be described as follows with the proof given in Appendix 1.

Theorem 1. *Assume that the regularity conditions (C1) to (C6) described above hold. Then as $n \to \infty$ and with probability tending to 1, the BAR estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ exists and has the following properties:*

(i) $\hat{\boldsymbol{\beta}}_2 = 0$.

(ii) $\hat{\boldsymbol{\beta}}_1$ *is the unique fixed point of* $f(\boldsymbol{\beta}_1)$, *where* $f(\boldsymbol{\beta}_1)$ *is a solution to* $\dot{Q}_{n1}(\boldsymbol{\theta}_1) = 0$, *wherein* $\dot{Q}_{n1}(\boldsymbol{\theta}_1)$ *is the first derivative of* $Q_{n1}(\boldsymbol{\theta}_1)$.

(iii) $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})$ *converges in distribution to the multivariate normal distribution* $N_{q+1}(0, I_1(\boldsymbol{\beta}_0)^{-1})$, *where* $I_1(\boldsymbol{\beta}_0)$ *denotes the up-left* $(q+1) \times (q+1)$ *submatrix of* $I(\boldsymbol{\beta}_0)$.

5 A simulation study

In this section, we present some results obtained from an extensive simulation study conducted to assess the performance of the variable selection approach proposed in the previous sections. In the study, we first generated the covariates z_i's from the multivariate normal distribution with mean zero, variance one, and the correlation between z_j and z_k being $\rho^{|j-k|}$ with $\rho = 0.5$, $j, k = 1, \ldots, p$, and the latent variables u_i's through assuming that $u_i^* = \exp(u_i)$ follows the gamma distribution with mean 4 and variance 8. The failure times of interest were then generated from model (1) with $\Lambda_0(t) = t$ or $\Lambda_0(t) = t^2$ and it was assumed that the $\tau_i's$ follow the uniform distribution over the interval $[3, 4]$.

For the observation process, it was supposed that $\tilde{N}_i(t)$ follows model (2) with $\lambda_{0h} = 1/4$. Then, given z_i, u_i and τ_i, K_i, the number of observation times for subject i was generated from the Poisson distribution with mean

$$\Lambda_{ih}(\tau_i|z_i, u_i) = \frac{\tau_i \exp(z_i^T \boldsymbol{\alpha} + u_i)}{4}$$

and the observation times $(U_{i1}, \ldots, U_{iK_i})$ were taken to be the order statistics of a random sample of size K_i from the uniform distribution over $(0, \tau_i)$, $i = 1, 2, \ldots, n$. The results given below are based on $n = 100$ or 300 with 100 replications.

Table 1 presents the results obtained on the covariate selection with $n = 100$ or 300, $p = 8$, $\boldsymbol{\beta} = (0.2, 1, 1, 0, 0, 0, 0, 0, 1)^T$, and all components of $\boldsymbol{\alpha}$ being 0.1. In the table, we calculated the median (MMSE) of the mean weighted squared errors (MSE) defined to be $(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*)^T \Sigma (\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*)$ with $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_p)$ and the standard deviation (SD) of the MSE, where Σ denotes the covariance matrix of the covariates given at the beginning of this section. Also, we computed the average number of the correctly selected covariates whose true coefficient are not zero (TP) and the average number of incorrectly selected covariates whose true coefficients

Table 1. Simulation results with $n = 100$ or 300 and, $p = 8$ and $\Lambda_0(t) = t$.

Penalty	MMSE(SD)	TP	FP
	$n = 100$		
LASSO	0.283(0.450)	3	2.07
ALASSO	0.232(0.687)	3	0.79
SCAD	0.183(0.983)	2.97	0.39
SELO	0.193(1.145)	2.81	0.27
SICA	0.233(0.883)	2.83	0.29
BAR	0.142(0.757)	3	0.20
	$n = 300$		
LASSO	0.093(0.073)	3	2.20
ALASSO	0.061(0.077)	3	0.67
SCAD	0.040(0.070)	3	0.24
SELO	0.038(0.173)	2.98	0.22
SICA	0.040(0.169)	2.97	0.13
BAR	0.039(0.053)	3	0.24

Table 2. Simulation results with $n = 300$, $p = 30$ or 50 and $\Lambda_0(t) = t$.

Penalty	MMSE(SD)	TP	FP
	$p = 30$		
LASSO	0.305(0.152)	4	7.55
ALASSO	0.149(0.134)	4	2.67
SCAD	0.060(0.196)	4	0.51
SELO	0.067(0.178)	3.97	0.18
SICA	0.063(0.204)	3.98	0.39
BAR	0.078(0.102)	4	0.33
	$p = 50$		
LASSO	1.215(0.586)	8	13.14
ALASSO	0.403(0.454)	8	4.32
SCAD	0.262(1.313)	7.99	0.10
SELO	0.292(0.493)	7.90	0.12
SICA	0.313(0.499)	7.87	0.13
BAR	0.215(0.267)	8	0.45

are zero (FP). Here we considered six penalty functions, LASSO, ALASSO, SCAD, SELO, SICA, and BAR, and for the results, we took m, the degree of Bernstein polynomials, to be 3. For the selection of the tuning parameter λ_n, the 5-fold cross-validation based on the grid search was used, and for ξ_n in the BAR penalty, we set $\xi_n = 100$ since, as mentioned above, the results are not sensitive to the choice of ξ_n.

The results given in Table 2 were obtained in the same way as above except that $n = 300$, $p = 30$ or $p = 50$, $\boldsymbol{\beta} = (0.2, 1, 1, \mathbf{0}_{p-4}, 1, 1)^T$ or $(0.2, 1, 1, 1, 1, \mathbf{0}_{p-8}, 1, 1, 1, 1)^T$. Table 3 gives the results obtained under the same set-up as that used for Table 2 expect that $\Lambda_0(t) = t^2$. One can see from Tables 1 to 3 that the proposed procedure seems to perform well no matter which penalty function was used, especially in terms of TP, measuring the true positive selection. In terms of MMSE and FP, measuring the false positive selection, as expected, the proposed methods with ALASSO, SCAD, SELO, SICA and BAR seem to give better performance than with LASSO.

Note that in the proposed variable selection procedure, it has been assumed that the observation process follows the non-homogeneous Poisson process and it is apparent that sometimes this may not be true. To assess the robustness of the procedure with respect to the assumption, we repeated the study above in the same way except generating the observation times from a renew process. More specifically, the gap times were set to be $4\exp(-z_i\boldsymbol{\alpha} - u_i)v_i$ with v_i generated from the uniform distribution over $[0, 2]$ until the summation of the generated gap times being larger than τ_i. Table 4 gives the results obtained under the set-up similar to that in Table 1 with $n = 300$ and $p = 8$, and one can see that they gave similar conclusions as before. We also considered other set-ups and obtained similar results.

