VARIABLE SELECTION FOR CASE-COHORT STUDIES WITH FAILURE TIME OUTCOME

Ai Ni

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Jianwen Cai

Jeannette Bensen

Amy Herring

Wei Sun

Donglin Zeng

ABSTRACT

Ai Ni: Variable Selection For Case-Cohort Studies With Failure Time Outcome
(Under the direction of Jianwen Cai)

Case-cohort design is widely used in large cohort studies with failure time data to reduce the cost associated with covariate measurement. Many of those studies collect a large number of covariates. Therefore, an efficient variable selection method is needed for the case-cohort design. In this dissertation, we study the properties of the Smoothly Clipped Absolute Deviation (SCAD) penalty based variable selection procedure in Cox proportional hazards model and additive hazards model in a case-cohort design with a diverging number of parameters.

We prove that the SCAD penalized variable selection procedure can identify the true model with probability tending to one as $n \to \infty$ under Cox proportional hazards model. We then establish the consistency and asymptotic normality of the penalized estimator. We show via simulation that the BIC-based tuning parameter selection method outperforms the AIC-based method under typical case-cohort study settings. The proposed procedure is applied to the Busselton Health Study (Cullen 1972, Knuiman et al. 2003).

Additive hazards model is a useful alternative to the Cox model for analyzing failure time data. In the second part of the dissertation, we extend the SCAD-penalized variable selection procedure to the additive hazards model with a stratified case-cohort design and a diverging number of parameters. We again establish variable selection consistency, estimation consistency, and asymptotic normality of the penalized estimator under this setting. We propose a new tuning parameter selection method and evaluate its performance via simulation. We show that the proposed tuning parameter selection method outperforms

the conventional k-fold cross-validation method. The proposed procedure is applied to the Atherosclerosis Risk in Communities (ARIC) study (Ballantyne et al. 2004).

Tuning parameter selection is critical to the success of a regularized variable selection method. A consistent tuning parameter selection method has not been established for the SCAD-penalized Cox model with a diverging dimension. In the last part of the dissertation, we propose a generalized information criterion (GIC) for tuning parameter selection and establish conditions required for its variable selection consistency under this setting. Simulation study shows that GIC performs well under the required conditions with finite sample size. It is then applied to the Framingham Heart Study (Dawber 1980).

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my dissertation advisor, Dr. Jianwen Cai, for her guidance, advices, patience, and encouragement during my PhD study. I have learned tremendously under her direction. My thanks are also due to my committee members: Drs. Jeannette Bensen, Amy Herring, Wei Sun, and Donglin Zeng for their valuable comments and suggestions. I am also thankful for my fellow students from whom I gained courage and ideas. Last but not least, my special thanks go to my wife, Xiaokun Qian, who always gave me unconditional support for my endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Modern epidemiological cohort studies and disease prevention trials often need to follow thousands of subjects for many years. There are two typical features of large-scale cohort studies and prevention trials. First, the investigators are usually interested in the association of a large number of risk factors with an outcome. However, the assembly of some covariates may require the analysis of previously stored precious biological samples such as serum and genetic materials using expensive bioassays, genotyping, or sequencing technology. Therefore, it can be prohibitively expensive to collect all covariates from every subject in the study. Second, the rate of occurrence of event of interest is usually low, especially for such events as cancer or death. Consequently, subjects without the event of interest (noncases) constitute a predominant portion of the cohort, and if the covariates were to be measured for every subject, most of the associated cost would be spent on the noncases, which do not contribute as much information as subjects with the event of interest (cases) in the analysis of failure time data. To reduce the cost and effort in collecting expensive covariates without decreasing much efficiency in the analysis of failure time data, Prentice (1986) proposed the case-cohort design, where the complete covariate information is only obtained from a random subcohort sample plus all cases. Case-cohort design has been widely used in practice. For example, in the Busselton Health Study (Cullen 1972, Knuiman et al. 2003) a cohort of 1,401 Australian from Busselton in West Australia was followed for 15 years, and the time to stroke was analyzed under case-cohort design where the main risk factor serum ferritin level was only measured for the case-cohort of size 513.

In case-cohort studies where a large number of covariates are collected, researchers are

often interested in selecting a subset of the covariates that are related to the event of interest. With the inclusion of interaction terms and polynomial terms, the number of candidate covariates can be very large. In the aforementioned Busselton Health Study, there are a number of potential confounders or effect modifiers that need to be considered in the modeling process. With the pairwise interactions between ferritin level and all the other covariates as well as the squared continuous covariates, the total number of terms in the model exceeds 30, which is fairly high considering that there are only 118 incidence of stroke in the cohort. As Huber (1973) argued, in the context of variable selection the number of parameters should be considered as increasing with sample size, and goes to infinity as sample size goes to infinity. Therefore, an efficient variable selection procedure that allows a diverging number of parameters is needed for the case-cohort design. Although we consider the dimension of the parameter to increase with sample size, we restrict ourselves to the $p \ll n$ scenario in this dissertation. The traditional variable selection methods such as stepwise and best subset selection suffer from two major drawbacks. First, they are unstable in that covariates are either retained or dropped from the model, and therefore small changes in the data can result in very different models being selected. Second, they are computationally intensive, and becomes infeasible when the number of covariates increases with sample size. To overcome these drawbacks, penalized likelihood based variable selection procedures have been developed over the last few decades. Under certain regularity conditions, these procedures can automatically and simultaneously select variables and estimate their coefficients. The penalty-based variable selection procedures have been successfully applied to linear, generalized linear, Cox proportional hazards, and additive hazards model. However, to our knowledge, the properties of these procedures have not been studied under proportional hazards or additive hazards model with case-cohort design and a diverging number of parameters. This dissertation intends to fill in this gap.

The properties of the regularized variable selection procedures depend on the penalty function that is applied to the likelihood function. Many penalty functions have been proposed in the literature. Among them, the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) has been shown to possess the so-called oracle property, namely, as sample size goes to infinity, the procedure correctly identifies the true model with probability one and estimates the standard errors of nonzero parameters as efficiently as if the zero parameters were never included in the estimation process. For the first topic of the dissertation, we investigate both the asymptotic and finite sample properties of the SCAD estimator under Cox proportional hazards model with a case-cohort design and a diverging number of parameters. We first establish the rate of convergence of the maximum SCAD-penalized pseudo-partial likelihood estimator. We then prove its oracle property and establish its asymptotic distribution. As mentioned before, the rate of event of interest is often very low in case-cohort studies (typically over 90% censoring rate). However, most previous studies on regularized variable selection in survival analysis investigated its finite sample properties with fairly low censoring percentage. The performance of the method in high censoring percentage situation is largely unknown. We conduct extensive simulation studies to assess its finite sample properties under a case-cohort design with high censoring percentages.

Although Cox proportional hazards model has gained tremendous popularity in the analysis of time-to-event data, its proportional hazards assumption may fail to hold in many situations. The additive hazards model was developed as a useful alternative to the proportional hazards model. It does not require the assumption that the covariate effect on the hazard function is proportional. It has sound biological and empirical basis. The additive covariate effect on the hazard function is easier to interpret and communicate with investigators. In fact, investigators are sometimes more interested in the risk difference attributed to the covariates. The risk difference is more relevant to public health because it

translates directly into the number of disease cases that would be avoided by eliminating a particular exposure (Kulich and Lin 2000). Over the years, estimators for additive hazards model with full cohort and case-cohort have been proposed and their asymptotic properties studied (Lin and Ying 1994, Kulich and Lin 2000). Variable selection procedures under additive hazards model have also been extensively studied with Lasso (Leng and Ma 2007), adaptive Lasso (Martinussen and Scheike 2009), and SCAD penalty (Lin and Lv 2013). However, to our knowledge, variable selection under additive hazards model with case-cohort design has not been studied. As the second topic of the dissertation, we theoretically and empirically investigate the properties of SCAD-penalized variable selection procedure in additive hazards model with a stratified case-cohort design and a diverging number of parameters. We also propose an effective tuning parameter selection method for the SCAD-based variable selection procedure in additive hazards model under case-cohort design.

All regularized variable selection procedures involve one or several tuning parameters that control the complexity of the selected model by adjusting the magnitude of the penalty. The optimal performance of these variable selection procedure are heavily dependent on the selection of the tuning parameters. There are mainly two data-driven tuning parameter selection methods: K-fold cross-validation (Efron and Tibshirani 1993) and generalized cross-validation (GCV) (Craven and Wahba 1979). The latter is more computationally efficient and is analogous to the Akaike information criteria (AIC) whose properties are thoroughly studied in the traditional variable selection literature. It has been shown that the original GCV is selection inconsistent. That is, the tuning parameter selected from GCV identifies a model different from the true one with probability tending to one as sample size goes to infinity. A number of authors developed various modified tuning parameter selection method that is selection consistent. However, their work lies in the framework of linear and generalized linear model. The tuning parameter selection method has not been theoretically studied for Cox proportional hazards model with a diverging number of

parameters. The third topic of the dissertation is devoted to developing a variable selection consistent tuning parameter selection method. We provide theoretical justification and empirical evidence via simulation that the proposed tuning parameter selection method leads to the correct tuning parameter that identifies the true model with probability tending to one under Cox proportional hazards model with a diverging number of parameters.

The overall goal of this dissertation is to provide theoretical foundation as well as practical guidance for regularized variable selection in a case-cohort design, and thereby facilitates large-scale epidemiological studies on public health issues.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we review the literature on the following topics: 1) estimation method for Cox proportional hazards model under a case-cohort design; 2) regularized variable selection procedures for Cox proportional hazards model; 3) estimation method for additive hazards model under a case-cohort design; 4) regularized variable selection procedures for additive hazards model; 5) tuning parameter selection for regularized variable selection procedures.

2.1 Estimation Method for Cox Proportional Hazards Model under a Case-Cohort Design

The Cox proportional hazards model (Cox 1972) has been the most widely used model to study the effect of covariates on failure times. Under Cox model, the hazard function for the failure time T given time-dependent covariate vector $Z(\cdot)$ is given by

$$\lambda\{t|Z(t)\} = \lambda_0(t)\exp\{\beta_0^T Z(t)\},$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β is a vector of regression coefficients.

Let C be the censoring time and $X = \min(T, C)$ be the observed time and $\Delta = I(T \leq C)$ be the failure indicator, where $I(\cdot)$ is an indicator function. T and C are assumed to be independent conditional on Z. Define the counting process $N(t) = I(X \leq t, \Delta = 1)$, and the at risk process $Y(t) = I(X \geq t)$. The partial likelihood function introduced by Cox

(1972) is given by

$$\ell_n(\beta) = \sum_{i=1}^{n} \left[\beta' Z_i(t) - \log \sum_{j=1}^{n} Y_j(t) \exp\{\beta^T Z_j(t)\} \right] \Delta_i.$$

The maximum partial likelihood estimator of β_0 can be obtained by solving the score equation

$$U_n(\beta) = \sum_{i=1}^{n} \left\{ Z_i(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} \Delta_i = 0,$$

where $S^{(0)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) \exp\{\beta^T Z_i(t)\}$ and
$S^{(1)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i(t) \exp\{\beta^T Z_i(t)\}$. If the longest follow-up time is τ, then the score equation can be equivalently written in the counting process format as

$$U_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dN_i(t) = 0.$$

The covariance matrix of the above estimator $\hat{\beta}$ can be consistently estimated by the inverse of the observed information matrix $\hat{\Sigma}^{-1} = -\{\partial U_n(\beta)/\partial\beta|_{\beta=\hat{\beta}}\}^{-1}$ (Andersen and Gill 1982).

In the case-cohort design, the covariate information is available only for a random subcohort plus all cases. As a result, the risk set at each failure time needs to be modified so that only subjects with available covariate information are used. Prentice (1986) introduced a pseudolikelihood to estimate the regression coefficients,

$$\tilde{\ell}_n(\beta) = \sum_{i=1}^{n} \left[\beta' Z_i(t) - \log \sum_{j \in \tilde{R}(t)} Y_j(t) \exp\{\beta^T Z_j(t)\} \right] \Delta_i,$$

where $\tilde{R}(t) = D(t) \cup C$, $D(t) = \{i : N_i(t) \neq N_i(t^-)\}$, and C is the random subcohort. In words, the risk set at each failure time t includes all subcohort members at risk at t and any subjects outside the subcohort that fail at t. The author provided a heuristic

estimation procedure for the pseudolikelihood. Self and Prentice (1988) slightly modified the risk set by setting $\tilde{R}(t) = C$. That is, only subcohort members at risk are included in the risk set of each failure time. While the estimator of Prentice is score-unbiased, that of Self and Prentice is not. Nevertheless, the latter is asymptotically equivalent to the former provided an individual's contributions to $S^{(1)}$ and $S^{(0)}$ are asymptotically negligible. Under mild regularity conditions, the authors used a combination of martingale and finite population convergence results to prove that the maximum pseudolikelihood estimator has an asymptotic normal distribution with mean β_0 and covariance matrix of the form $n^{-1}\Sigma^{-1}(\Sigma + \Delta)\Sigma^{-1}$. The matrix Σ can be consistently estimated by the observed information matrix. The matrix Δ takes on a very complicated expression and reflects the extra variance induced by the sampling of the subcohort. To circumvent direct estimation of Δ, Wacholder et al. (1989) developed a bootstrap estimate of the variance of the maximum pseudolikelihood estimator. Their method imitates the original sampling scheme by resampling separately cases and subcohort controls. However, it is very computationally intensive. Barlow (1994) and Lin and Ying (1993) proposed different variance estimators that are easily computed.

Barlow (1994) proposed a robust estimator of the variance based on the influence of an individual observation on the overall score function. The author also proposed a slightly different pseudolikelihood function than those of Prentice (1986) and Self and Prentice (1988). In the modified pseudolikelihood function, the author introduced a time-dependent weight for individual i given by $w_i(t) = dN_i(t) + \{1 - dN_i(t)\}\xi_i m(t)/\tilde{m}(t)$, where $\xi_i = 1$ if individual i belongs to the subcohort and 0 otherwise, $m(t)$ is the number of individuals in the full cohort at risk at time t, and $\tilde{m}(t)$ is the number of individuals in the subcohort at risk at time t. This weight is different from that in Prentice (1986) in that individuals with $dN_i(t) = 0$ and $\xi_i = 1$ receives a weight of $m(t)/\tilde{m}(t)$ instead of 1. The log-pseudolikelihood

function is then given by

$$\tilde{\ell}_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta^T Z_i(t) - \log \sum_{j=1}^{n} Y_j(t) w_j(t) \exp\{\beta^T Z_i(t)\} \right] dN_i(t) = 0. \qquad (2.1)$$

Lin and Ying (1993) developed a general solution to the problem of missing covariates under the Cox proportional hazards model. It approximates the partial likelihood score function with full covariate measurements and includes case-cohort design as a special case. Let the p–dimensional covariate vector (possibly time-dependent) for individual i be $Z_i(\cdot) = \{Z_{1i}(\cdot), ..., Z_{pi}(\cdot)\}^T$. Let $H_{0i}(t)$ be an indicator function that equals 1 if $Z_i(t)$ is completely observed and 0 otherwise. Let $H_i(\cdot)$ be a $p \times p$ diagonal matrix with indicator functions $\{H_{1i}(\cdot), ..., H_{pi}(\cdot)\}$ as the diagonal elements, where $H_{ji}(t) = 1$ if $Z_{ji}(t)$ is available and 0 otherwise. The authors proposed the following approximate partial-likelihood score function for estimation of β_0

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \Delta_i H_i(X_i) \{Z_i(X_i) - \mathrm{E}(\beta, X_i)\},$$

where X_i is the observed time for individual i, $\mathrm{E}(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$, and $S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} H_{0i}(t) Y_i(t) \exp\{\beta^T Z_i(t)\} Z_i(t)^{\otimes d}$, $d = 0, 1$. Let $\hat{\beta}$ be the root of the above score function. Under certain regularity conditions, the authors showed that $n^{1/2}(\hat{\beta} - \beta_0)$ converges to a zero-mean normal distribution. Under case-cohort design, the covariance matrix of the limiting distribution is much easier to estimate than those in Prentice (1986) and Self and Prentice (1988).