Table 3. Simulation results with $n = 300$, $p = 30$ or 50 and $\Lambda_0(t) = t^2$.

Penalty	MMSE(SD)	TP	FP
	$p = 30$		
LASSO	0.277(0.181)	4	7.16
ALASSO	0.114(0.149)	4	3.31
SCAD	0.044(0.261)	4	0.26
SELO	0.048(0.267)	3.96	0.10
SICA	0.049(0.186)	3.98	0.17
BAR	0.044(0.074)	4	0.27
	$p = 50$		
LASSO	1.468(0.910)	8	13.65
ALASSO	0.364(0.998)	8	5.38
SCAD	0.200(0.906)	7.94	0.15
SELO	0.208(0.994)	7.89	0.10
SICA	0.202(0.909)	7.95	0.14
BAR	0.191(0.358)	8	0.39

Table 4. Simulation results with $n = 300$, $p = 8$ and the renew observation process.

Penalty	MMSE(SD)	TP	FP
LASSO	0.134(0.096)	3	2.26
ALASSO	0.059(0.089)	3	0.59
SCAD	0.045(0.126)	3	0.43
SELO	0.055(0.144)	2.99	0.30
SICA	0.046(0.187)	2.97	0.24
BAR	0.048(0.101)	3	0.34

6 An application

Now we apply the methodology proposed in the previous sections to a set of real data arising from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal follow-up study that started in 2004 and was designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of the Alzheimer's disease (AD). In the study, the participants were examined periodically for their AD status and the collection of related information and also were initially grouped based on their cognitive conditions into three groups, cognitively normal, mild cognitive impairment and Alzheimer's disease. Among others, one variable of interest is the time (in year) from the baseline visit date to the AD conversion. As expected, many patients dropped out of the study early and some missed their scheduled visits. Thus, the patients have different observation times and only case K interval-censored data are available on the AD conversion time.

For the analysis here, by following Li et al.,[28] we will consider 292 participants with mild cognitive impairment initial status and for whom the information on 24 covariates is complete to identify important prognostic factors for the AD conversion. These 24 demographic and clinical covariates were identified as possible important factors associated with the AD conversion by Li et al.,[28] who considered a similar problem by performing a simple or individual analysis. In addition to the information on the covariates, the observed data for each participant include the number of observations K_i, the observation times U_{ij}'s, and the event indicators δ_{ij}'s. For the analysis results given below, as in the simulation study, we considered six penalty functions, LASSO, ALASSO, SCAD, SICA, SELO and BAR. Also, as in the previous section, the 5-fold cross-validation was used to select the optimal λ_n with setting $\xi_n = 100$ and $m = 3$, the degrees of freedom for the Bernstein polynomial approximation.

Table 5 presents the selection results and for the selected covariates, it also gives the estimated covariate effects along with the estimated standard errors, given in the parentheses and obtained by using the bootstrap procedure with 100 bootstrap samples randomly drawn with replacement from the data. One can see from the table that six

Table 5. Variable selection results in ADNI study.

Covariate	LASSO	ALASSO	SCAD	SELO	SICA	BAR
Age	$-0.229_{(0.110)}$	$-0.228_{(0.154)}$	$-0.279_{(0.153)}$	$-0.362_{(0.172)}$	$-0.356_{(0.178)}$	$-0.322_{(0.152)}$
Gender	–(–)	$-0.221_{(0.238)}$	$-0.383_{(0.287)}$	–(–)	–(–)	–(–)
Years of education	$0.041_{(0.094)}$	$0.031_{(0.083)}$	$0.088_{(0.099)}$	–(–)	–(–)	$0.077_{(0.110)}$
Marital status	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
APOE-ϵ4	$0.265_{(0.134)}$	$0.419_{(0.168)}$	$0.467_{(0.182)}$	–(–)	–(–)	$0.471_{(0.170)}$
CDR-SB	$0.095_{(0.109)}$	$0.172_{(0.163)}$	$0.203_{(0.150)}$	$0.123_{(0.127)}$	$0.099_{(0.127)}$	$0.211_{(0.177)}$
ADAS11	$0.106_{(0.132)}$	–(–)	$0.132_{(0.230)}$	–(–)	–(–)	–(–)
ADAS13	$0.176_{(0.135)}$	$0.365_{(0.266)}$	$0.099_{(0.121)}$	$0.327_{(0.132)}$	$0.372_{(0.147)}$	$0.276_{(0.173)}$
ADASQ4	$0.033_{(0.101)}$	–(–)	$0.085_{(0.137)}$	–(–)	–(–)	–(–)
MMSE	$-0.087_{(0.096)}$	–(–)	$-0.091_{(0.128)}$	–(–)	–(–)	$-0.126_{(0.124)}$
RAVLT.i	$-0.294_{(0.139)}$	$-0.337_{(0.222)}$	$-0.394_{(0.257)}$	$-0.469_{(0.219)}$	$-0.437_{(0.228)}$	$-0.577_{(0.212)}$
RAVLT.l	–(–)	$0.286_{(0.209)}$	$0.401_{(0.232)}$	–(–)	–(–)	$0.231_{(0.269)}$
RAVLT.f	–(–)	$-0.368_{(0.276)}$	$-0.485_{(0.313)}$	–(–)	–(–)	–(–)
RAVLT.p.f	$0.101_{(0.235)}$	$0.547_{(0.307)}$	$0.649_{(0.374)}$	–(–)	–(–)	$0.142_{(0.328)}$
DIGITSCOR	–(–)	–(–)	$-0.080_{0.114}$	–(–)	–(–)	–(–)
TRABSCOR	$0.012_{(0.094)}$	–(–)	–(–)	–(–)	–(–)	–(–)
FAQ	$0.181_{(0.122)}$	$0.316_{(0.163)}$	$0.317_{(0.181)}$	$0.273_{(0.163)}$	$0.205_{(0.179)}$	$0.358_{(0.198)}$
Ventricles	–(–)	–(–)	$-0.053_{(0.172)}$	–(–)	–(–)	–(–)
Hippocampus	$-0.242_{(0.168)}$	$-0.086_{(0.192)}$	$-0.172_{(0.230)}$	$-0.283_{(0.233)}$	$-0.433_{(0.263)}$	$-0.070_{(0.182)}$
WholeBrain	–(–)	–(–)	$0.004_{(0.193)}$	–(–)	–(–)	–(–)
Entorhinal	$-0.114_{(0.123)}$	$-0.233_{(0.151)}$	$-0.264_{(0.192)}$	$-0.253_{(0.188)}$	–(–)	$-0.306_{(0.165)}$
Fusiform	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
MidTemp	$-0.408_{(0.163)}$	$-0.604_{(0.247)}$	$-0.576_{(0.238)}$	$-0.545_{(0.278)}$	$-0.527_{(0.323)}$	$-0.652_{(0.228)}$
ICV	$0.205_{(0.140)}$	$0.321_{(0.215)}$	$0.420_{(0.262)}$	$0.312_{(0.218)}$	$0.292_{(0.227)}$	$0.311_{(0.254)}$

factors, Age, APOE-ϵ4, ADAS13, RAVLT.i, FAQ and MidTemp, seem to have had some significant prognostic effects on the AD conversion, and two factors, Marital status and Fusiform, were not selected by any penalty function. Also three factors, CDR-SB, Hippocampus and ICV, were selected by all penalty functions and may have some mild effects on the AD conversion. The conclusions above are similar to those given by the previous simplified or single variable analyses.[28]

7 Discussion and concluding remarks

This paper discussed the variable or covariate selection problem when one faces a general type of failure time data and a unified variable selection procedure was proposed under the Cox's proportional hazards model. The proposed penalized approach can accommodate any penalty function and makes use of the sieve approach. Unlike the existing variable selection procedures for failure time data, one major advantage of the approach is that it allows for the dependent or informative censoring, which can easily occur in medical studies as well as other studies and it has been shown to cause biased estimation or misleading inference conclusions if not taken into account. For the implementation, a coordinate descent algorithm was developed and the proposed procedure with the use of the BAR penalty function was shown to have the oracle property. In addition, the numerical studies indicated that the proposed approach works well for practical situations.

A main contribution of the proposed variable selection procedure is that it applies to various types of failure time data, including right-censored data and case II interval-censored data.[20] As discussed above, although right-censored data and interval-censored may seem to be similar, the latter has much more complicated structures and thus their analysis is also much more difficult than the former. For example, with the former, a partial likelihood function is available for regression analysis under the proportional hazards model, and in contrast, one has to work with some complex full likelihood functions for the latter. In terms of the censoring or observation process, for the former, it can simply be described by one censoring variable, while one has to use and deal with two censoring variables or a stochastic process for the latter.

Note that the proposed approach is essentially a two-step procedure in terms of estimation of models (1) and (2) or parameters. Instead of this, one may consider an alternative method that makes use of the observed full likelihood function as the objective function in the proposed penalized procedure. It is easy to see that this would

be much more complicated in terms of implementation and also it would be much more difficult to establish the asymptotic properties of the resulting variable selection procedure. In addition, unlike the proposed approach, one would need to make some assumptions about the distribution of the latent variables.

In the preceding sections, the focus has been on the proportional hazards model and it is apparent that sometimes a different model may be preferred or more appropriate such as the additive hazards model or the linear transformation model. It is easy to see that the idea discussed above can still be applied to these situations but the development of a new implementation algorithm may be needed and also the derivation of the asymptotic property of the resulting variable selection approach may be different. To describe the correlation between the failure time variable of interest and the observation process, the latent variable approach was employed above and as mentioned before, an alternative is to employ the copula model approach.[5,29,30] One advantage of the latter is that it allows for the direct estimation of the association but it usually requires more assumptions that cannot usually be verified based on available information.

Authors' note

Mingyue Du is now affiliated with Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China.

Acknowledgements

The authors wish to thank the Associate Editor and two reviewers for their many comments and suggestions that greatly improved the article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Hui Zhao https://orcid.org/0000-0003-0070-0130

References

1. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B* 1972; **34**: 187–220.
2. Jewell NP. Non-parametric estimation and doubly-censored data: general ideas and applications to AIDS. *Stat Med* 1994; **13**: 2081–2095.
3. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*. 2nd ed. New York: Wiley, 2002.
4. Sun J. *The statistical analysis of interval-censored failure time data*. New York: Springer, 2006.
5. Zhou Q, Hu T and Sun J. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *J Am Stat Assoc* 2017; **112**: 664–672.
6. Wang PJ, Zhao H and Sun J. Regression analysis of case K interval-censored failure time data in the presence if informative censoring. *Biometrics* 2016; **72**: 1103–1112.
7. Chen D, Sun J and Peace KE. *Interval-censored time-to-event data: methods and applications*. London: Chapman & Hall/CRC, 2012.
8. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* 1996; **58**: 267–288.
9. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle property. *J Am Stat Assoc* 2001; **96**: 1348–1360.
10. Zou H. The adaptive Lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
11. Lv J and Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. *Ann Stat* 2009; **37**: 3498–3528.
12. Dicker L, Huang B and Lin X. Variable selection and estimation with seamless-L_0 penalty. *Statist Sinica* 2013; **23**: 929–962.
13. Liu Z and Li G. Efficient regularized regression with L_0 penalty for variable selection and network construction. *Comput Math Meth Med*; 2016: Article ID 3456153. DOI: 10.1155/2016/3456153
14. Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med* 1997; **16**: 385–395.
15. Fan J and Li R. Variable selection for Cox's proportional hazards model and frailty model. *Ann Stat* 2002; **30**: 74–99.
16. Zhang H and Lu WB. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 2007; **94**: 1–13.