If the complete covariate history is available for the cases outside the subcohort, then a more efficient pseudolikelihood function can be constructed as proposed by Kalbfleisch and Lawless (1988). Their original pseudolikelihood function (13) can be equivalently expressed as (2.1) with the weight function $w_i = \Delta_i + (1 - \Delta_i)\xi_i/\alpha$, where $\alpha = \tilde{n}/n$ is the sampling probability of the subcohort and Δ_i and ξ_i are the same as defined before. With

this weight function, cases outside the subcohort are always included in the risk set for all failure times rather than only the ones at which they fail. Borgan et al. (2000) considered a time-varying version of this weight in their Estimator II in which the true sampling probability α is replaced with its sample estimate $\hat{\alpha}(t) = \sum_{i=1}^{n} \xi_i(1-\Delta_i)Y_i(t) / \sum_{i=1}^{n}(1-\Delta_i)Y_i(t)$. Using an estimated rather than the known true sampling probability can actually improve efficiency (Robins et al. 1994). Kulich and Lin (2004) rigorously proved the asymptotic properties of the estimator based on this efficient time-varying weight function and generalized it to doubly weighted estimator by replacing the scalar $\hat{\alpha}(t)$ with a matrix $\hat{\alpha}(t) = \{\sum_{i=1}^{n}(1-\Delta_i)A_i(t)\}^{-1}\{\sum_{i=1}^{n}\xi_i(1-\Delta_i)A_i(t)\}$, where $A_i(t)$ is a $p \times p$ diagonal matrix with p potentially different random processes on the diagonal to capture the covariate information that is available for all cohort members as well as surrogate measurements of the expensive covariates. Kang and Cai (2009) extended the efficiently weighted estimator of Borgan et al. (2000) to studies with multiple outcomes of interest. The authors used a marginal model to handle the correlation among multiple outcomes and derived a sandwich estimator of the covariance matrix of the estimated parameters. Kim et al. (2013) further improved the efficiency of the estimators for case-cohort studies with multiple outcomes by replacing the weight function $w_i(t)$ in (2.1) with a modified one that uses the covariate information from cases of all types. Let K be the number of outcome types. $\Delta_{ij} = 1$ if individual i has the outcome j and 0 otherwise. The modified weight function for outcome type k is given by

$$\psi_{ik}(t) = \left\{1 - \prod_{j=1}^{K}(1-\Delta_{ij})\right\} + \prod_{j=1}^{K}(1-\Delta_{ij})\xi_i\tilde{\alpha}_k^{-1}(t),$$

where $\tilde{\alpha}_k(t) = \sum_{i=1}^{n}\xi_i\{\prod_{j=1}^{K}(1-\Delta_{ij})\}Y_{ik}(t) / \sum_{i=1}^{n}\{\prod_{j=1}^{K}(1-\Delta_{ij})\}Y_{ik}(t)$. In words, this weight function makes use of the complete covariate history of cases of all other types that are outside the subcohort when constructing the pseudolikelihood function for a specific outcome type.

In this dissertation, we use the time-varying efficient weight considered in the Estimator II in Borgan et al. (2000) and Kulich and Lin (2004) in a univariate case-cohort design.

2.2 Regularized Variable Selection Procedures for Cox Proportional Hazards Model

Variable selection is an important component of statistical modeling. The idea of penalization has long been used in the modeling process to achieve the balance between goodness-of-fit and model complexity. Among others, Akaike's information criterion (AIC) (Akaike 1973), Mallows' C_p (Mallows 1973), and Bayesian information criterion (BIC) (Schwarz 1978) are probably the most commonly used traditional penalty-based variable selection criteria. These criteria, however, rely on stepwise or subset selection procedures and are separated from the parameter estimation procedure. As a result, they are computationally intensive and unstable (Breiman 1996), and their sampling properties are hard to derive. Tibshirani (1996) proposed a seminal method for variable selection in linear models based on penalized sum of squares. The author named the procedure least absolute shrinkage and selection operator or Lasso. Let $X_i = (X_{i1}, ..., X_{ip})^T$ be the p-dimensional covariate vector for individual i ($i = 1, ..., n$), y_i be the response variable for individual i, and $\beta_0 = (\beta_{01}, ..., \beta_{0p})^T$ be p-dimensional regression coefficients. In its original form, the Lasso estimate $\hat{\beta}$ is defined by

$$\hat{\beta} = \text{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{2.2}$$

where $t \geq 0$ is a tuning parameter that controls model complexity. The Lasso estimator can be more generally expressed as the maximizer of the L_1 penalized log-likelihood function

$$\hat{\beta} = \text{argmax} \left\{ \ell_n(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where λ is a tuning parameter that has a one-to-one relationship with t in (2.2). Tibshirani (1996) showed that the Lasso procedure shrinks all parameter estimates towards 0 and sets some estimates to exactly 0, thus achieves model selection purpose. There are a number of algorithms proposed in the literature to compute the Lasso estimator. Tibshirani (1996) used an iterative reweighted least squares (IRLS) method. Fu (1998) developed a "shooting algorithm" in the linear model framework. Efron et al. (2004) proposed an elegant and powerful variable selection algorithm named least angle regression or LARS that computes Lasso estimator as a special case. Moreover, LARS can compute the entire solution path as a function of the tuning parameter. Tibshirani (1997) extended the Lasso variable selection method to the Cox proportional hazards model, where the Lasso estimator is the maximizer of the L_1 penalized log-partial likelihood function. Park and Hastie (2007) introduced a L_1 penalty solution path algorithm for generalized linear and Cox proportional hazards model.

Fan and Li (2001) proposed a new penalty function under linear and generalized linear models, which they named Smoothly Clipped Absolute Deviation Penalty or SCAD. The SCAD estimator is the maximizer of the following penalized likelihood function

$$Q_n(\beta) = \ell_n(\beta) - n\sum_{j=1}^{p} P_\lambda(|\beta_j|),$$

where the first derivative of the penalty function satisfies

$$P'_\lambda(\theta) = \lambda I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{a-1} I(\theta > \lambda)$$

for some $a > 2$, $\lambda > 0$, and $\theta > 0$, with $P_\lambda(0) = 0$.

The SCAD penalty is different from the Lasso penalty in that it does not over penalize large β's. The authors showed that, under some regularity conditions, SCAD estimator

correctly shrinks zero-valued parameters to 0, and consistently estimates the non-zero parameters. Moreover, it estimates the non-zero parameters as efficient as if the underlying true model is known *a priori*, a property often called *oracle property* in the literature. As pointed out by the authors, the Lasso estimator does not possess oracle property because it underestimates the non-zero parameters due to its over-penalization on large parameters. Fan and Li (2001) also proposed a new unified algorithm to compute the estimates from penalty functions that are singular at the origin, which include Lasso and SCAD. In this *local quadratic approximations* or LQA algorithm, the penalty function is locally approximated by a quadratic function as follows. Suppose an initial value $\beta^{(0)}$ is obtained. If $\beta_j^{(0)}$ is very close to 0 by a pre-specified threshold value, then it is set to 0. Otherwise the penalty function for β_j is approximated by

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\left\{P'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\right\}\left\{\beta_j^2 - (\beta_j^{(0)})^2\right\} \qquad \text{for } \beta_j \approx \beta_j^{(0)}.$$

With the approximated penalty function, the minimization problem becomes a quadratic minimization problem and standard Newton-Raphson algorithm can be used to solve for the minimizer, which is used as the new initial value $\beta^{(0)}$. These steps are iterated until convergence. In practice, the authors suggested using the unpenalized maximum likelihood estimator as the initial value $\beta^{(0)}$. It should be noted that the SCAD penalty is not convex on $(-\infty, \infty)$, and therefore the SCAD penalized likelihood function is not concave. As a result, the SCAD estimator obtained by the above algorithm cannot be guaranteed to be the global maximizer. In practice it is suggested that different initial values be used to increase the probability of obtaining the global maximizer. Fan and Li (2002) extended the SCAD estimator to Cox proportional hazards model and proved its oracle property.

Several other penalty functions have been proposed and their properties studied in Cox proportional hazards model. Zou (2006) proposed an adaptive Lasso method for variable selection where the L_1 penalty for β_j is multiplied by a weight defined by $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$,

where $\hat{\beta}_j$ is a root-n-consistent estimator of the true parameter β_0 and γ is a positive constant that is chosen by the analyst. Under certain regularity conditions, the author established the oracle property of the adaptive Lasso estimator. Zhang and Lu (2007) extended the adaptive Lasso estimator to the Cox proportional hazards model and proved its oracle property. Zou and Hastie (2005) proposed a new penalty function that is a linear combination of L_2 and L_1 penalties. The authors named their penalty *elastic net*. The elastic net penalty successfully addresses the $p \gg n$ scenario and high correlation among groups of covariates. Wu (2012) recently extended the elastic net method to Cox model and developed a path algorithm for it.

As mentioned in the introduction, in many real data applications the number of covariates should be modeled as diverging with sample size. On this frontier, Peng and Fan (2004) provided a rather complete theoretical framework for the asymptotic properties of nonconcave penalized likelihood under generalized linear model with a diverging number of parameters. Cai et al. (2005) investigated the SCAD penalty in Cox proportional hazards model with correlated outcomes and a diverging number of parameters. The authors proved the oracle property of the variable selection procedure and derived the asymptotic distribution of the parameter estimates. Zou and Zhang (2009) proposed an adaptive elastic net penalty which is a modified elastic net penalty with the L_1 penalty component replaced by a weighted L_1 penalty as in the adaptive Lasso. The authors established the oracle property of the procedure with a diverging number of parameters and showed by simulations that the proposed method dealt with collinearity problem better than other oracle-possessing variable selection methods.

2.3 Estimation Method for Additive Hazards Model under a Case-Cohort Design

Additive hazards model is an important alternative to the Cox proportional hazards model. It models risk difference, which bares more intuitive interpretation than risk ratio in many epidemiological and biological studies (Huffer and McKeague 1991). Additive hazards model was originally proposed by Aalen (1980). The hazard function under the additive hazards model for the failure time T given time-dependent covariate vector $Z(\cdot)$ is given by

$$\lambda(t|Z(t)) = \lambda_0(t) + \beta_0^T Z(t), \tag{2.3}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β is a vector of regression coefficients. Lin and Ying (1994) proposed an estimator for model (2.3) and derived its asymptotic properties. The authors proposed the following score equation under counting process framework

$$U(\beta) = \sum_{i=1}^{n} \int_0^\tau \left\{ Z_i(t) - \bar{Z}(t) \right\} \left\{ dN_i(t) - Y_i(t)\beta^T Z_i(t)dt \right\},$$

where $\bar{Z}(t) = \sum_{j=1}^n Y_j(t)Z_j(t) / \sum_{j=1}^n Y_j(t)$. The estimator $\hat{\beta}$ is obtained by solving $U(\beta) = 0$, which has a closed form

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^\tau Y_i(t) \left\{ Z_i(t) - \bar{Z}(t) \right\}^{\otimes 2} dt \right]^{-1} \left[\sum_{i=1}^{n} \int_0^\tau \left\{ Z_i(t) - \bar{Z}(t) \right\} dN_i(t) \right]. \tag{2.4}$$

Under some regularity conditions, $n^{1/2}(\hat{\beta} - \beta_0)$ has been shown to converge in distribution to a p-dimensional normal distribution with mean 0 and covariance matrix that can

be consistently estimated by a sandwich type estimator $A^{-1}BA^{-1}$, where

$$A = n^{-1}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}dt, \qquad B = n^{-1}\sum_{i=1}^{n}\int_0^\tau \left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}dN_i(t).$$

Kulich and Lin (2000) extended the additive hazards model to case-cohort studies. Sharing the same spirit of Kalbfleisch and Lawless (1988), the authors proposed a weighted pseudo-score function

$$U_H(\beta) = \sum_{i=1}^{n}\rho_i\int_0^\tau \left\{Z_i(t)-\bar{Z}_H(t)\right\}\left\{dN_i(t)-Y_i(t)\beta^T Z_i(t)dt\right\},$$

where $\bar{Z}_H(t) = \sum_{j=1}^n \rho_j Y_j(t)Z_j(t)/\sum_{j=1}^n \rho_j Y_j(t)$, $\rho_i = \Delta_i + (1-\Delta_i)\xi_i/p_i$ and $p_i = \Pr(\xi_i = 1)$. The estimator $\hat{\beta}$ solves $U_H(\beta)$ and takes a similar closed form as (2.4). Under some regularity conditions, the authors showed that $n^{1/2}(\hat{\beta}-\beta_0)$ converges to a p-dimensional normal distribution with mean 0 and covariance matrix $D_A^{-1}(\Sigma_A+\Sigma_H)D_A^{-1}$, where

$$D_A = \mathrm{E}\left[\int_0^\tau \{Z_1(t)-e(t)\}^{\otimes 2}Y_1(t)dt\right], \qquad \Sigma_A = \mathrm{E}\left[\int_0^\tau \{Z_1(t)-e(t)\}^{\otimes 2}dN_1(t)\right],$$

$$\Sigma_H(\beta_0) = \mathrm{E}\left\{\frac{(1-p_1)(1-\Delta_1)S_1^{\otimes 2}(\beta_0)}{p_1}\right\},$$

where $e(t) = \mathrm{E}\{Z_1(t)Y_1(t)\}/\mathrm{E}\{Y_1(t)\}$, $S_i(\beta_0) = \int_0^\tau\{Z_i(t)-e(t)\}dM_i(t)$, and $M_i(t) = N_i(t) - \int_0^\tau Y_i(s)d\Lambda_0(s) - \int_0^\tau \beta_0^T Z_i(s)Y_i(s)ds$.