17. Scolas S, EI Ghouch A, Legrand C, et al. Variable selection in a flexible parametric mixture cure model with interval-censored data. *Stat Med* 2016; **35**: 1210–1225.
18. Wu Y and Cook R. Penalized regression for interval-censored times of disease progression: selection of HLA markers in psoriatic arthritis. *Biometrics* 2015; **71**: 782–791.
19. Zhao H, Wu Q, Gilbert PB, et al. A regularized estimation approach for case-cohort periodic follow-up studies with an application to HIV vaccine trials. *Biom J* 2020; **62**: 1176–1191.
20. Zhao H, Wu Q, Li G, et al. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *J Am Stat Assoc* 2020; **115**: 204–216.
21. Huang CY and Wang MC. Joint modeling and estimation for recurrent event processes and failure time data. *J Am Stat Assoc* 2004; **99**: 1153–1165.
22. Dai L, Chen K, Sun Z, et al. Broken adaptive ridge regression and its asymptotic properties. *J Multivar Anal* 2018; **168**: 334–351.
23. Shen X. On methods of sieves and penalization. *Ann Stat* 1997; **25**: 2555–2591.
24. Kiefer J. Sequential minimax search for a maximum P. *Am Math Soc* 1953; **4**: 502–506.
25. Wang H, Li R and Tsai C. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 2007; **94**: 553–568.
26. Fan J and Lv J. Nonconcave penalized likelihood with NP-dimensionality. *IEEE T Inform Theory* 2011; **57**: 5467–5484.
27. Lin W and Lv J. High-dimensional sparse additive hazards regression. *J Am Stat Assoc* 2013; **108**: 247–264.
28. Li K, Chan W, Doody RS, et al. Prediction of conversion to alzheimer's disease with longitudinal measures and time-to-event data. *J Alzheimer Dis* 2017; **58**: 361–371.
29. Hu T, Zhou Q and Sun J. Regression analysis of bivariate current status data under the proportional hazards model. *Canad J Stat* 2017; **45**: 410–424.
30. Ma L, Hu T and Sun J. Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* 2015; **102**: 731–738.
31. Cai J, Fan J, Li R, et al. Variable selection for multivariate failure time data. *Biometrika* 2005; **92**: 303–316.

Appendix 1. Proof of Theorem 1

In this appendix, we will sketch the proof of Theorem 1. For this and also the completeness, we will first describe the regularity conditions needed for the asymptotic properties of $\hat{\Lambda}_{0h}(t)$ and $\hat{\boldsymbol{\alpha}}$ and the properties.

(A1) For the follow-up time τ and latent variable u, we have $P(\tau \geq \tau_0, \exp(u) > 0) > 0$.
(A2) The covariate z is uniformly bounded.
(A3) For the latent variable u, the variance of $\exp(u)$ is bounded and there exists a positive small constant $\epsilon > 0$ such that $\exp(u) > \epsilon$ almost surely.
(A4) Also for τ and u, the function $G(s) = E[\exp(u)I(\tau \geq s)]$ is continuous for $s \in [0, \tau_0]$.

It has been pointed out [6,21] that if the regularity conditions (A1) to (A4) hold, then as $n \to \infty$ and with probability tending to 1, $\hat{\Lambda}_{0h}(t)$ and $\hat{\boldsymbol{\alpha}}$ are consistent and possess the asymptotical normality. Hence, one can treat the functions of the u_i's as the ones as if the u_i's were observed and rewrite model (1) as

$$\lambda_i(t|\hat{x}_i) = \lambda_0(t)\exp(\hat{x}_i^T\boldsymbol{\beta}) \tag{4}$$

To prove Theorem 1, we need the following three lemmas.

Lemma 1. *(Consistency of the ridge estimator). Let $\boldsymbol{\beta}_{ridge}$ denote the ridge estimator defined in (3) and suppose that the conditions (C1) to (C6) hold. Then we have that*

$$||\boldsymbol{\beta}_{\text{ridge}} - \boldsymbol{\beta}_0|| = O_p(\sqrt{p/n}) \tag{5}$$

Proof. Denote

$$\mathcal{L}(\boldsymbol{\beta}) = l_p(\boldsymbol{\beta}|\hat{u}) - \xi_n\sum_{j=1}^{p}\beta_j^2,$$
$$a_n = \max_{1\leq j\leq q}\{|\dot{p}_{\xi_n}(\beta_{0j})| : \beta_{0j} \neq 0\},$$
$$b_n = \max_{1\leq j\leq q}\{|\ddot{p}_{\xi_n}(\beta_{0j})| : \beta_{0j} \neq 0\}$$

For ridge regression, we can see that $p_{\xi_n}(\beta_{0j}) = \beta_{0j}^2\xi_n/n$ for $j = 1, \ldots, p$. Thus the first and second derivatives of $p_{\xi_n}(\beta_{0j})$ are $\dot{p}_{\xi_n}(\beta_{0j}) = 2\beta_{0j}\xi_n/n$ and $\ddot{p}_{\xi_n}(\beta_{0j}) = 2\xi_n/n$ respectively. From Conditions (C5) and (C6), we have that $a_n \leq 2a_1\xi_n/n = o(n^{-1/2})$ and $b_n \leq 2\xi_n/n = o(n^{-1/2})$. Therefore $a_n \to 0$ and $b_n \to 0$.

Let $\alpha_n = \sqrt{p}(n^{-1/2} + a_n)$, then using the similar manipulation as those in Cai et al.,[31] we can prove that, for any given $\epsilon > 0$, there exists a large constant C_0 such that

$$P\left\{\sup_{||v||=C_0} \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n v) < \mathcal{L}(\boldsymbol{\beta}_0)\right\} \geq 1 - \epsilon$$

which implies that there exists a local maximiser, $\boldsymbol{\beta}_{\text{ridge}}$, such that $||\boldsymbol{\beta}_{ridge} - \boldsymbol{\beta}_0|| = O_p(\sqrt{p/n})$. ■

To describe Lemma 2, for a vector of $\boldsymbol{\theta}$ and given $\boldsymbol{\beta}$, define

$$Q_n(\boldsymbol{\theta}) \equiv Q_n(\boldsymbol{\theta}; \boldsymbol{\beta}, \hat{u}) = l_p(\boldsymbol{\theta}|\hat{u}) - \lambda_n\boldsymbol{\theta}^T D(\boldsymbol{\beta})\,\boldsymbol{\theta}$$

where $D(\boldsymbol{\beta}) = \text{diag}\{0, \beta_1^{-2}, \ldots, \beta_p^{-2}\}$. Then the first and second derivatives of $Q_n(\boldsymbol{\theta})$ are

$$\dot{Q}_n(\boldsymbol{\theta}) = \dot{l}_p(\boldsymbol{\theta}|\hat{u}) - 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\theta} \tag{6}$$

and

$$\ddot{Q}_n(\boldsymbol{\theta}) = \ddot{l}_p(\boldsymbol{\theta}|\hat{u}) - 2\lambda_n D(\boldsymbol{\beta}) \tag{7}$$

Lemma 2. *Suppose* $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta})^T, g_2(\boldsymbol{\beta})^T)^T$ *is a solution to* $\dot{Q}_n(\boldsymbol{\theta}) = 0$ *and let* $\{\delta_n\}$ *be a sequence of positive real numbers satisfying* $\delta_n \to \infty$ *and* $\delta_n^2 p/\lambda_n \to 0$. *Furthermore, define* $\mathcal{H}_n \equiv \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T : |\boldsymbol{\beta}_1| = (|\beta_0|, |\beta_1|, \ldots, |\beta_q|)^T \in [1/K_0, K_0]^{q+1}, ||\boldsymbol{\beta}_2|| \leq \delta_n\sqrt{p/n}\}$, *where* $K_0 > 1$ *is a constant such that* $|\boldsymbol{\beta}_{01}| \in [1/K_0, K_0]^{q+1}$. *Then under the regularity conditions (C1) to (C6) and with probability tending to 1, we have that*

(i) $\sup_{\boldsymbol{\beta}\in H_n} \frac{||g_2(\boldsymbol{\beta})||}{||\boldsymbol{\beta}_2||} < \frac{1}{C_0}$ *for some constant* $C_0 > 1$;

(ii) $g(\cdot)$ *is a mapping from* $\mathcal{H}_n$ *to itself.*

Proof. Taking the first-order Taylor expansion for $\dot{Q}_n(\boldsymbol{\theta})$ at $\boldsymbol{\beta}_0$ in a neighborhood of $g(\boldsymbol{\beta})$, we have that

$$\dot{Q}_n(\boldsymbol{\beta}_0) = \dot{Q}_n(g(\boldsymbol{\beta})) + \ddot{Q}_n(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_0 - g(\boldsymbol{\beta}))$$

where $\boldsymbol{\beta}_0$ is the true parameter vector, and $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $g(\boldsymbol{\beta})$. Then

$$\ddot{Q}_n(\boldsymbol{\beta}^*)g(\boldsymbol{\beta}) = -\dot{Q}_n(\boldsymbol{\beta}_0) + \ddot{Q}_n(\boldsymbol{\beta}^*)\boldsymbol{\beta}_0$$

since $\dot{Q}_n(g(\boldsymbol{\beta})) = 0$. Substituting equations (6) and (7) to the above equation, we have

$$\left[\frac{1}{n}\ddot{l}_p(\boldsymbol{\beta}^*|\hat{u}) - \frac{2\lambda_n}{n}D(\boldsymbol{\beta})\right]g(\boldsymbol{\beta}) = \frac{1}{n}\ddot{l}_p(\boldsymbol{\beta}^*|\hat{u})\boldsymbol{\beta}_0 - \frac{1}{n}\dot{l}_p(\boldsymbol{\beta}_0|\hat{u}) \tag{8}$$

Denote $H_n(\boldsymbol{\beta}^*) = -\frac{1}{n}\ddot{l}_p(\boldsymbol{\beta}^*|\hat{u})$ and from (C3), $H_n(\boldsymbol{\beta}^*)^{-1}$ exists. Then multiplying both sides of (8) by $H_n(\boldsymbol{\beta}^*)^{-1}$

$$g(\boldsymbol{\beta}) - \boldsymbol{\beta}_0 + \frac{2\lambda_n}{n}H_n(\boldsymbol{\beta}^*)^{-1}D(\boldsymbol{\beta})g(\boldsymbol{\beta}) = \frac{1}{n}H_n(\boldsymbol{\beta}^*)^{-1}\dot{l}_p(\boldsymbol{\beta}_0|\hat{u}) \tag{9}$$

Partition $H_n(\boldsymbol{\beta}^*)^{-1}$ and $D(\boldsymbol{\beta})$ into

$$H_n(\boldsymbol{\beta}^*)^{-1} = \begin{pmatrix} A & B \\ B^T & G \end{pmatrix} \text{ and } D(\boldsymbol{\beta}) = \begin{pmatrix} D_1(\boldsymbol{\beta}_1) & 0 \\ 0 & D_2(\boldsymbol{\beta}_2) \end{pmatrix}$$

where A is a $(q+1) \times (q+1)$ matrix, $D_1(\boldsymbol{\beta}_1) = \text{diag}\{0, \beta_1^{-2}, \ldots, \beta_q^{-2}\}$ and $D_2(\boldsymbol{\beta}_2) = \text{diag}\{\beta_{q+1}^{-2}, \ldots, \beta_p^{-2}\}$. Then equation (9) can be rewritten as

$$\begin{pmatrix} g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01} \\ g_2(\boldsymbol{\beta}) \end{pmatrix} + \frac{2\lambda_n}{n} \begin{pmatrix} AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta}) \\ B^T D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta}) \end{pmatrix} = \frac{1}{n} H_n(\boldsymbol{\beta}^*)^{-1} \dot{l}_p(\boldsymbol{\beta}_0|\hat{u}) \tag{10}$$

By arguments similar to those in Theorem 1 of Cai et al.,[31] Conditions (C1) to (C6) guarantee that $||\frac{1}{n} H_n(\boldsymbol{\beta}^*)^{-1} \dot{l}_p(\boldsymbol{\beta}_0|\hat{u})|| = O_p(\sqrt{p/n})$, therefore

$$\sup_{\boldsymbol{\beta} \in H_n} \left\| g_2(\boldsymbol{\beta}) + \frac{2\lambda_n}{n} B^T D_1(\boldsymbol{\beta}_1) g_1(\boldsymbol{\beta}) + \frac{2\lambda_n}{n} GD_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta}) \right\| = O_p(\sqrt{p/n}) \tag{11}$$

Note that $|\boldsymbol{\beta}_1| \in [1/K_0, K_0]^{q+1}$, $||g_1(\boldsymbol{\beta})|| \leq ||g(\boldsymbol{\beta})|| \leq ||\hat{\boldsymbol{\beta}}|| = O_p(\sqrt{p})$, and furthermore, from

$$||BB^T|| - ||A^2|| \leq ||BB^T + A^2|| \leq ||H_n(\boldsymbol{\beta}^*)^{-2}|| < C^2$$

we can derive $||B|| \leq \sqrt{2}C$ and

$$\sup_{\boldsymbol{\beta} \in H_n} \left\| \frac{2\lambda_n}{n} B^T \boldsymbol{D}_1(\boldsymbol{\beta}_1) g_1(\boldsymbol{\beta}) \right\| \leq \frac{2\lambda_n}{n} \sup_{\boldsymbol{\beta} \in H_n} \|B^T\| \|D_1(\boldsymbol{\beta}_1)\| \|g_1(\boldsymbol{\beta})\| = o_p(\sqrt{p/n}) \tag{12}$$

then equation (11) can be rewritten as

$$\sup_{\boldsymbol{\beta} \in H_n} \left\| g_2(\boldsymbol{\beta}) + \frac{2\lambda_n}{n} GD_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta}) \right\| = O_p(\sqrt{p/n}) \tag{13}$$