2.4 Regularized Variable Selection Procedures for Additive Hazards Model

Many researchers have applied the penalty-based variable selection procedures to the additive hazards model to achieve a sparse model from a large number of candidate covariates. Ma and Huang (2005) proposed a Lasso type estimator to select important genes

under additive hazards model. The authors applied an L_1 constraint that $\sum_{s=1}^{d} |\beta_s| \leq u$ (d is the number of the covariates; u is a tuning parameter) to the loss function

$$M(\beta) = \sum_{s=1}^{d} \left\{ \left(\sum_{i=1}^{n} L_{s,1}^i \right) \beta_1 + ... + \left(\sum_{i=1}^{n} L_{s,d}^i \right) \beta_d - \sum_{i=1}^{n} R_s^i \right\}^2,$$

where $L_{s,l}^i$ is the (s,l) component of matrix $L^i = \int_0^\infty Y_i(t) \left\{ Z_i - \bar{Z}(t) \right\}^{\otimes 2} dt$ and R_s^i is the s^{th} component of $R^i = \int_0^\infty \left\{ Z_i - \bar{Z}(t) \right\} dN_i(t)$. The Lasso type estimator is the minimizer of the above loss function under the L_1 constraint. The authors proposed using weighted bootstrap technique to compute the covariance matrix of the Lasso type estimator. This estimator shares the same drawback of regular Lasso estimator that it is not path consistent. In other words, there is a positive probability that the solution path of this procedure does not contain the true model. Leng and Ma (2007) proposed a weighted Lasso estimator under additive hazards model which is the maximizer of the following objective function

$$\frac{1}{2}(\beta^T A_n \beta - 2\beta^T b_n) + n\lambda_n \sum_{j=1}^{p} \omega_j |\beta_j|,$$

where $A_n = \sum_{i=1}^{n} \int_0^\infty Y_i(t) \left\{ Z_i - \bar{Z}(t) \right\}^{\otimes 2} dt$, $b_n = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \bar{Z}(t) \right\} dN_i(t)$, and ω_j is a non-negative weight whose inverse is a consistent estimator of β_j. The authors showed that the weighted Lasso estimator is path consistent and possesses the oracle property. Martinussen and Scheike (2009) independently proposed the same weighted Lasso estimator. They formally justified the choice of the loss function $L(\beta) = \beta^T A_n \beta - 2\beta^T b_n$ used in the variable selection procedure. Unlike the Cox proportional hazards model where the log-partial likelihood function is a natural choice of loss function for variable selection, under additive hazards model the likelihood function is difficult to work with due to the nonparametric baseline function and the additive structure of the model. Motivated by the similarity between the Lin-Ying estimator (Lin and Ying 1994) and the least square estimator, Martinussen and Scheike (2009) argued that the above loss function $L(\beta)$ should

be used in variable selection for additive hazards model. In fact, $L(\beta)$ can be obtained by integrating the Lin-Ying score function $U(\beta)$ with respect to β, which further justifies the use of $L(\beta)$ as the loss function.

Lin and Lv (2013) applied a class of penalty function that includes Lasso and SCAD to the aforementioned loss function $L(\beta)$, and investigated their variable selection properties in a high dimensional framework. Under mild regularity conditions, they proved the weak oracle property (Lv and Fan 2009) and the oracle property for the penalized estimators. Gaiffas and Guilloux (2012) applied the same weighted L_1 penalty as in Leng and Ma (2007) and Martinussen and Scheike (2009) to a more general form of loss function which includes the $L(\beta)$ used in Leng and Ma (2007) and Martinussen and Scheike (2009) as a special case. The authors established non-asymptotic sharp oracle inequalities for the estimator under high dimensional setting using a new version of Bernstein's inequality.

2.5 Tuning Parameter Selection for Regularized Variable Selection Procedures

Tuning parameter selection plays a central role in the implementation of penalty based variable selection procedures. The realization of the desirable theoretical properties of the variable selection procedures in real data analyses is heavily dependent on the selection of the correct tuning parameters. In practice, tuning parameters are usually selected by a data-driven fashion that involves minimization of a certain criterion over the tuning parameter space. A grid search method is typically used to identify the minimizer of the selection criterion. There are two major categories of tuning parameter selection methods: K-fold cross-validation (Efron and Tibshirani 1993) and generalized cross-validation (GCV) (Craven and Wahba 1979). In K-fold cross-validation method, the full dataset D is evenly divided into K random subsets D^k ($k = 1, ..., K$). Denote the training and test set by $D - D^k$ and D^k, respectively. Denote the observed response and covariate vector for individual i

by (y_i, x_i). For each tuning parameter value λ over a pre-specified grid and subset D^k, a penalized estimator $\hat{\beta}^{(k)}(\lambda)$ is obtained using the training set $D - D^k$. Then the cross-validation criterion is given by

$$\mathrm{CV}(\lambda) = \sum_{k=1}^{K} \sum_{(y_i,x_i)\in D^k} \{y_i - x_i^T \hat{\beta}^{(k)}(\lambda)\}^2.$$

The $\hat{\lambda}$ is chosen as the minimizer of $\mathrm{CV}(\lambda)$. The K-fold cross-validation method is computationally intensive, and $\mathrm{CV}(\lambda)$ is less intuitive for right censored outcome such as survival time. As an alternative, generalized cross-validation has been widely used in tuning parameter selection for various penalty based variable selection procedures. The GCV criterion is defined in linear model as

$$\mathrm{GCV}(\lambda) = \frac{\|Y - X\hat{\beta}_\lambda\|^2}{n\{1 - e(\lambda)/n\}^2},$$

and defined in generalized linear model as

$$\mathrm{GCV}(\lambda) = \frac{-\ell_n(\hat{\beta}_\lambda)}{n\{1 - e(\lambda)/n\}^2},$$

where $\ell_n(\hat{\beta}_\lambda)$ is the log-likelihood function evaluated at the penalized estimates, $e(\lambda)$ is the effective number of parameters given by $e(\lambda) = \mathrm{tr}[X\{X^TX + n\Sigma_\lambda(\hat{\beta}_\lambda)\}^{-1}X^T]$ for linear model and $e(\lambda) = \mathrm{tr}[\{\ell_n''(\hat{\beta}_\lambda) - n\Sigma_\lambda(\hat{\beta}_\lambda)\}^{-1}\ell_n''(\hat{\beta}_\lambda)]$ for generalized linear model, and $\Sigma_\lambda(\hat{\beta}_\lambda) = \mathrm{diag}\{P_\lambda'(|\hat{\beta}_{\lambda 1}|)/|\hat{\beta}_{\lambda 1}|, ..., P_\lambda'(|\hat{\beta}_{\lambda p}|)/|\hat{\beta}_{\lambda p}|\}$. The GCV criterion can be deemed as a weighted version of the leave-one-out cross-validation criterion (Craven and Wahba 1979). In Cox proportional hazards model, the partial likelihood is used in the numerator of the GCV statistic.

Wang et al. (2007) demonstrated in linear model the similarity between GCV and traditional AIC criterion with a logarithm transformation of GCV

$$\log\{\mathrm{GCV}(\lambda)\} = \log(\|Y - X\hat{\beta}_\lambda\|^2/n) - 2\log\{1 - e(\lambda)/n\}.$$

When $e(\lambda) \gg n$ we have

$$\log\{\mathrm{GCV}(\lambda)\} \approx \log(\|Y - X\hat{\beta}_\lambda\|^2/n) + 2e(\lambda)/n,$$

which is analogous to the traditional AIC criterion. The authors showed with SCAD penalty that GCV is not a consistent selection criterion. Namely, tuning parameter selected by GCV criterion results in overfitted model with a positive probability as sample size goes to infinity. The authors proposed a new criterion that is analogous to the traditional BIC criterion, which is defined in linear model as

$$\mathrm{BIC}(\lambda) \equiv \log(\|Y - X\hat{\beta}_\lambda\|^2/n) + \log(n)e(\lambda)/n.$$

They showed that the BIC criterion can identify the true model with probability 1 as n goes to infinity. Zhang et al. (2010) obtained similar results in generalized linear models with nonconcave penalized likelihood. The authors introduced a generalized information criterion (GIC) defined as

$$\mathrm{GIC}(\lambda) \equiv D(y; \hat{\beta}_\lambda)/n + \kappa_n e(\lambda)/n, \tag{2.5}$$

where $D(y; \hat{\beta}_\lambda) = 2\{\ell_n(y; y) - \ell_n(\hat{\beta}_\lambda; y)\}$ is the deviance and κ_n is a positive constant chosen by the analyst. The tuning parameter is selected as the minimizer of GIC. The authors showed that when κ_n is bounded above, then the selected tuning parameter overfits the model with a positive probability, whereas when $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, then the selected

tuning parameter identifies the true model with probability tending to one.

Both Wang et al. (2007) and Zhang et al. (2010) considered tuning parameter selection in a finite dimensional setting where the number of candidate covariates is a fixed finite constant. Wang et al. (2009) extended the investigation on tuning parameter selection into the realm of diverging number of parameters. In linear model framework, the authors defined a slightly modified BIC criterion as

$$\mathrm{BIC}_\lambda(\hat{\beta}_\lambda) \equiv \log(\|Y - X\hat{\beta}_\lambda\|^2/n) + C_n \log(n)|S_\lambda|/n,$$

where $|S_\lambda|$ is the size of the model identified by tuning parameter λ and C_n is a positive constant chosen by the analyst. The selected tuning parameter minimizes this criterion. The authors showed that, under some regularity conditions, the BIC criterion consistently identifies the true model as n goes to infinity given that $C_n \to \infty$ and $C_n p \log(n)/n \to 0$, where p is the dimension of the parameters that goes to infinity with sample size, and $\|P'_\lambda(\hat{\beta}_{\lambda,a})\|^2 = o_p\{\log(n)/n\}$, where $\hat{\beta}_{\lambda,a}$ is the penalized estimates of the non-zero components of β_0. The authors showed that both SCAD and adaptive Lasso penalties satisfy the last condition.

Chen and Chen (2008) investigated the selection property of an extended BIC in high dimensional linear model where p grows at a polynomial rate with n. Assume that the model space S is partitioned into $\bigcup_{j=1}^{p} S_j$ such that models in each S_j have equal dimension. Let $\tau(S_j)$ be the size of S_j. Therefore, $\tau(S_j) = \binom{p}{j}$. The authors proposed an extended BIC for a model $s \in S_j$

$$\mathrm{BIC}_\gamma(s) \equiv -2\ell_n\{\hat{\theta}(s)\} + \nu(s)\log(n) + 2\gamma \log \tau(S_j),$$

where $\hat{\beta}(s)$ is the maximum likelihood estimator for model s, $\nu(s)$ is the size of model s, $\gamma \in [0, 1]$ is a constant that is related to the divergence rate of the model dimension. The

authors showed in linear model that when $p = O(n^\kappa)$ for some constant $\kappa \geq 0$, $\gamma > 1 - 1/(2\kappa)$, and certain asymptotic identifiability condition is satisfied, the extended BIC can identify the true model with probability tending to one as sample size goes to infinity. Wang and Zhu (2011) further proposed a new family of BIC-like criteria for ultra-high dimensional variable selection in linear model where $\log(p) = O(n^\kappa)$. The new criteria they proposed is defined as

$$\mathrm{HBIC}_\gamma(M) \equiv n \log\left(\frac{1}{n} RSS_M\right) + 2\gamma \log(p)|M|,$$

where $|M|$ is the size of model M, RSS_M is the residual sum of squares of model M, and $\gamma \geq 1$ is a constant as in the above definition of extended BIC (Chen and Chen 2008). Let integer K be the upper bound of the true model M_0 that is set by the researcher. This bound relieves the searching endeavor by focusing exclusively on the class of sub-models $\{M : |M| \leq K\}$. Under some regularity conditions, if $\gamma > 1$, for any K satisfying $K \log(p) = o(n)$, the authors proved that the HBIC_γ consistently selects the true model from the model space $\{M : |M| \leq K\}$ as sample size goes to infinity.

Fan and Tang (2013) studied tuning parameter selection in generalized linear model under ultra-high dimensional setting with $\log(p) = o(n)$. They used the GIC defined in (5.3) as the selection criterion. They introduced a quantity δ_n which they call the signal strength of the true model. For any model α, let $|\alpha|$ be the size of model. Define its "population parameter" $\beta^*(\alpha)$ to be the minimizer of the Kullback-Leibler (KL) distance $I\{\beta(\alpha)\} = \mathrm{E}_{\beta_0}[\log\{f_0(\beta_0)/f_\alpha(\beta(\alpha))\}]$, where f_0 and f_α are the density under the true model and model α, respectively. Note that the expectation is taken under the true model. Let K be the upper bound of the true model as described in Wang and Zhu (2011). For $K = o(n)$, the signal strength δ_n is defined as

$$\delta_n \equiv \inf_{\alpha \not\supset \alpha_0,\ |\alpha| \leq K} \frac{1}{n} I\{\beta^*(\alpha)\}.$$

Then under some regularity conditions, the authors showed that GIC is a consistent tuning parameter selector provided the constant κ_n diverges to infinity at a rate that is a function of δ_n, K, parameter dimension p, and true model size s, and the form of the function depends on whether the outcome variable is bounded, Gaussian, or unbounded non-Gaussian. They recommended for practical implementation to use a uniform choice of $\kappa_n = \log\{\log(n)\}\log(p)$.

CHAPTER 3: REGULARIZED VARIABLE SELECTION FOR COX PROPORTIONAL HAZARDS MODEL WITH A CASE-COHORT DESIGN

3.1 Introduction

Large-scale epidemiological studies and disease prevention trials often need to follow thousands of subjects for an extended period of time. The assembly of covariates for the entire study cohort can be prohibitively expensive, especially when it requires precious biological samples or expensive bioassays. Moreover, the occurrence rate of the event of interest is usually low in these studies, especially for such events as cardiovascular disease, cancer, or death. We refer to subjects who develop the event during the study as cases and the others as noncases. If the covariates were to be measured for everyone in the study, most of the cost would be spent on noncases, which do not contribute as much information as cases. To reduce the cost and effort in collecting expensive covariates without decreasing much efficiency in the analysis of time-to-event data, Prentice (1986) proposed the case-cohort design, where the complete covariate information is only obtained from a random subcohort of the sample plus all cases.

Various estimation methods have been developed for case-cohort studies under the proportional hazard model (Cox 1972). Prentice (1986) and Self and 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 estimate for the regression parameters. Kalbfleisch and Lawless (1988) proposed a more efficient weight that uses the complete covariate history of all cases. Borgan et al. (2000) further studied several types of weight

under the stratified case-cohort design. Kulich and Lin (2004) rigorously proved the asymptotic properties of the efficiently weighted estimator (Kalbfleisch and Lawless 1988). Kang and Cai (2009) extended the weighted estimator to studies with multivariate failure time outcome. Kim et al. (2013) further improved the efficiency of the estimators for case-cohort studies with multivariate failure time outcome. In this chapter, we focus on the efficient weight proposed by Kalbfleisch and Lawless (1988) in a univariate unstratified case-cohort design.

In the large epidemiological studies that use the case-cohort design a large number of covariates are usually collected, especially with the increasing availability of the electronic medical record data. Thus, one research goal is often to identify a subset of them that are related to the event of interest. With the inclusion of interaction terms 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 sample size n. Therefore, an efficient variable selection procedure that allows a diverging number of parameters is needed for the case-cohort design. In this chapter of the dissertation, we consider the scenario where the model size d_n diverges to infinity at a slower rate than the sample size. Therefore, $d_n \to \infty$ but $d_n \ll n$. The traditional variable selection methods such as stepwise and best subset selection are known to be computationally intensive and unstable. Since the introduction of Lasso method by Tibshirani (1996), penalty-based variable selection procedures have achieved great success. Under certain regularity conditions, these procedures can simultaneously select variables and estimate their coefficients. Many penalty functions have been proposed, among which the smoothly clipped absolute deviation (Fan and Li 2001), adaptive Lasso (Zou 2006), adaptive elastic net (Zou and Zhang 2009), and minimax concave (Zhang 2010) penalties have been shown to possess the so-called oracle property, namely, as n goes to infinity, the procedure correctly identifies the true model with probability one and estimates the

standard errors of nonzero parameters as efficiently as if the zero parameters were never included in the estimation process. Fan and Li (2002) applied the smoothly clipped absolute deviation penalty to the proportional hazard model and proved its oracle property. Cai et al. (2005) further extended the penalized partial likelihood procedure to multivariate models with a diverging number of parameters. However, to our knowledge, the properties of penalized variable selection procedure have not been studied under the case-cohort design where not all covariates are fully observed. This chapter intends to fill this gap.