At the same time

$$\frac{2\lambda_n}{n} \|GD_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta})\| \geq \frac{2\lambda_n}{n} \frac{1}{C} \|D_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta})\| \tag{14}$$

and thus

$$\frac{2\lambda_n}{n} \frac{1}{C} ||D_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta})|| - ||g_2(\boldsymbol{\beta})|| \leq \sup_{\boldsymbol{\beta} \in H_n} \left\| g_2(\boldsymbol{\beta}) + \frac{2\lambda_n}{n} GD_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta}) \right\| \leq \delta_n(\sqrt{p/n}) \tag{15}$$

Let $m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2} = (g_2(\beta_{q+1})/\beta_{q+1}, g_2(\beta_{q+2})/\beta_{q+2}, \ldots, g_2(\beta_p)/\beta_p)^T$, then

$$g_2(\boldsymbol{\beta}) = D_2(\boldsymbol{\beta}_2)^{-1/2} m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}$$

Furthermore, it follows from the Cauchy-Schwarz inequality and the assumption $||\boldsymbol{\beta}_2|| \leq \delta_n \sqrt{p/n}$ that

$$\frac{1}{C} \left\| \frac{2\lambda_n}{n} D_2(\boldsymbol{\beta}_2) g_2(\boldsymbol{\beta}) \right\| \geq \frac{2\lambda_n}{nC} \frac{\sqrt{n}}{\delta_n \sqrt{p}} ||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \tag{16}$$

and

$$||g_2(\boldsymbol{\beta})|| = ||(D_2(\boldsymbol{\beta}_2))^{-1/2} m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \leq ||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \cdot ||\boldsymbol{\beta}_2|| \leq ||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \delta_n \sqrt{p/n} \tag{17}$$

By equations (15), (16) and (17), we have the following inequality

$$\frac{2\lambda_n}{nC} \frac{\sqrt{n}}{\delta_n \sqrt{p}} ||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| - \frac{\delta_n \sqrt{p}}{\sqrt{n}} ||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \leq \frac{\delta_n \sqrt{p}}{\sqrt{n}}$$

Immediately from $p\delta_n^2/\lambda_n \to 0$, we have

$$||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \leq \frac{1}{\frac{2\lambda_n}{p\delta_n^2 C} - 1} < \frac{1}{C_0}, \ \ (C_0 > 1)$$

with probability tending to 1. Hence with probability tending to 1

$$||g_2(\boldsymbol{\beta})|| \leq ||\boldsymbol{\beta}_2||||m_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}|| \leq \frac{1}{C_0}||\boldsymbol{\beta}_2|| \text{ as } n \to \infty$$

which implies that conclusion (i) holds and $||g_2(\boldsymbol{\beta})|| \leq \delta_n\sqrt{p/n}$ with probability tending to 1.

To prove (ii), we only need to verify that $||g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}|| \leq \delta_n\sqrt{p/n}$ with probability tending to 1. Analogously, from equation (10), we have

$$\sup_{\boldsymbol{\beta}\in H_n}\left\|\frac{2\lambda_n}{n}AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})\right\| = o_p(\sqrt{p/n})$$

and

$$\sup_{\boldsymbol{\beta}\in H_n}\left\|g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01} + \frac{2\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\| = O_p(\sqrt{p/n}) \leq \delta_n\sqrt{p/n}$$

Again by equation (15) and Condition (C4), we know that as $n \to \infty$ and with probability tending to 1

$$\sup_{\boldsymbol{\beta}\in H_n}\|\frac{2\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\| \leq C(||g_2(\boldsymbol{\beta})|| + \delta_n\sqrt{p/n})||B|| \leq 2\sqrt{2}C^2\delta_n\sqrt{p/n}$$

Since

$$\|g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}\| - \frac{2\lambda_n}{n}\|BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\| \leq \sup_{\boldsymbol{\beta}\in H_n}\left\|g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01} + \frac{2\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|,$$

then

$$\sup_{\boldsymbol{\beta}\in H_n}\|g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}\| \leq (2\sqrt{2}C^2 + 1)\delta_n\sqrt{p/n} \to 0$$

with probability tending to 1, which implies that for any $\epsilon > 0$, $P(||g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}|| \leq \epsilon) \to 1$. Thus, it follows from $\boldsymbol{\beta}_{01} \in [1/K_0, K_0]^{q+1}$ that $g_1(\boldsymbol{\beta}) \in [1/K_0, K_0]^{q+1}$ holds for large n, which implies that conclusion (ii) holds. This completes the proof. ■

Since $\boldsymbol{\beta}_{02} = 0$, we can express the objective function of this reduced model as

$$Q_{n1}(\boldsymbol{\theta}_1) = l_{p1}(\boldsymbol{\theta}_1) - \lambda_n\boldsymbol{\theta}_1^T D_1(\boldsymbol{\beta}_1)\boldsymbol{\theta}_1 \tag{18}$$

Lemma 3. *Let $f(\boldsymbol{\beta}_1)$ be a solution to $\dot{Q}_{n1}(\boldsymbol{\theta}_1) = 0$, then under regularity conditions (C1) to (C6) and with probability tending to 1*

(i) *$f(\boldsymbol{\beta}_1)$ is a contraction mapping from $[1/K_0, K_0]^{q+1}$ to itself;*

(ii) *$\sqrt{n}(\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_{01}) \xrightarrow{D} N(0, I_1(\boldsymbol{\beta}_0)^{-1})$, where $\hat{\boldsymbol{\beta}}_1^o$ is the unique fixed point of $f(\boldsymbol{\beta}_1)$ and $I_1(\boldsymbol{\beta}_0)$ is the leading $(q+1) \times (q+1)$ submatrix of $I(\boldsymbol{\beta}_0)$.*