3.2 Pseudo-Partial Likelihood for Case-Cohort Design

Suppose there are n independent subjects in a cohort. Let T and C be respectively the time to the outcome of interest and the censoring time. Let $Z_i(t)$ be the $d_n \times 1$ possibly time-dependent covariate vector for subject i at time t. Let $\beta = (\beta_1, ..., \beta_{d_n})^T$ be a vector of unknown regression coefficients. Let $X = \min(T, C)$ be the observed time and $\Delta = I(T \leq C)$ be the censoring indicator, where $I(\cdot)$ is an indicator function. T and C are assumed to be independent conditional on Z. Define for subject i the counting process $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, and the at risk process $Y_i(t) = I(X_i \geq t)$. Let $\lambda_i(t)$ denote the hazard function for subject i. Cox (1972) proposed the proportional hazard model where $\lambda_i\{t|Z_i(t)\} = \lambda_0(t)\exp\{\beta^T Z_i(t)\}$, where $\lambda_0(t)$ is an unspecified baseline hazard function. Under the case-cohort design, suppose we randomly select a subcohort of fixed size $\tilde{n}$ from the full cohort of size n. Let ξ_i denote the indicator for the ith subject being selected into the subcohort, and $\alpha = \tilde{n}/n = \Pr(\xi_i = 1)$ denote the selection probability of the ith subject. Here we consider simple random sampling without replacement with fixed subcohort size. Under this sampling scheme $(\xi_1, ..., \xi_n)$ are correlated. The covariate histories are not observed for censored subjects outside the subcohort. Assuming the complete covariate histories are available for all the cases, one can use the following pseudo-partial likelihood

to estimate the regression coefficients β (Kalbfleisch and Lawless 1988):

$$\tilde{\ell}_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta^T Z_i(t) - \log \sum_{j=1}^{n} \rho_j(t) Y_j(t) \exp\{\beta^T Z_j(t)\} \right] dN_i(t), \tag{3.1}$$

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

$$\tilde{U}_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \frac{\tilde{S}_n^{(1)}(\beta, t)}{\tilde{S}_n^{(0)}(\beta, t)} \right\} dN_i(t) = 0, \tag{3.2}$$

where $\tilde{S}_n^{(k)}(\beta, t) = n^{-1} \sum_{i=1}^{n} \rho_i(t) Y_i(t) Z_i(t)^{\otimes k} e^{\beta' Z_i(t)}$ for $k = 0, 1, 2$. For a vector a, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$.

3.3 Variable Selection with a Penalized Pseudo-Partial Likelihood

3.3.1 Penalized Pseudo-Partial Likelihood

We define a penalized pseudo-partial likelihood as

$$\tilde{Q}_n(\beta) = \tilde{\ell}_n(\beta) - n \sum_{j=1}^{d_n} P_{\lambda_{jn}}(|\beta_j|), \tag{3.3}$$

where $P_{\lambda_{jn}}(|\beta_j|)$ is a nonnegative penalty function with λ_{jn} as the nonnegative tuning parameter controlling the model complexity. We use smoothly clipped absolute deviation penalty proposed by Fan and Li (2001) with the modification of covariate-specific tuning parameters λ_{jn}, which allows different regression coefficients to have different penalty functions. When $\lambda_{jn} = 0$, no penalty is applied to β_j. The first derivative of the penalty is

$$P'_{\lambda_{jn}}(\theta) = \lambda_{jn} I(\theta \le \lambda_{jn}) + \frac{(a\lambda_{jn} - \theta)_+}{a - 1} I(\theta > \lambda_{jn})$$

for some $a > 2$ and $\theta > 0$, with $P_{\lambda_{jn}}(0) = 0$.

3.3.2 Notations and Regularity Conditions

We denote by $\hat{\beta}$ the penalized pseudo-partial likelihood estimator that maximizes (3.3). We denote by β_0 the true value of β. Let $\beta_0 = (\beta_{I0}^T, \beta_{II0}^T)^T$, where β_{I0} and β_{II0} are the nonzero and zero components of β_0, respectively. Let $\hat{\beta} = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$, where $\hat{\beta}_I$ and $\hat{\beta}_{II}$ are the penalized pseudo-partial likelihood estimators of β_{I0} and β_{II0}, respectively. Denote by k_n the dimension of β_{I0} and k_n/d_n converges to a constant $c \in [0, 1]$. For each n, we define the following notations:

$$S_n^{(k)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) Z_i(t)^{\otimes k} e^{\beta' Z_i(t)}, \quad k = 0, 1, 2,$$

$$\tilde{S}_n^{(k)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} \rho_i(t) Y_i(t) Z_i(t)^{\otimes k} e^{\beta' Z_i(t)}, \quad k = 0, 1, 2,$$

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

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

$$s_n^{(k)}(\beta, t) = \mathrm{E}\{S_n^{(k)}(\beta, t)\}, k = 0, 1, 2, \quad e_n(\beta, t) = s_n^{(1)}(\beta, t)/s_n^{(0)}(\beta, t),$$

$$I_n(\beta) = \mathrm{E}\left\{\int_0^\tau V_n(\beta, t) S_n^{(0)}(\beta, t) d\Lambda_0(t)\right\}, \qquad \Gamma_n(\beta) = \frac{1}{n}\mathrm{var}\{\tilde{\ell}_n'(\beta)\},$$

$$a_n = \max_{1 \le j \le k_n} \{|P'_{\lambda_{jn}}(|\beta_{j0}|)|\}, \qquad b_n = \max_{1 \le j \le k_n} \{|P''_{\lambda_{jn}}(|\beta_{j0}|)|\},$$

$$\Sigma_n = \mathrm{diag}\{P''_{\lambda_{1n}}(|\beta_{10}|), ..., P''_{\lambda_{k_n n}}(|\beta_{k_n 0}|)\},$$

$$B_n = \{P'_{\lambda_{1n}}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), ..., P'_{\lambda_{k_n n}}(|\beta_{k_n 0}|)\mathrm{sgn}(\beta_{k_n 0})\}^T.$$

We require the following regularity conditions:

(A) $\int_0^\tau \lambda_0(t) dt < \infty$.

(B) $\mathrm{E}\{Y(\tau)\} > 0$.

(C) $\mid Z_{ij}(0) \mid + \int_0^\tau |dZ_{ij}(t)| < C_1 < \infty$ almost surely for some constant C_1 and $i = 1, ..., n$ and $j = 1, ..., d_n$. That is, $Z_{ij}(t)$ has bounded variation almost surely.

(D) There exists a neighborhood $\mathscr{B}$ of β_0 such that for all $\beta \in \mathscr{B}$ and $t \in [0, \tau]$, $\partial s_n^{(0)}(\beta, t)/\partial\beta = s_n^{(1)}(\beta, t)$, and $\partial^2 s_n^{(0)}(\beta, t)/\partial\beta\partial\beta^T = s_n^{(2)}(\beta, t)$. The functions $s_n^{(k)}(\beta, t)$ $(k = 0, 1, 2)$ are continuous and bounded and $s_n^{(0)}(\beta, t)$ is bounded away from 0 on $\mathscr{B} \times [0, \tau]$.

(E) $\alpha = \tilde{n}/n$ converges to a constant $C_2 \in (0, 1]$ as $n \to \infty$.

(F) For each n, there exist positive constants C_3, C_4, C_5, and C_6 such that

$$0 < C_3 < \mathrm{eigen}_{\min}\{I_n(\beta_0)\} \leq \mathrm{eigen}_{\max}\{I_n(\beta_0)\} < C_4 < \infty,$$
$$0 < C_5 < \mathrm{eigen}_{\min}\{\Gamma_n(\beta_0)\} \leq \mathrm{eigen}_{\max}\{\Gamma_n(\beta_0)\} < C_6 < \infty,$$

where $\mathrm{eigen}_{\min}\{\cdot\}$ and $\mathrm{eigen}_{\max}\{\cdot\}$ are the minimum and maximum eigenvalues of a matrix.

(G) $\min_{1 \leq j \leq k_n} |\beta_{0j}|/\lambda_{jn} \to \infty$ as $n \to \infty$.

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

3.3.3 Asymptotic Properties of Penalized Pseudo-Partial Likelihood Estimator

Throughout this dissertation we use $O_p(\cdot)$ and $o_p(\cdot)$ to denote in probability order relations and $O(\cdot)$ and $o(\cdot)$ to denote almost sure order relations. We first prove the existence of a penalized pseudo-partial likelihood estimator that converges at rate $O_p\{d_n^{1/2}(n^{-1/2} + a_n)\}$, and then establish its oracle property. The proofs of Theorem 3.3.1 and 3.3.2 are provided in section 3.7.

Theorem 3.3.1. *Under Conditions (A) to (G), if* $b_n \to 0$ *and* $d_n^4/n \to 0$ *as* $n \to \infty$, *then with probability tending to one there exists a local maximizer* $\hat{\beta}$ *of* $\tilde{Q}_n(\beta) = \tilde{\ell}_n(\beta) - n\sum_{j=1}^{d_n} P_{\lambda_{jn}}(|\beta_j|)$, *such that* $\|\hat{\beta} - \beta_0\| = O_p\{d_n^{1/2}(n^{-1/2} + a_n)\}$.

From Theorem 3.3.1 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 smoothly clipped absolute deviation penalty under Condition (G). This consistency rate is the same as that of the maximum likelihood estimator for the exponential family (Portnoy 1988).

Theorem 3.3.2. *(Oracle property) Under Conditions (A) to (H), if* $b_n \to 0, d_n^5/n \to 0, \lambda_{jn} \to 0, \lambda_{jn}(n/d_n)^{1/2} \to \infty$, *and* $a_n = O(n^{-1/2})$ *as* $n \to \infty$, *the* $(n/d_n)^{1/2}$*-consistent local maximizer* $\hat{\beta} = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ *must satisfy that* $\hat{\beta}_{II} = 0$ *with probability tending to one and for any nonzero* $k_n \times 1$ *constant vector* u *with* $u^T u = 1$,

$$n^{1/2} u^T \Gamma_{n11}^{-1/2}(I_{n11} + \Sigma_n)\{\hat{\beta}_I - \beta_{I0} + (I_{n11} + \Sigma_n)^{-1} B_n\} \to N(0,1)$$

in distribution, where I_{n11} *consists of the first* $k_n \times k_n$ *components of* $I_n(\beta_0)$, *and* Γ_{n11} *consists of the first* $k_n \times k_n$ *components of* $\Gamma_n(\beta_0)$.

The matrix $I_n(\beta_0)$ can be estimated by $\hat{I}_n(\hat{\beta}) = n^{-1}\sum_{i=1}^n \int_0^\tau \tilde{V}_n(\hat{\beta}, t) dN_i(t)$. The estimation of matrix $\Gamma_n(\beta_0)$ is derived in section 3.7. For the smoothly clipped absolute deviation penalty, $a_n = 0$, $\Sigma_n = 0$, and $B_n = 0$ for large n under Condition (G). Therefore, the result of Theorem 3.3.2 reduces to

$$n^{1/2} u^T \Gamma_{n11}^{-1/2} I_{n11}(\hat{\beta}_I - \beta_{I0}) \to N(0,1)$$

in distribution. The conditions $d_n^4/n \to 0$ and $d_n^5/n \to 0$ in the above theorems only describe the divergence rate of d_n when sample size goes to infinity. They do not impose any one-to-one relationship between finite d_n and n.

3.4 Considerations in Practical Implementation

3.4.1 Local Quadratic Approximation and Variance Estimation

Since the smoothly clipped absolute deviation penalty function is singular at the origin, in practical implementation the Newton-Raphson algorithm cannot be directly applied to maximize (3.3). Instead, we use a modified Newton-Raphson algorithm with a local quadratic approximation to the penalty function. The unpenalized pseudo-partial likelihood (3.1) can be seen as a special case of the penalized pseudo-partial likelihood (3.3) with $P_{\lambda_{jn}}(|\beta_j|) = 0$ for all $j = 1, ..., d_n$. Applying Theorem 3.3.1 with $a_n = 0$, we know there exists a $(n/d_n)^{1/2}$-consistent maximizer of (3.1). We use this maximizer as the initial value $\beta^{(0)}$ for the modified Newton-Raphson algorithm. If $|\beta_j^{(0)}|$ is less than a pre-specified small positive constant c_j, then set $\hat{\beta}_j = 0$. Otherwise, the penalty function is locally approximated by a quadratic function as

$$P_{\lambda_{jn}}(|\beta_j|) \approx P_{\lambda_{jn}}(|\beta_j^{(0)}|) + P'_{\lambda_{jn}}(|\beta_j^{(0)}|)(2|\beta_j^{(0)}|)^{-1}(\beta_j^2 - \beta_j^{(0)2})$$

and therefore $P'_{\lambda_{jn}}(|\beta_j|) \approx \{P'_{\lambda_{jn}}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j$. With the approximated penalty function, one step Newton-Raphson algorithm is performed and the updated nonzero estimate is used as the new initial value. The process is iterated until convergence or no nonzero estimate is left.

The sandwich estimate of the covariance matrix for $\hat{\beta}$ can be directly obtained from the last iteration of the above algorithm as $\hat{\text{cov}}(\hat{\beta}) = \{\tilde{\ell}''_n(\hat{\beta}) - n\Sigma_\lambda(\hat{\beta})\}^{-1}\hat{\text{var}}\{\tilde{\ell}'_n(\hat{\beta})\}\{\tilde{\ell}''_n(\hat{\beta}) - n\Sigma_\lambda(\hat{\beta})\}^{-1}$, where $\Sigma_\lambda(\beta) = \text{diag}\{P'_{\lambda_{1n}}(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, ..., P'_{\lambda_{d_n n}}(|\beta_{d_n}^{(0)}|)/|\beta_{d_n}^{(0)}|\}$. The sandwich estimate of the covariance matrix is only applicable to the nonzero estimate of the parameters.

3.4.2 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 control the complexity of the selected model. In practical implementation, the properties of the penalized estimator heavily depend on the choice of the appropriate tuning parameters. The typical methods of selecting the tunng parameters are data-driven procedures such as K-fold cross-validation and generalized cross-validation (Craven and Wahba 1979). We use the generalized cross-validation method in our implementation. For proportional hazard model the effective number of parameters is defined as $e(\lambda_{1n}, ..., \lambda_{d_n n}) = \text{tr}[\{\tilde{\ell}_n''(\hat{\beta}) - n\Sigma_\lambda(\hat{\beta})\}^{-1}\tilde{\ell}_n''(\hat{\beta})]$. The generalized cross-validation statistic is defined as

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

The optimal tuning parameters are chosen as $\text{argmin}_{(\lambda_{1n},...,\lambda_{d_n n})}\text{GCV}(\lambda_{1n}, ..., \lambda_{d_n n})$. The d_n-dimensional optimization problem is difficult to solve in practice. We follow Cai et al. (2005) to take $\lambda_{jn} = \lambda_n \hat{\text{se}}(\beta_j^{(0)})$, where $\hat{\text{se}}(\beta_j^{(0)})$ is the estimated standard error of the unpenalized pseudo-partial likelihood estimator used in section 3.4.1. Then the optimization problem reduces to 1-dimensional search for the optimal λ_n.

When $e(\lambda_n)/n$ is small, as is the case under the conditions for Theorem 3.3.1 and 3.3.2, the log-transformation of $\text{GCV}(\lambda_n)$ can be approximated by

$$\log\{\text{GCV}(\lambda_n)\} = \log\{-\tilde{\ell}_n(\hat{\beta})/n\} - 2\log\{1 - e(\lambda_n)/n\} \approx \log\{-\tilde{\ell}_n(\hat{\beta})/n\} + 2e(\lambda_n)/n.$$

This expression is analogous to the Akaike information criterion (Akaike 1973). Therefore, we denote $\log\{\text{GCV}(\lambda_n)\}$ as $\text{AIC}(\lambda_n)$, and define $\lambda_n^{\text{AIC}} := \text{argmin}_{\lambda_n}\text{AIC}(\lambda_n)$. Wang et al. (2007) and Zhang et al. (2010) showed in linear and generalized linear models with finite number of parameters that $\text{AIC}(\lambda_n)$ overfits the model with a positive probability as

$n \to \infty$. Following the idea of Bayesian information criterion (Schwarz 1978), we define another tuning parameter selection criteria, where the optimal tuning parameter, denoted by λ_n^{BIC}, minimizes $\text{BIC}(\lambda_n) := \log\{-\tilde{\ell}_n(\hat{\beta})/n\} + \log(n)e(\lambda_n)/n$. In the simulation section that follows, we will empirically investigate the performance of the tuning parameter λ_n^{AIC} and λ_n^{BIC} in penalty-based variable selection. Following Fan and Li (2001), we set the second tuning parameter a in the smoothly clipped absolute deviation penalty function to 3.7 in our simulation.

In practice, researchers can perform a grid search to identify λ_n^{AIC} and λ_n^{BIC}. The lower limit of the search range is 0 and the upper limit is the minimum λ_n that gives an empty model. From our simulation experience, the upper limit rarely exceeds 2. Moreover, the model selection result is fairly insensitive to the fineness of the search grid.

3.5 Numerical Study and Application

3.5.1 Simulation Study

Independent failure times are generated from the proportional hazard model. We set $\lambda_0(t) = 2$ and model dimension $d_n = [5n_c^{1/5-1/500}]$ to reflect its dependence on sample size, where n_c is the number of cases and $[x]$ rounds x to the nearest integer. We relate the model dimension to the number of cases rather than sample size as the former better represents the amount of information in the dataset. The first component of β is the smallest nonzero parameter in terms of the absolute value and is set to either 0.34 (large effect scenario with corresponding hazard ratio of 1.4) or 0.18 (small effect scenario with corresponding hazard ratio of 1.2). There is one nonzero parameter for every two zero parameters, with the other nonzero parameters recycling from values 0.6 and -0.8. For example, when $d_n = 15$, $\beta_1 = 0.34$, then $\beta = (0.34, 0, 0, 0.6, 0, 0, -0.8, 0, 0, 0.6, 0, 0, -0.8, 0, 0)$. We generate the design matrix Z as a mixture of correlated binary and continuous variables. First, d_n-dimensional multivariate standard normal variable Z^* are generated with the correlation coefficient

between Z_i^* and Z_j^* being $0.5^{|i-j|}$. Then the first three components of Z^* are kept as continuous, and the next three components are dichotomized at 0, and this pattern is repeated for the rest of Z^*. Thus half of the covariates become binary with parameter 0.5. Censoring times C_i are generated from a uniform distribution $\mathrm{U}(0, c)$ where c is adjusted to achieve desired censoring percentage.

Two sample sizes, two censoring rates, and two noncase to case ratios are considered for each β_1 value (0.34 or 0.18). Performance of penalized variable selection procedures with tuning parameter λ_n^{AIC} and λ_n^{BIC} are assessed. As a benchmark, we include the hard threshold variable selection procedure, where the component of the the unpenalized maximum pseudo-partial likelihood estimator from the full model is selected if it p-value from the Wald test is less than 0.05. We also include the result from 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% and 90% in the simulation. For each setting 1000 replications are conducted.

We define model error of a variable selection procedure as $\mathrm{ME}(\hat{\mu}) = \mathrm{E}\{\mathrm{E}(T|z) - \hat{\mu}(z)\}^2$, and the relative model error as the ratio of its model error to that of the unpenalized pseudo-partial likelihood estimates from the full model. We use the median and the median absolute deviation of the relative model error to compare the performance of different variable selection procedures. We also calculate the average number of parameters correctly estimated as 0, the average number of parameters erroneously estimated as 0, and the overall rate of identifying the true model. Point estimates, empirical and model-based standard errors, and the empirical 95% confidence interval coverage are also calculated for $\hat{\beta}_1$ using replications with nonzero $\hat{\beta}_1$.

Table 3.1 summarizes the variable selection performance under large effect size (β_1 = 0.34). Larger sample size, lower censoring rate, and higher noncase to case ratio are associated with better variable selection performance in all three methods. The penalized

method with λ_n^{BIC} outperforms the other two methods in all settings. The inferior performance of λ_n^{AIC} is apparently due to its overfitting effect as shown by the low average number of correctly identified zero parameters. This is consistent with the theoretical findings from Wang et al. (2007) and Zhang et al. (2010) that λ_n^{AIC} overfits the model with a positive probability when n goes to infinity in linear and generalized linear models. Table 3.2 summarizes the parameter estimation of β_1 under the same settings as in Table 3.1. Given that β_1 is correctly identified as nonzero, all procedures produce approximately unbiased point and standard error estimates and the 95% confidence interval coverage is close to the nominal level. The parameter is slightly overestimated under 90% censoring rate. This is due to the fact that very small $\hat{\beta}_1$ are set to 0 in the variable selection algorithm and therefore excluded from the computation of the average of point estimates. This bias decreases as the variable selection performance improves.

Table 3.3 summarizes the variable selection performance under small effect size scenario ($\beta_1 = 0.18$). Similar patterns are observed as in Table 3.1, although the variable selection performance of all three procedures decreases substantially. Nevertheless, the procedure with λ_n^{BIC} outperforms the other procedures in all settings. Even with small effect size, λ_n^{BIC} method performs almost as well as the oracle procedure when $n = 10000$ with 80% censoring rate, which is a reasonable setting for case-cohort study. Table 3.4 shows the parameter estimation of β_1 under settings in Table 3.3. Conditional on correctly identifying β_1 all procedures perform reasonably well in parameter estimation. Again, slight overestimation is observed under 90% censoring rate for the same reason as described before, which disappears when the variable selection performance increases.

We also conducted simulation with smaller effect size ($\beta_1 = 0.095$ corresponding to hazard ratio= 1.1). The sample size and censoring rate needed to achieve reasonable variable selection performance under this effect size become unrealistic. The result is not shown due to space limit.

Finally, the normality of the sampling distributions of $\hat{\beta}_1$ under all scenarios is graphically assessed by Q-Q plots (Figure 3.1 to 3.4). It can be seen that the sampling distribution of $\hat{\beta}_1$ is a mixture of a point mass at 0 and a left-truncated distribution, which is well approximated by a truncated normal distribution as indicated by the straight line in the plots and the conditional 95% confidence interval coverage in Table 3.2 and 3.4. Furthermore, from the Q-Q plots and Table 3.1 and 3.3, the number of 0 estimates decreases as the rate of identifying the true model increases.

Table 3.1: Model selection performance with large effect size (β_1 = 0.34, hazard ratio = 1.4)

	80% Censored				90% Censored			
	RME	Zero Parm.		RITM	RME	Zero Parm.		RITM
Method	median (MAD)	C	I	(%)	median (MAD)	C	I	(%)
	n = 2500, noncase:case = 1:1, d_n = 17 for 80% censored, d_n = 15 for 90% censored							
HT	0.69 (0.23)	10.28	0.05	47.9	0.92 (0.31)	9.09	0.75	15.7
SCAD(AIC)	0.67 (0.23)	9.75	0.02	29.3	0.93 (0.16)	6.36	0.22	0.9
SCAD(BIC)	0.46 (0.3)	10.97	0.27	74.8	0.77 (0.35)	9.24	0.7	21.8
Oracle	0.35 (0.18)	11	0	100	0.33 (0.18)	10	0	100
	n = 2500, noncase:case = 2:1, d_n = 17 for 80% censored, d_n = 15 for 90% censored							
HT	0.69 (0.21)	10.35	0	52.9	0.78 (0.33)	9.33	0.36	35.5
SCAD(AIC)	0.54 (0.22)	10.46	0	58.4	0.82 (0.2)	7.58	0.06	8
SCAD(BIC)	0.39 (0.21)	11	0.14	86.7	0.58 (0.38)	9.75	0.47	49.3
Oracle	0.37 (0.18)	11	0	100	0.32 (0.16)	10	0	100
	n = 5000, noncase:case = 1:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored							
HT	0.67 (0.21)	12.23	0	46.2	0.8 (0.28)	10.2	0.25	35.1
SCAD(AIC)	0.64 (0.21)	11.79	0	31.5	0.89 (0.13)	7.2	0.03	2.1
SCAD(BIC)	0.35 (0.17)	12.99	0.01	98.1	0.57 (0.29)	10.48	0.23	49.1
Oracle	0.34 (0.17)	13	0	100	0.35 (0.16)	11	0	100
	n = 5000, noncase:case = 2:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored							
HT	0.65 (0.21)	12.25	0	48.2	0.68 (0.21)	10.32	0.06	48.5
SCAD(AIC)	0.48 (0.2)	12.49	0	62	0.8 (0.17)	8.41	0.01	7
SCAD(BIC)	0.34 (0.15)	13	0	100	0.42 (0.21)	10.85	0.08	81.2
Oracle	0.34 (0.15)	13	0	100	0.35 (0.16)	11	0	100

RME: relative model error; MAD: median absolute deviation; C: average number of 0 parameters correctly identified as 0; I: average number of nonzero parameters incorrectly identified as 0; RITM: rate of identifying true model; HT: hard threshold method; SCAD(AIC): smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC): smoothly clipped absolute deviation with λ_n^{BIC}.

Table 3.2: Parameter estimation for β_1 with large effect size (β_1 = 0.34, hazard ratio = 1.4)

	80% Censored				90% Censored			
Method	$\hat{\beta}_1$	se_e	se_m	95% CI_e	$\hat{\beta}_1$	se_e	se_m	95% CI_e
n = 2500, noncase:case = 1:1, d_n = 17 for 80% censored, d_n = 15 for 90% censored								
HT	0.35	0.08	0.07	93.1	0.41	0.11	0.12	92.4
SCAD(AIC)	0.35	0.07	0.06	92.1	0.37	0.12	0.11	91.2
SCAD(BIC)	0.35	0.07	0.06	95.3	0.38	0.1	0.11	93
Oracle	0.34	0.07	0.06	93.7	0.34	0.12	0.11	92.4
n = 2500, noncase:case = 2:1, d_n = 17 for 80% censored, d_n = 15 for 90% censored								
HT	0.35	0.07	0.06	91.8	0.37	0.09	0.1	94.5
SCAD(AIC)	0.34	0.06	0.05	92.4	0.35	0.1	0.09	91.4
SCAD(BIC)	0.34	0.06	0.05	94.5	0.36	0.08	0.09	95
Oracle	0.34	0.06	0.05	93.7	0.35	0.09	0.09	93.3
n = 5000, noncase:case = 1:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored								
HT	0.34	0.05	0.05	93.6	0.36	0.09	0.09	92.5
SCAD(AIC)	0.34	0.05	0.05	93.1	0.36	0.09	0.08	90.3
SCAD(BIC)	0.34	0.05	0.05	94.5	0.36	0.08	0.08	93
Oracle	0.34	0.05	0.05	94.6	0.35	0.09	0.08	92.7
n = 5000, noncase:case = 2:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored								
HT	0.34	0.04	0.04	95.5	0.35	0.07	0.07	94.1
SCAD(AIC)	0.34	0.04	0.04	94	0.35	0.07	0.06	92.7
SCAD(BIC)	0.34	0.04	0.04	94.8	0.34	0.06	0.06	94.2
Oracle	0.34	0.04	0.04	94.8	0.34	0.06	0.06	94

se_e: empirical standard error; se_m: model-based standard error; 95% CI_e: empirical 95% confidence interval coverage; HT: hard threshold method; SCAD(AIC): smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC): smoothly clipped absolute deviation with λ_n^{BIC}. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_1$.

Table 3.3: Model selection performance with small effect size (β_1 = 0.18, hazard ratio = 1.2)

	80% Censored				90% Censored			
	RME	Zero Parm.		RITM	RME	Zero Parm.		RITM
Method	median (MAD)	C	I	(%)	median (MAD)	C	I	(%)
	n = 5000, noncase:case = 1:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored							
HT	0.66 (0.21)	12.21	0.06	43.4	0.79 (0.26)	10.17	0.63	21.5
SCAD(AIC)	0.63 (0.22)	11.75	0.02	29.1	0.89 (0.14)	7.27	0.17	1.6
SCAD(BIC)	0.42 (0.22)	12.98	0.33	66.7	0.6 (0.28)	10.45	0.6	30.5
Oracle	0.35 (0.16)	13	0	100	0.36 (0.16)	11	0	100
	n = 5000, noncase:case = 2:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored							
HT	0.65 (0.21)	12.27	0.01	48.2	0.7 (0.21)	10.29	0.32	33.6
SCAD(AIC)	0.5 (0.22)	12.5	0.01	61.6	0.79 (0.18)	8.49	0.09	8.4
SCAD(BIC)	0.44 (0.22)	13	0.26	74.2	0.48 (0.23)	10.83	0.45	51.2
Oracle	0.35 (0.16)	13	0	100	0.35 (0.16)	11	0	100
	n = 10000, noncase:case = 1:1, d_n = 23 for 80% censored, d_n = 20 for 90% censored							
HT	0.66 (0.18)	14.15	0	43.8	0.7 (0.2)	12.1	0.17	33.9
SCAD(AIC)	0.61 (0.18)	13.74	0	30.3	0.89 (0.14)	8.75	0.03	0.6
SCAD(BIC)	0.38 (0.17)	15	0.03	96.7	0.49 (0.21)	12.51	0.18	53.2
Oracle	0.37 (0.16)	15	0	100	0.33 (0.15)	13	0	100
	n = 10000, noncase:case = 2:1, d_n = 23 for 80% censored, d_n = 20 for 90% censored							
HT	0.66 (0.17)	14.16	0	44.8	0.67 (0.19)	12.26	0.07	44.8
SCAD(AIC)	0.49 (0.2)	14.55	0	65.3	0.79 (0.18)	10.27	0.02	6.3
SCAD(BIC)	0.39 (0.17)	15	0.02	98.4	0.42 (0.19)	12.85	0.12	77.4
Oracle	0.38 (0.17)	15	0	100	0.35 (0.16)	13	0	100

RME: relative model error; MAD: median absolute deviation; C: average number of 0 parameters correctly identified as 0; I: average number of nonzero parameters incorrectly identified as 0; RITM: rate of identifying true model; HT: hard threshold method; SCAD(AIC): smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC): smoothly clipped absolute deviation with λ_n^{BIC}.