Proof. (i) Similar as the derivation of equation (9), through the first-order Taylor expansion, we have that

$$f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01} + \frac{2\lambda_n}{n}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1) = \frac{1}{n}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_{p1}(\boldsymbol{\beta}_{01}) \tag{19}$$

where $H_{n1}(\boldsymbol{\beta}_1^*) = -n^{-1}\ddot{l}_{p1}(\boldsymbol{\beta}_1^*)$ and $\boldsymbol{\beta}_1^*$ lies between $\boldsymbol{\beta}_{01}$ and $f(\boldsymbol{\beta}_1)$. From $n^{-1}\dot{l}_{p1}(\boldsymbol{\beta}_{01}) = O_p(q/n)$, we know that

$$\sup_{|\boldsymbol{\beta}_1|\in[1/K_0,K_0]^{q+1}} \left\| f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01} + \frac{2\lambda_n}{n} H_{n1}(\boldsymbol{\beta}_1^*)^{-1} D_1(\boldsymbol{\beta}_1) f(\boldsymbol{\beta}_1) \right\| = O_p(\sqrt{q/n})$$

By (C4) and similar as the proof process of Lemma 2, we have that

$$\sup_{|\boldsymbol{\beta}_1|\in[1/K_0,K_0]^{q+1}} \left\| \frac{2\lambda_n}{n} H_{n1}(\boldsymbol{\beta}_1^*)^{-1} D_1(\boldsymbol{\beta}_1) f(\boldsymbol{\beta}_1) \right\| = o_p(\sqrt{q/n})$$

Thus, with the probability tending to 1

$$\sup_{|\boldsymbol{\beta}_1|\in[1/K_0,K_0]^{q+1}} ||f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01}|| \le \delta_n\sqrt{q/n} \to 0$$

which implies that $P\{f(\boldsymbol{\beta}_1) \in [1/K_0, K_0]^{q+1}\} \to 1$ as $n \to \infty$. That is, $f(\boldsymbol{\beta}_1)$ is a mapping from $[1/K_0, K_0]^{q+1}$ to itself. Next to prove $f(\boldsymbol{\beta}_1)$ is a contraction mapping, we need show that $\sup_{|\boldsymbol{\beta}_1|\in[1/K_0,K_0]^{q+1}} ||\dot{f}(\boldsymbol{\beta}_1)|| = o_p(1)$.

From $\dot{Q}_{n1}(f(\boldsymbol{\beta}_1)) = 0$ we have

$$\dot{l}_{p1}(f(\boldsymbol{\beta}_1)) = 2\lambda_n D_1(\boldsymbol{\beta}_1) f(\boldsymbol{\beta}_1) \tag{20}$$

Taking the derivative with respect to $\boldsymbol{\beta}_1^T$ on both sides of equation (20) and rearranging terms, we obtain that

$$\left[\frac{2\lambda_n}{n} D_1(\boldsymbol{\beta}_1) + H_{n1}(f(\boldsymbol{\beta}_1))\right] \dot{f}(\boldsymbol{\beta}_1) = \frac{4\lambda_n}{n} f(\boldsymbol{\beta}_1)\mathrm{diag}(0, \boldsymbol{\beta}_1^{-3}, \ldots, \boldsymbol{\beta}_q^{-3}) \tag{21}$$

where $\dot{f}(\boldsymbol{\beta}_1) = \partial f(\boldsymbol{\beta}_1)/\partial \boldsymbol{\beta}_1{}^T$. From the fact that $\lambda_n/\sqrt{n} \to 0$, $||f(\boldsymbol{\beta}_1)||$ and $||\boldsymbol{\beta}_1||$ are bounded, we have

$$\sup_{|\boldsymbol{\beta}_1|\in[1/K_0,K_0]^{q+1}} \frac{4\lambda_n}{n} \|f(\boldsymbol{\beta}_1)\mathrm{diag}(0, \boldsymbol{\beta}_1^{-3}, \ldots, \boldsymbol{\beta}_q^{-3})\| = o_p(1)$$

Again, since

$$1/K_0^2||\dot{f}(\boldsymbol{\beta}_1)|| \le ||H_{n1}(f(\boldsymbol{\beta}_1))\dot{f}(\boldsymbol{\beta}_1)|| \le K_0^2||\dot{f}(\boldsymbol{\beta}_1)||$$

and

$$1/C||\dot{f}(\boldsymbol{\beta}_1)|| \le ||D_1(\boldsymbol{\beta}_1)\dot{f}(\boldsymbol{\beta}_1)|| \le C||\dot{f}(\boldsymbol{\beta}_1)||$$

and from equation (21), we can reach the conclusion that

$$\sup_{|\boldsymbol{\beta}_1|\in[1/K_0,K_0]^{q+1}} ||\dot{f}(\boldsymbol{\beta}_1)|| = o_p(1)$$

which implies that $f(\cdot)$ is a contraction mapping from $[1/K_0, K_0]^{q+1}$ to itself with probability tending to 1. Hence, according to the contraction mapping theorem, there exists one unique fixed-point $\hat{\boldsymbol{\beta}}_1^o \in [1/K_0, K_0]^{q+1}$ such that $f(\hat{\boldsymbol{\beta}}_1^o) = \hat{\boldsymbol{\beta}}_1^o$.

(ii) From equation (19) we have $f(\boldsymbol{\beta}_1) = [H_{n1}(\boldsymbol{\beta}_1^*) + \frac{2\lambda_n}{n} D_1(\boldsymbol{\beta}_1)]^{-1}[H_{n1}(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01} + \frac{1}{n}\dot{l}_{p1}(\boldsymbol{\beta}_{01})]$. Denote $\Phi(\hat{\boldsymbol{\beta}}_1^o) = [H_{n1}(\boldsymbol{\beta}_1^*) + \frac{2\lambda_n}{n} D_1(\hat{\boldsymbol{\beta}}_1^o)]^{-1}$, then we have

$$\hat{\boldsymbol{\beta}}_1^o = f(\hat{\boldsymbol{\beta}}_1^o) = \Phi(\hat{\boldsymbol{\beta}}_1^o) H_{n1}(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01} + \frac{1}{n}\Phi(\hat{\boldsymbol{\beta}}_1^o)\dot{l}_{p1}(\boldsymbol{\beta}_{01})]$$

and

$$\begin{aligned}\sqrt{n}H_{n1}(\boldsymbol{\beta}_1^*)^{1/2}(\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_{01}) &= \sqrt{n}H_{n1}(\boldsymbol{\beta}_1^*)^{1/2}\{\Phi(\hat{\boldsymbol{\beta}}_1^o)H_{n1}(\boldsymbol{\beta}_1^*) - I_{q+1}\}\boldsymbol{\beta}_{01} \\ &\quad + \frac{1}{\sqrt{n}}H_{n1}(\boldsymbol{\beta}_1^*)^{1/2}\Phi(\hat{\boldsymbol{\beta}}_1^o)\dot{l}_{p1}(\boldsymbol{\beta}_{01}) \\ &= \Pi_1 + \Pi_2\end{aligned}$$

with $\Pi_1 = \sqrt{n}H_{n1}(\boldsymbol{\beta}_1^*)^{1/2}\{\Phi(\hat{\boldsymbol{\beta}}_1^o)H_{n1}(\boldsymbol{\beta}_1^*) - I_{q+1}\}\boldsymbol{\beta}_{01}$ and $\Pi_2 = \frac{1}{\sqrt{n}}H_{n1}(\boldsymbol{\beta}_1^*)^{1/2}\Phi(\hat{\boldsymbol{\beta}}_1^o)\dot{l}_{p1}(\boldsymbol{\beta}_{01})$.

Furthermore, it follows from Woodbury matrix identity and Condition (C6) that

$$||\Pi_1|| = \frac{2\lambda_n}{\sqrt{n}}||H_{n1}(\boldsymbol{\beta}_1^*)^{-\frac{1}{2}}D_1(\hat{\boldsymbol{\beta}}_1^o)\Phi(\hat{\boldsymbol{\beta}}_1^o)H_{n1}(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01}|| \le \frac{2\lambda_n}{\sqrt{n}}K_0^2||H_{n1}(\boldsymbol{\beta}_1^*)^{\frac{1}{2}}||||\boldsymbol{\beta}_{01}|| = o_p(1)$$

Similarly using Woodbury matrix identity to Π_2, we have

$$\begin{aligned}||\Pi_2|| &= \frac{1}{\sqrt{n}}H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}\left\{aI_{q+1} - \frac{2\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^o)\Phi(\hat{\boldsymbol{\beta}}_1^o)\right\}\dot{l}_{p1}(\boldsymbol{\beta}_{01}) \\ &= \frac{1}{\sqrt{n}}H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}\dot{l}_{p1}(\boldsymbol{\beta}_{01}) - \frac{2\lambda_n}{\sqrt{n}}H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}D_1(\hat{\boldsymbol{\beta}}_1^o)\Phi(\hat{\boldsymbol{\beta}}_1^o)\frac{1}{n}\dot{l}_{p1}(\boldsymbol{\beta}_{01}) = \frac{1}{\sqrt{n}}H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}\dot{l}_{p1}(\boldsymbol{\beta}_{01}) + o_p(1) \\ &\to N_{q+1}(0, I_{q+1})\end{aligned}$$

Therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_{01}) \to N_{q+1}(0, H_{n1}(\boldsymbol{\beta}_1^*)^{-1})$. ■

Proof of Theorem 1. (i) According to the definitions of the BAR estimator $\hat{\boldsymbol{\beta}}$ and Lemma 1 and Lemma 2(i), we have that

$$\hat{\boldsymbol{\beta}}_2 = \lim_{k\to\infty} g_2(\hat{\boldsymbol{\beta}}^{(k)}) = 0$$

holds with the probability tending to 1.

(ii) Since $\hat{\boldsymbol{\beta}}_1 = \lim_{k\to\infty} g_1(\hat{\boldsymbol{\beta}}^{(k)})$, next we should show that

$$P(\lim_{k\to\infty} ||g_1(\hat{\boldsymbol{\beta}}^{(k)}) - \hat{\boldsymbol{\beta}}_1^o|| = 0) \to 1$$

where $\hat{\boldsymbol{\beta}}_1^o$ is the unique fixed point of $f(\boldsymbol{\beta}_1)$ defined in Lemma 3.

According to the definition of $g(\boldsymbol{\beta})$, $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta})^T, g_2(\boldsymbol{\beta})^T)^T$ is the solution of $\dot{Q}_n(\boldsymbol{\theta}) = 0$, that is, $g_1(\boldsymbol{\beta})$ is the solution of $\dot{Q}_{n1}(\boldsymbol{\theta}_1) = 0$ and $g_2(\boldsymbol{\beta})$ is the solution of $\dot{Q}_{n2}(\boldsymbol{\theta}_2) = 0$.

From (i) we can see that $\lim_{\boldsymbol{\beta}_2\to 0} g_2(\boldsymbol{\beta}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = 0$, and thus $\lim_{\boldsymbol{\beta}_2\to 0} g_1(\boldsymbol{\beta}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = f(\boldsymbol{\beta}_1)$ holds. Also, for any $\hat{\boldsymbol{\beta}}_2^{(k)}$, $g(\boldsymbol{\beta}; \boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_2^{(k)})$ is a mapping of $\boldsymbol{\beta}_1$, and with $k \to \infty$ and probability tending to 1, we have that

$$\eta_k \equiv \sup_{g_1(\boldsymbol{\beta})\in[1/K_0, K_0]^{q+1}} \|f(\boldsymbol{\beta}_1) - g_1(\boldsymbol{\beta}; \boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_2^{(k)})\| \to 0 \tag{22}$$

On the other hand, since $f(\cdot)$ is a contraction mapping, there exists a constant $C_1 > 1$ such that

$$||f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\boldsymbol{\beta}}_1^o|| = ||f(\hat{\boldsymbol{\beta}}_1^{(k)}) - f(\hat{\boldsymbol{\beta}}_1^o)|| \le \frac{1}{C_1}||\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^o|| \tag{23}$$

Let $h_k = ||\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^o||$, then it follows from equations (22) and (23) that

$$h_{k+1} = ||\hat{\boldsymbol{\beta}}_1^{(k+1)} - \hat{\boldsymbol{\beta}}_1^o|| \quad \leq ||g_1(\hat{\boldsymbol{\beta}}^{(k)}) - f(\hat{\boldsymbol{\beta}}_1^{(k)})|| + ||f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\boldsymbol{\beta}}_1^o|| \leq \eta_k + \frac{1}{C_1} h_k$$

From equation (22), for any $\epsilon \geq 0$, there exists $N > 0$ such that when $k > N$, $0 \leq \eta_k < \epsilon$.
Employing some recursive calculation, we have $h_k \to 0$ as $k \to \infty$. Hence, with probability tending to 1, we have

$$||\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^o|| \to 0 \,\text{as}\, k \to \infty$$

Since $\hat{\boldsymbol{\beta}}_1 \equiv \lim_{k\to\infty} \hat{\boldsymbol{\beta}}_1^{(k)}$, it follows from the uniqueness of the fixed-point that

$$P(\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^o) \to 1, \qquad k \to \infty$$

(iii) The asymptotic normality of $\hat{\boldsymbol{\beta}}_1$ follows from part (ii) of Lemma 3. ■