Table 3.4: Parameter estimation for β_1 with small effect size (β_1 = 0.18, hazard ratio = 1.2)

	80% Censored				90% Censored			
Method	$\hat{\beta}_1$	se_e	se_m	95% CI_e	$\hat{\beta}_1$	se_e	se_m	95% CI_e
n = 5000, noncase:case = 1:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored								
HT	0.19	0.05	0.05	96.4	0.26	0.06	0.09	92.2
SCAD(AIC)	0.19	0.05	0.05	94.8	0.22	0.08	0.08	92.6
SCAD(BIC)	0.21	0.03	0.05	95.9	0.25	0.06	0.08	90.7
Oracle	0.18	0.05	0.05	94.5	0.19	0.08	0.08	92.3
n = 5000, noncase:case = 2:1, d_n = 20 for 80% censored, d_n = 17 for 90% censored								
HT	0.18	0.04	0.04	96.3	0.22	0.06	0.07	94.5
SCAD(AIC)	0.18	0.04	0.04	93.9	0.2	0.06	0.06	94.8
SCAD(BIC)	0.2	0.03	0.04	96.8	0.22	0.04	0.06	95.7
Oracle	0.18	0.04	0.04	94.1	0.18	0.06	0.06	94.8
n = 10000, noncase:case = 1:1, d_n = 23 for 80% censored, d_n = 20 for 90% censored								
HT	0.18	0.04	0.04	95	0.21	0.05	0.06	95.8
SCAD(AIC)	0.18	0.03	0.03	94	0.19	0.06	0.06	94.7
SCAD(BIC)	0.19	0.03	0.03	97.1	0.2	0.05	0.06	95.9
Oracle	0.18	0.03	0.03	94.9	0.19	0.06	0.06	94.7
n = 10000, noncase:case = 2:1, d_n = 23 for 80% censored, d_n = 20 for 90% censored								
HT	0.18	0.03	0.03	94.9	0.19	0.05	0.05	95.9
SCAD(AIC)	0.18	0.03	0.03	93.4	0.18	0.05	0.05	94.1
SCAD(BIC)	0.19	0.03	0.03	95.2	0.19	0.04	0.05	96.6
Oracle	0.18	0.03	0.03	93.7	0.18	0.05	0.05	94.7

se_e: empirical standard error; se_m: model-based standard error; 95% CI_e: empirical 95% confidence interval coverage; HT: hard threshold method; SCAD(AIC): smoothly clipped absolute deviation with λ_n^{AIC}; SCAD(BIC): smoothly clipped absolute deviation with λ_n^{BIC}. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_1$.

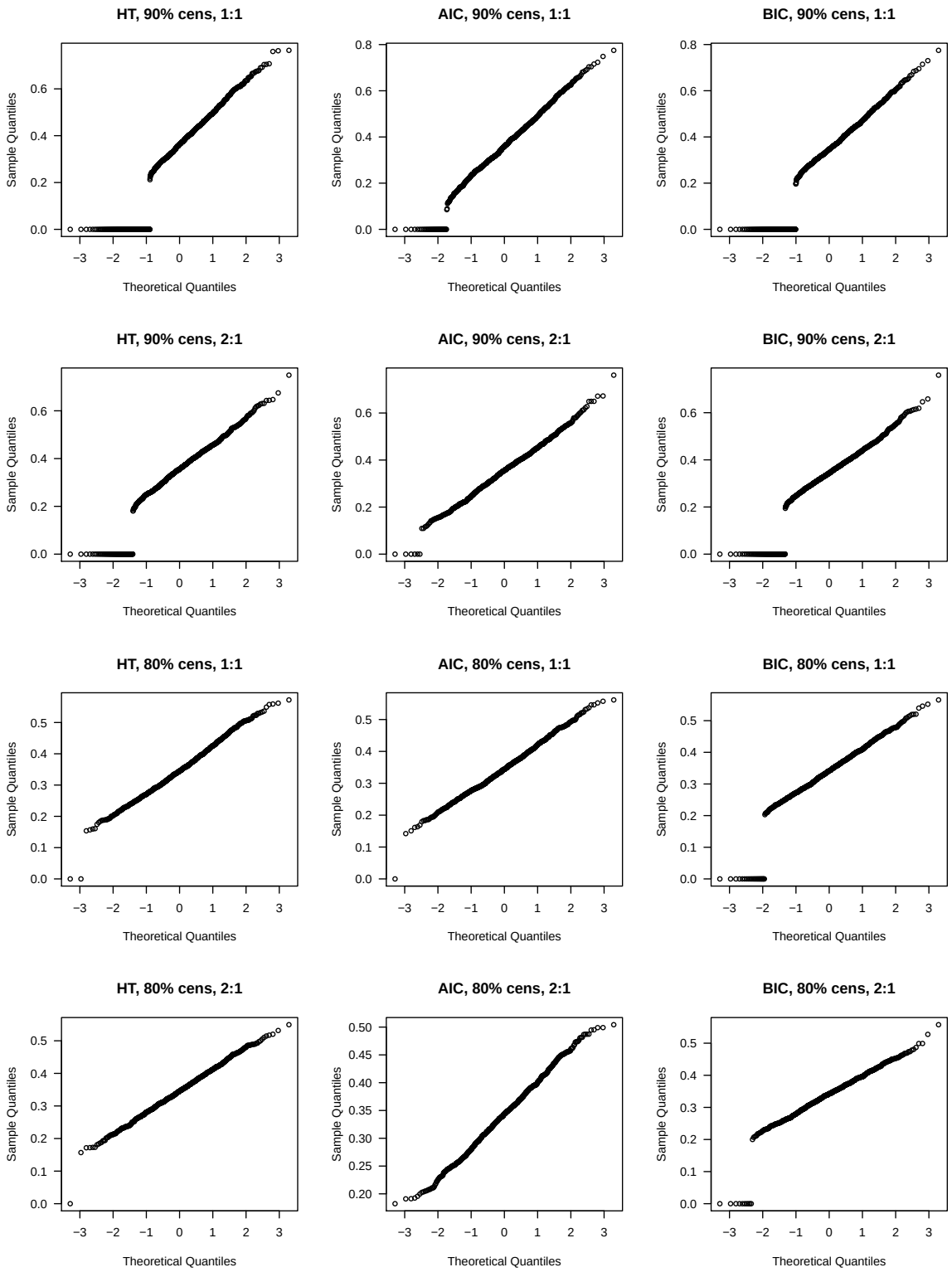

Figure 3.1: Q-Q plot of $\hat{\beta}$ for the smallest nonzero parameter by three procedures (Hard threshold, λ_n^{AIC}, λ_n^{BIC}). Sample size $n = 2500$. True $\beta = 0.34$ (hazard ratio= 1.4).

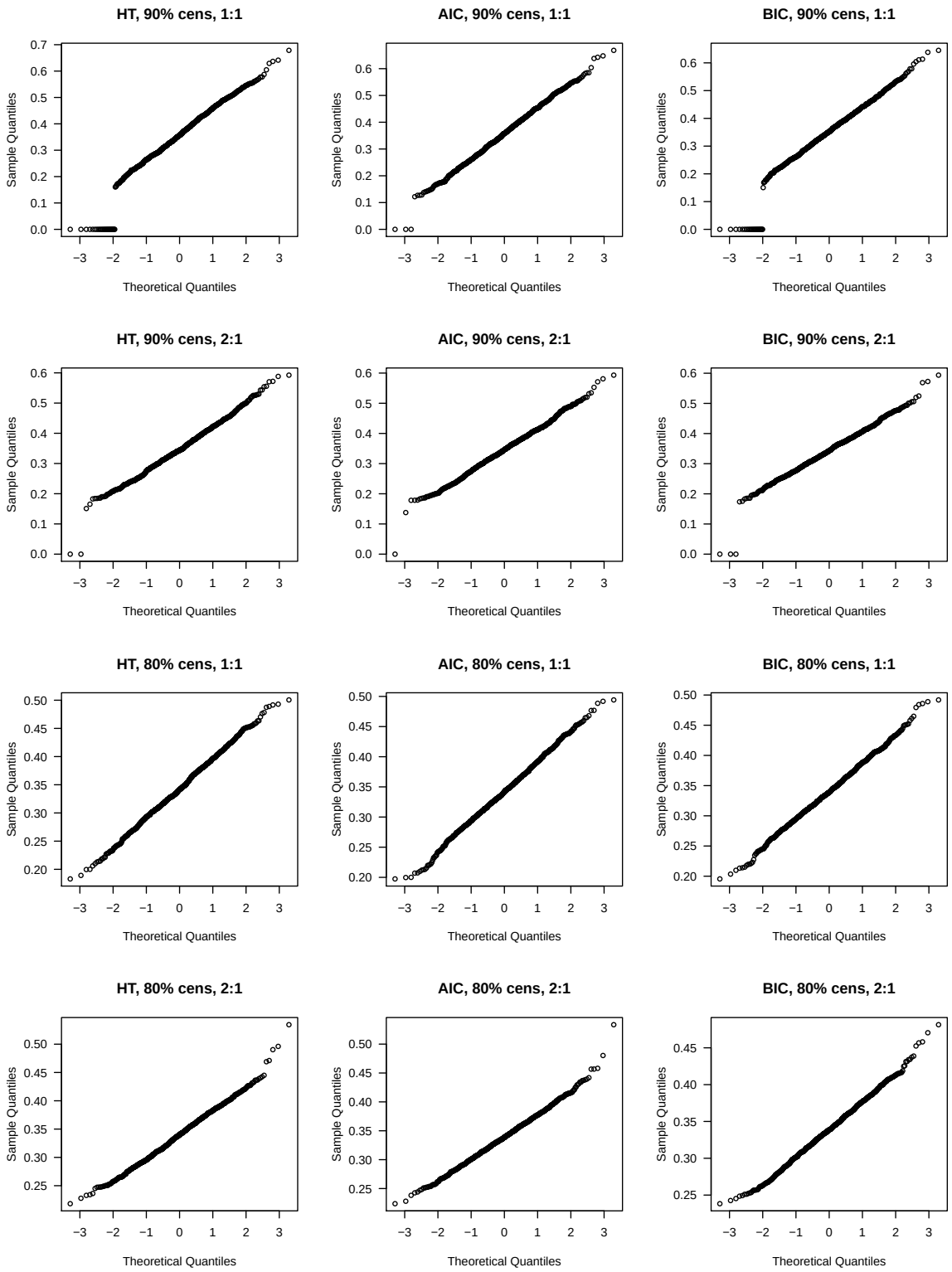

Figure 3.2: Q-Q plot of $\hat{\beta}$ for the smallest nonzero parameter by three procedures (Hard threshold, λ_n^{AIC}, λ_n^{BIC}). Sample size $n = 5000$. True $\beta = 0.34$ (hazard ratio= 1.4).

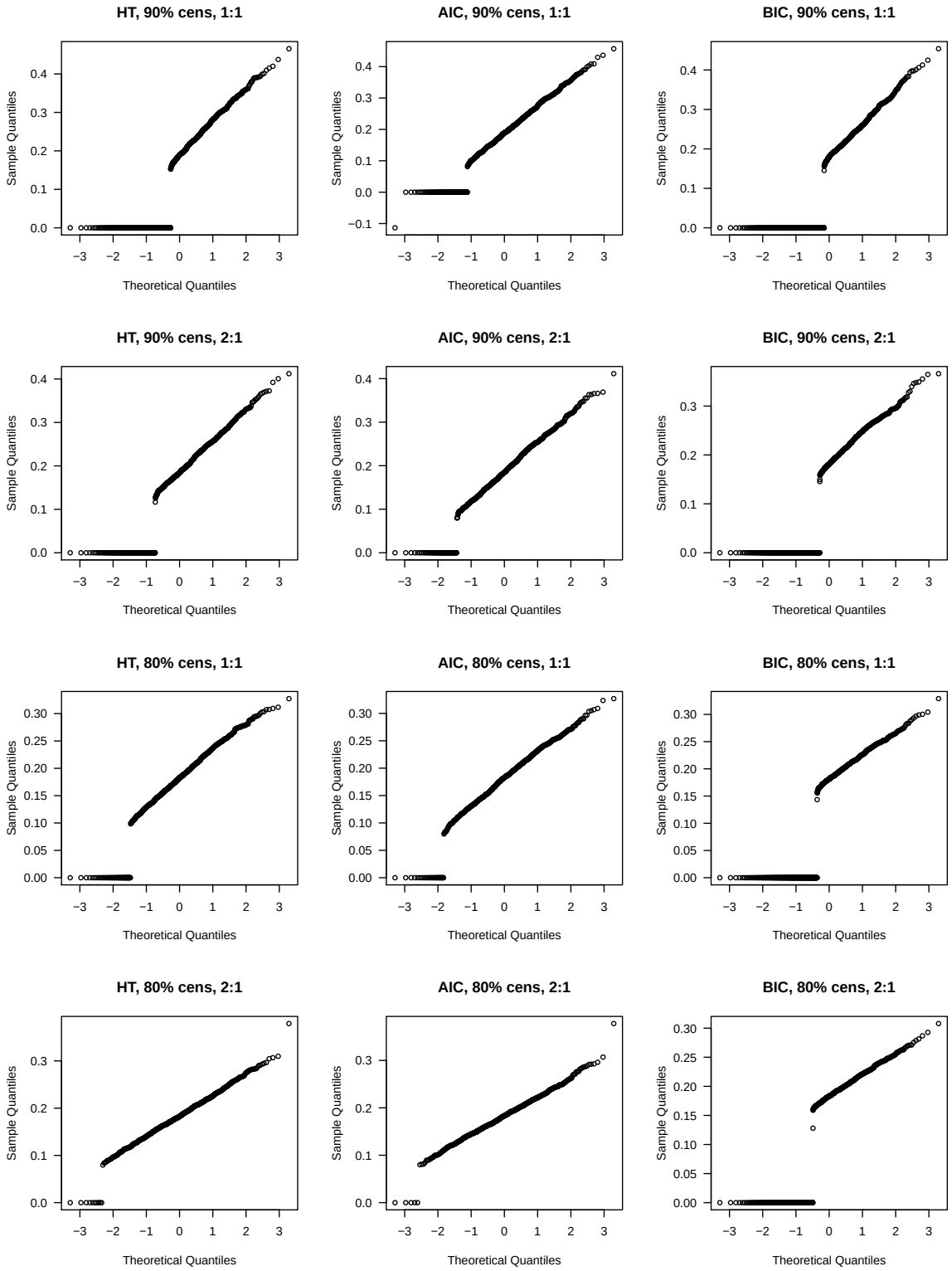

Figure 3.3: Q-Q plot of $\hat{\beta}$ for the smallest nonzero parameter by three procedures (Hard threshold, λ_n^{AIC}, λ_n^{BIC}). Sample size $n = 5000$. True $\beta = 0.18$ (hazard ratio= 1.2).

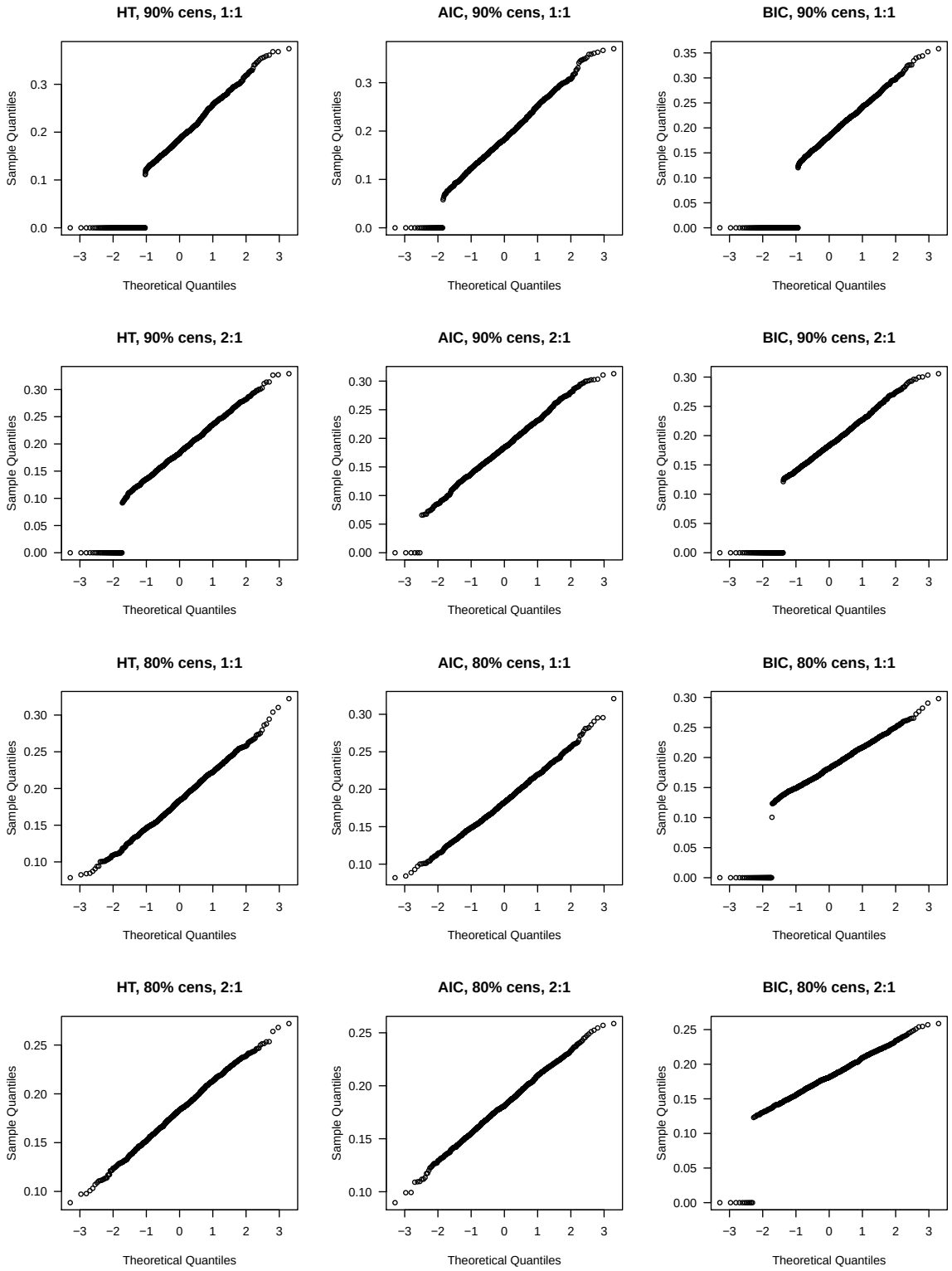

Figure 3.4: Q-Q plot of $\hat{\beta}$ for the smallest nonzero parameter by three procedures (Hard threshold, λ_n^{AIC}, λ_n^{BIC}). Sample size $n = 10000$. True $\beta = 0.18$ (hazard ratio= 1.2).

3.5.2 Analysis of Busselton Health Study

We use the proposed variable selection procedures to analyze the Busselton Health Study data (Cullen 1972, Knuiman et al. 2003). The study is a series of cross-sectional health surveys conducted in the town of Busselton in Western Australia. Every 3 years from 1966 to 1981, general health information for adult participants were collected by questionnaire and clinical visit. In this analysis we are interested in the effect of cardiovascular risk factors on the risk of stroke. In particular, the main risk factor of interest is the serum ferritin level. We also consider several other risk factors in the variable selection process: age (years), body mass index (BMI), blood pressure treatment (0=no, 1=yes), systolic blood pressure (mmHg), cholesterol (mmol/L), triglycerides (mmol/L), hemoglobin (g/100ml), and smoking (1=never, 2=former, 3=current). 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 December 31, 1998, and their time to stroke was recorded if any. They were treated as censored if they left Western Australia during the follow-up period. There were 118 (8.4%) 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 a randomly selected subcohort plus all stroke cases only. The random subcohort size was 450, and the case-cohort size was 513.

Table 3.5 summarizes the baseline characteristics of the full cohort and the subcohort. The average ferritin level is not available for the full cohort due to the case-cohort design. The summary statistics of the baseline characteristics are similar between the full cohort and sub-cohort, suggesting that the subcohort is representative of the full cohort.

We apply the hard threshold, penalty with tuning parameter λ_n^{AIC} and λ_n^{BIC} variable selection procedures to the Busselton Health Study data to identify important risk factors for stroke. In order not to miss any potentially important effects, we also include the

Table 3.5: Baseline characteristics of the Busselton Health Study

Variables	Full cohort (n=1401) Mean (SD) or %	Subcohort ($\tilde{n}$=450) Mean (SD) or %
Age (yrs)	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)
Hemoglobin (g/100ml)	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)

quadratic terms of all continuous covariates as well as interactions between ferritin and all covariates in the initial model. The total number of parameters is 32. To decrease the skewness in the distribution we log-transform ferritin and triglycerides values. The following continuous covariates are standardized: age, body mass index, systolic blood pressure, cholesterol, log(triglycerides), and hemoglobin. The tuning parameter selector identified $\lambda_n^{\text{AIC}} = 0.1724$ and $\lambda_n^{\text{BIC}} = 0.2405$. Table 3.6 shows the selected terms and their estimated coefficients and standard errors by the two penalized procedures with λ_n^{AIC} and λ_n^{BIC}. The λ_n^{BIC} selected 16 terms and λ_n^{AIC} selected additional 6 terms. This is consistent with the fact that λ_n^{AIC} tends to select more variables than λ_n^{BIC}. Both methods selected the main effect of log(ferritin) and a number of interaction, suggesting that the effect of ferritin on risk of stroke is modified by other risk factors. The 6 terms selected by only λ_n^{AIC} are squared systolic blood pressure, squared log(triglycerides), hemoglobin, log(ferritin)*hemoglobin, log(ferritin)*squared log(triglycerides), and log(ferritin)*sex. Hard threshold method only selected the blood pressure treatment into the final model.

Table 3.6: Estimated coefficients and standard errors from Busselton Health Study data

Variable	SCAD (BIC) $\hat{\beta}$ (ŝe)	SCAD (AIC) $\hat{\beta}$ (ŝe)
Age (yrs)	1.76 (0.28)	1.55 (0.27)
Age^2	-0.58 (0.02)	-0.57 (0.4)
Sex (1=female)	0 (–)	0 (–)
Body mass index	0 (–)	0 (–)
Body mass index^2	0 (–)	0 (–)
Blood pressure treatment	0.73 (0.26)	0.80 (0.27)
Systolic blood pressure	1.06 (0.06)	1.04 (0.71)
Systolic blood pressure^2	0 (–)	0.12 (0.01)
Cholesterol	0 (–)	0 (–)
Cholesterol^2	-0.59 (0.01)	-0.62 (0.03)
log(triglycerides)	0 (–)	0 (–)
log^2(triglycerides)	0 (–)	-0.30 (0.02)
Hemoglobin	0 (–)	0.24 (0.004)
Hemoglobin^2	0.19 (0.06)	0.25 (0.07)
Smoking (former vs. never)	2.12 (1.42)	2.04 (1.43)
Smoking (current vs. never)	2.23 (1.20)	2.26 (1.22)
log(ferritin)	0.40 (0.13)	0.27 (0.09)
log(ferritin)*body mass index	0 (–)	0 (–)
log(ferritin)*body mass index^2	0 (–)	0 (–)
log(ferritin)*age	-0.20 (0.03)	-0.14 (0.02)
log(ferritin)*age^2	0.12 (0.03)	0.11 (0.09)
log(ferritin)*cholesterol	0 (–)	0 (–)
log(ferritin)*cholesterol^2	0.11 (0.02)	0.12 (0.02)
log(ferritin)*hemoglobin	0 (–)	-0.05 (0.02)
log(ferritin)*hemoglobin^2	-0.03 (0.02)	-0.05 (0.02)
log(ferritin)*systolic blood pressure	-0.16 (0.02)	-0.19 (0.15)
log(ferritin)*systolic blood pressure^2	0 (–)	0 (–)
log(ferritin)*log(triglycerides)	0 (–)	0 (–)
log(ferritin)*log^2(triglycerides)	0 (–)	0.08 (0.01)
log(ferritin)*sex	0 (–)	-0.10 (0.02)
log(ferritin)*smoking (former vs. never)	-0.38 (0.28)	-0.41 (0.29)
log(ferritin)*smoking (current vs. never)	-0.42 (0.26)	-0.46 (0.26)

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

3.6 Discussion

In this chapter of the dissertation we proposed a variable selection procedure based on smoothly clipped absolute deviation penalized pseudo-partial likelihood in case-cohort studies with failure time outcome. We showed that under certain regularity conditions, as sample size goes to infinity, the variable selection procedure identifies the true model with probability tending to one, and the nonzero estimate from this procedure is consistent and asymptotically normally distributed. Moreover, the nonzero estimate is estimated as efficient as if the true model is known by the investigator. The theorems presented in this chapter only establish local consistency and oracle property in the neigborhood of β_0. Due to the non-convexity of the penalty function, there may be multiple maximizers for the penalized objective function. However, since the initial value $\beta^{(0)}$ for the local quadratic approximation algorithm is $(n/d_n)^{1/2}$-consistent, the maximizer identified by this algorithm will also be likely to converge to β_0.

Our simulation study found that the penalized variable selection procedure with tuning parameter selected by Bayesian information criteria performs much better than that selected by Akaike information criterion. The poor performance of variable selection with tuning parameter λ_n^{AIC} may seem inconsistent with previous simulation studies such as Fan and Li (2002), Cai et al. (2005) where the finite sample performance of λ_n^{AIC} is quite good despite its theoretical property of overfitting the model with positive probability. However, those studies used much lower censoring rates (15-40%) than our simulations. Our results demonstrate that in survival analysis with high censoring rate, as is usually the case in case-cohort studies, the overfitting effect of λ_n^{AIC} becomes prominent, and λ_n^{BIC} works much better in comparison. Based on our simulation results of different noncase to case ratios, we also recommend including more noncases in a case-cohort design if possible to improve the accuracy of the proposed variable selection procedure.

Since the smoothly clipped absolute deviation penalty is a non-linear function of the

parameter, the variable selection result is not invariant to covariate standardization. In practice, we recommend standardization of continuous covariates before carrying out the proposed variable selection procedure so that the estimated coefficients are comparable across covariates. For covariates that are not available for all subjects due to the case-cohort design, the random sub-cohort should be used to compute the sample mean and standard deviation for standardization. Another practical issue is that as the number of noncases in the random subcohort becomes small, $\hat{\alpha}(t)^{-1}$ becomes less reliable. When there is no noncase left in the subcohort, $\hat{\alpha}(t)^{-1}$ is not well defined. In practice, to avoid this difficulty, we recommend selecting the stopping time τ such that there are at least 10 subjects at risk from the subcohort on $[0, \tau]$.

The proposed variable selection procedure does not guarantee a hierarchical final model. Although it does not pose any theoretical difficulty, it makes the interpretation less straightforward. This is a future research topic that could incorporate a group penalized variable selection method into case-cohort design to ensure hierarchical model structure.

With any given sample size the proposed procedure may not be able to detect some very small effect, resulting in false negative finding. By decreasing the tuning parameter size one can decrease the false negative rate but it also increases the false positive rate. Therefore, the proposed procedure bares a trade-off between the two types of error under a finite sample as any other variable selection methods do. If some covariates are known scientifically to be associated with the risk of outcome, the investigator can set the tuning parameters to 0 for them to ensure their inclusion in the final model.

3.7 Proof of Theorems

Throughout the proofs, we denote $\tilde{\ell}'_n(\beta_0)_j = \partial\tilde{\ell}_n(\beta_0)/\partial\beta_j$, $\tilde{\ell}''_n(\beta_0)_{jk} = \partial^2\tilde{\ell}_n(\beta_0)/\partial\beta_j\partial\beta_k$, and $\tilde{\ell}'''_n(\beta_0)_{jkl} = \partial^3\tilde{\ell}_n(\beta_0)/\partial\beta_j\partial\beta_k\partial\beta_l$. We let $\tilde{V}_{njk}(\beta_0, t)$, $V_{njk}(\beta_0, t)$, $\tilde{S}^{(2)}_{njk}(\beta_0, t)$, and $S^{(2)}_{njk}(\beta_0, t)$ be the (j, k) component of corresponding matrices. For a matrix $A = \{a_{ij}\}, (i, j = 1, ..., n)$,

the norm is defined as $\|A\| = (\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2)^{1/2}$. The following two lemmas will be used repeatedly in the proof of the theorems.

Lemma 3.7.1. *Let $\xi = (\xi_1, ..., \xi_n)$ be a random vector containing $\tilde{n}$ ones and $n - \tilde{n}$ zeros, with each permutation equally likely. Let $B_i(t), i = 1, ..., n$ be i.i.d. real-valued random processes on $[0, \tau]$ with $\mathrm{E}\{B(t)\} = \mu_B(t), \mathrm{var}\{B(0)\} < \infty$ and $\mathrm{var}\{B(\tau)\} < \infty$. Let $B(t) = (B_1(t), ..., B_n(t))$ and ξ be independent. Suppose that almost all paths of $B_i(t)$ have finite variation. Then $n^{-1/2}\sum_{i=1}^{n}\xi_i\{B_i(t) - \mu_B(t)\}$ converges weakly to a tight zero mean Gaussian process and therefore $n^{-1}\sum_{i=1}^{n}\xi_i\{B_i(t) - \mu_B(t)\}$ converges in probability to 0 uniformly in t.*

The proof of this lemma can be found in Lemma A1 in Kang and Cai (2009). Under finite population sampling, $\mu_B(t) = n^{-1}\sum_{i=1}^{n} B_i(t)$. It follows that $n^{-1/2}\sum_{i=1}^{n}\xi_i\{B_i(t) - \mu_B(t)\} = n^{-1/2}\sum_{i=1}^{n}(\xi_i - \tilde{n}/n)B_i(t) = n^{-1/2}\alpha\sum_{i=1}^{n}(\xi_i/\alpha - 1)B_i(t)$.

Lemma 3.7.2. *Let $W_n(t)$ and $G_n(t)$ be two sequences of processes with bounded variation almost surely, and $G_n(t)$ is progressively measurable and cadlag. For some constant τ, assume that $\sup_{0\le t\le\tau}\|W_n(t) - W(t)\| \to 0$ in probability for some bounded process $W(t)$, $W_n(t)$ is monotone on $[0, \tau]$, and $G_n(t)$ converges to a zero mean process with continuous sample paths. Then both $\sup_{0\le t\le\tau}\left\|\int_0^t\{W_n(s) - W(s)\}dG_n(s)\right\|$ and $\sup_{0\le t\le\tau}\left\|\int_0^t G_n(s)d\{W_n(s) - W(s)\}\right\|$ converge to 0 in probability.*

The proof of this lemma can be found in Lemma 1 in Lin (2000).

We also need the following lemmas.

Lemma 3.7.3. *Given that ξ is independent of Δ and $Y(t)$, $n^{1/2}\{\hat{\alpha}^{-1}(t) - \alpha^{-1}\}$ converges to a zero-mean Gaussian process.*

Proof. By Taylor expansion of $\hat{\alpha}(t)$ around α,

$$n^{1/2}\{\hat{\alpha}^{-1}(t) - \alpha^{-1}\} = -\frac{n^{1/2}}{\alpha^*(t)^2}\{\hat{\alpha}(t) - \alpha\}$$

$$= -\frac{n^{1/2}}{\alpha^*(t)^2}\left\{\frac{\sum_{i=1}^n(1-\Delta_i)\xi_i Y_i(t)}{\sum_{i=1}^n(1-\Delta_i)Y_i(t)} - \alpha\right\}$$
$$= \frac{\alpha}{\alpha^*(t)^2}\frac{n}{\sum_{i=1}^n(1-\Delta_i)Y_i(t)} n^{-1/2}\sum_{i=1}^n\left(1-\frac{\xi_i}{\alpha}\right)(1-\Delta_i)Y_i(t),$$

where $\alpha^*(t)$ lies between $\hat{\alpha}(t)$ and α. Since $\text{var}\{(1-\Delta_i)Y_i(0)\} < \infty$, $\text{var}\{(1-\Delta_i)Y_i(\tau)\} < \infty$, and $(1-\Delta)Y(t)$ is of bounded variation, by Lemma 3.7.1, $n^{-1/2}\sum_{i=1}^n(\xi_i/\alpha - 1)(1-\Delta_i)Y_i(t)$ converges weakly to a tight zero mean Gaussian process. This implies that $n^{-1}\sum_{i=1}^n(\xi_i/\alpha - 1)(1-\Delta_i)Y_i(t)$ converges to 0 in probability uniformly in $t \in [0,\tau]$. Since $n^{-1/2}\sum_{i=1}^n[(1-\Delta_i)Y_i(t) - \text{E}\{(1-\Delta)Y(t)\}]$ can be seen as a special case of the expression $n^{-1/2}\sum_{i=1}^n\xi_i[(1-\Delta_i)Y_i(t) - \text{E}\{(1-\Delta)Y(t)\}]$ with $\xi_i = 1$ for all i, by Lemma 3.7.1 it converges weakly to a zero mean Gaussian process. This implies that $n^{-1}\sum_{i=1}^n(1-\Delta_i)Y_i(t)$ converges to $\text{E}\{(1-\Delta)Y(t)\}$ in probability uniformly in t. Under Conditions (A) and (B), $\text{E}\{(1-\Delta)Y(t)\}$ is uniformly bounded away from 0 on $[0,\tau]$. By law of large numbers and Slutsky's theorem, under Condition (E), it follows that $\hat{\alpha}(t)$ and α converge to the same constant limit C_2 uniformly in t. Therefore, $\alpha^*(t)$ and α also converge to the same limit. By Slutsky's theorem,

$$n^{1/2}\{\hat{\alpha}^{-1}(t) - \alpha^{-1}\} = \frac{1}{\alpha\text{E}\{(1-\Delta)Y(t)\}} n^{-1/2}\sum_{i=1}^n\left(1-\frac{\xi_i}{\alpha}\right)(1-\Delta_i)Y_i(t) + o_p(1),$$

which converges to a zero mean Gaussian process. □

Lemma 3.7.4. *Under Conditions (C) and (D), for any nonzero $d_n \times 1$ constant vector u with $\|u\| = C < \infty$ and $\|u\|_0 = c_n > 0$ where $\|\cdot\|_0$ denotes the number of nonzero components of a vector, $n^{1/2}\{\tilde{S}_n^{(0)}(\beta_0,t) - S_n^{(0)}(\beta_0,t)\}$, $(n/c_n)^{1/2}u^T\{\tilde{S}_n^{(1)}(\beta_0,t) - S_n^{(1)}(\beta_0,t)\}$, and $n^{1/2}c_n^{-1}u^T\{\tilde{S}_n^{(2)}(\beta_0,t) - S_n^{(2)}(\beta_0,t)\}u$ all converge to tight zero mean Gaussian processes.*

Proof. The three processes can be written in a unified form as $(k = 0,1,2)$,

$$n^{1/2}\left[n^{-1}\sum_{i=1}^n\rho_i(t)Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k - n^{-1}\sum_{i=1}^n Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k\right]$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{i=1}^{n}\{\Delta_i + (1-\Delta_i)\xi_i\hat{\alpha}(t)^{-1}\}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&\quad - n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\xi_i\alpha^{-1}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&\quad + n^{-1/2}\sum_{i=1}^{n}\Big[(1-\Delta_i)\xi_i\alpha^{-1}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k - Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k\Big] \\
&= n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\xi_i\hat{\alpha}(t)^{-1}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&\quad - n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\xi_i\alpha^{-1}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&\quad + n^{-1/2}\sum_{i=1}^{n}\Big[(1-\Delta_i)\xi_i\alpha^{-1}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&\quad -(1-\Delta_i)Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k\Big] \\
&= n^{-1/2}\sum_{i=1}^{n}\{\hat{\alpha}(t)^{-1}-\alpha^{-1}\}(1-\Delta_i)\xi_i Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&\quad - n^{-1/2}\sum_{i=1}^{n}\left(1-\frac{\xi_i}{\alpha}\right)(1-\Delta_i)Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k \\
&= \left[\frac{n^{-1/2}}{\mathrm{E}\{(1-\Delta)Y(t)\}}\sum_{j=1}^{n}\left(1-\frac{\xi_j}{\alpha}\right)(1-\Delta_j)Y_j(t) + o_p(1)\right]\times \\
&\quad \left[\frac{1}{n}\sum_{i=1}^{n}(1-\Delta_i)\frac{\xi_i}{\alpha}Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k\right] \\
&\quad - n^{-1/2}\sum_{i=1}^{n}\left(1-\frac{\xi_i}{\alpha}\right)(1-\Delta_i)Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k. \qquad (3.4)
\end{aligned}
$$

The last equality holds by Lemma 3.7.3. By Cauchy-Schwarz inequality, $u^T Z_i(t) \le \|u\|\|Z_i(t)\| = C\{\sum_{j=1}^{d_n} Z_{ij}^2(t)\}^{1/2}$. Under Condition (C), $Z_{ij}^2(t)$ has bounded variation, and therefore $c_n^{-1/2}u^T Z_i(t)$ has bounded variation. This along with Condition (D) gives that $(1-\Delta_i)Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k$ is of bounded variation for $i = 1, ..., n$. Therefore, by Lemma 3.7.1, $n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\xi_i/\alpha Y_i(t)e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k$ converges to a deterministic process $L(t)$ in probability uniformly on $[0,\tau]$. Therefore,

$$
\begin{aligned}
(3.4) &= n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\left(1-\frac{\xi_i}{\alpha}\right)Y_i(t)\left[\frac{L(t)}{\mathrm{E}\{(1-\Delta)Y(t)\}} - e^{\beta_0^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k\right] \\
&\quad + o_p(1). \qquad (3.5)
\end{aligned}
$$

Under Conditions (C) and (D) the term in the square brackets of (3.5) is of bounded variation. It follows by Lemma A3.7.1 that (3.5) converges weakly to a tight zero mean Gaussian process. Therefore, $n^{1/2}\{\tilde{S}_n^{(0)}(\beta_0,t)-S_n^{(0)}(\beta_0,t)\}$, $(n/c_n)^{1/2}u^T\{\tilde{S}_n^{(1)}(\beta_0,t)-S_n^{(1)}(\beta_0,t)\}$, and $n^{1/2}c_n^{-1}u^T\{\tilde{S}_n^{(2)}(\beta_0,t)-S_n^{(2)}(\beta_0,t)\}u$ all converge weakly to tight zero mean Gaussian processes. □

Lemma 3.7.5. *Under Conditions (A) to (D), for any nonzero $d_n \times 1$ constant vector u with $\|u\| = 1$, $n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}_n'(\beta_0)$ converges to a standard normal distribution, where $\Gamma_n(\beta_0)$ is the covariance matrix of $n^{-1/2}\tilde{\ell}_n'(\beta_0)$.*

Proof. Let $c_n = \|u\|_0$, the number of nonzero components of u. We first consider the quantity $(nc_n)^{-1/2}u^T\tilde{\ell}_n'(\beta_0)$, which can be decomposed as

$$\begin{aligned}(nc_n)^{-1/2}u^T\tilde{\ell}_n'(\beta_0) &= (nc_n)^{-1/2}u^T\sum_{i=1}^n\int_0^\tau\left\{Z_i(t)-\frac{S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}\right\}dN_i(t)\\ &\quad+(nc_n)^{-1/2}u^T\sum_{i=1}^n\int_0^\tau\left\{\frac{S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}-\frac{\tilde{S}_n^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)}\right\}dN_i(t)\\ &= I_1+I_2.\end{aligned}$$

I_1 is a linear combination of the partial likelihood score vector of the full cohort data. The score vector was shown by Andersen and Gill (1982) to converge to a zero mean multivariate normal distribution. Therefore, I_1 converges to a zero mean normal distribution.

I_2 can be further decomposed as

$$\begin{aligned}I_2 &= \int_0^\tau c_n^{-1/2}\left\{\frac{u^TS_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}-\frac{u^T\tilde{S}_n^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)}\right\}d\left\{n^{-1/2}\sum_{i=1}^n M_i(t)\right\}\\ &\quad+\int_0^\tau(nc_n)^{-1/2}\left\{\frac{u^TS_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}-\frac{u^T\tilde{S}_n^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)}\right\}\sum_{i=1}^n Y_i(t)e^{\beta_0^TZ_i(t)}d\Lambda_0(t).\end{aligned}\tag{3.6}$$

The first term on the right-hand side of (3.6) can be written as

$$\begin{aligned}
&\int_0^\tau c_n^{-1/2} \left\{ \frac{u^T S_n^{(1)}(\beta_0, t)}{S_n^{(0)}(\beta_0, t)} - \frac{u^T \tilde{S}_n^{(1)}(\beta_0, t)}{\tilde{S}_n^{(0)}(\beta_0, t)} \right\} d\left\{ n^{-1/2} \sum_{i=1}^n M_i(t) \right\} \\
&= \int_0^\tau c_n^{-1/2} \left\{ \frac{u^T S_n^{(1)}(\beta_0, t)}{S_n^{(0)}(\beta_0, t)} - u^T e_n(\beta_0, t) \right\} d\left\{ n^{-1/2} \sum_{i=1}^n M_i(t) \right\} \\
&\quad - \int_0^\tau c_n^{-1/2} \left\{ \frac{u^T \tilde{S}_n^{(1)}(\beta_0, t)}{\tilde{S}_n^{(0)}(\beta_0, t)} - u^T e_n(\beta_0, t) \right\} d\left\{ n^{-1/2} \sum_{i=1}^n M_i(t) \right\}. \qquad (3.7)
\end{aligned}$$

Under Conditions (C) and (D) along with $\|u\| = 1$, $c_n^{-1/2} u^T S_n^{(1)}(\beta_0, t)/S_n^{(0)}(\beta_0, t)$, and $c_n^{-1/2} u^T \tilde{S}_n^{(1)}(\beta_0, t)/\tilde{S}_n^{(0)}(\beta_0, t)$ are of bounded variation, so they can both be written as sum of two monotone functions in t. By the fact that $s_n^{(k)}(\beta, t) = \mathrm{E}\{S_n^{(k)}(\beta, t)\}$ for $k = 0, 1, 2$ and Lemma 3.7.1 (with $\xi_i = 1$ for all i) it is easy to show that $n^{1/2}\{S_n^{(0)}(\beta_0, t) - s_n^{(0)}(\beta_0, t)\}$ and $(n/c_n)^{1/2} u^T \{S_n^{(1)}(\beta_0, t) - s_n^{(1)}(\beta_0, t)\}$ converge weakly to tight zero mean Gaussian processes. It is then straightforward from Lemma 3.7.4 that $n^{1/2}\{\tilde{S}_n^{(0)}(\beta_0, t) - s_n^{(0)}(\beta_0, t)\}$ and $(n/c_n)^{1/2} u^T \{\tilde{S}_n^{(1)}(\beta_0, t) - s_n^{(1)}(\beta_0, t)\}$ converge weakly to tight zero mean Gaussian processes. Thus, we have that $c_n^{-1/2} u^T S_n^{(1)}(\beta_0, t)/S_n^{(0)}(\beta_0, t) - c_n^{-1/2} u^T e_n(\beta_0, t)$ and $c_n^{-1/2} u^T \tilde{S}_n^{(1)}(\beta_0, t)/\tilde{S}_n^{(0)}(\beta_0, t) - c_n^{-1/2} u^T e_n(\beta_0, t)$ both converge to 0 in probability uniformly in $t \in [0, \tau]$.

On the other hand, $\sum_{i=1}^n M_i(t)$ is a sum of i.i.d. random processes whose sample paths are of bounded variation under Condition (C). Therefore, $M_i(t)$ can be decomposed into two monotone functions in t. Since $\mathrm{E}\{M_i(t)\} = 0$, it follows from the Example 2.11.16 of van der Vaart and Wellner (1996) (p215) that $n^{-1/2} \sum_{i=1}^n M_i(t)$ converges weakly to a tight zero mean Gaussian process, say $G_M(t)$. It can be shown that $\mathrm{E}\{G_M(t) - G_M(s)\}^4 \leq C_M(t - s)^2$ for all $t, s \in [0, \tau]$ and some constant C_M. Therefore, by Kolmogorov-Centsov Theorem (Karatzas and Shereve, 1988, p53), $G_M(t)$ has continuous sample path almost surely. Since $G_M(t)$ is also of bounded variation almost surely, it follows from Lemma 3.7.2 that both terms of (3.7) converge to 0 in probability. Therefore, the first term on the

right-hand side of (3.6) converges to 0 in probability.

For the second term on the right-hand side of (3.6) we have

$$
\begin{aligned}
&(nc_n)^{-1/2}\int_0^\tau \left\{\frac{u^T S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)} - \frac{u^T \tilde{S}_n^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)}\right\}\sum_{i=1}^n Y_i(t)e^{\beta_0^T Z_i(t)}d\Lambda_0(t)\\
&=\left(\frac{n}{c_n}\right)^{1/2}\int_0^\tau \left\{u^T S_n^{(1)}(\beta_0,t) - u^T \tilde{S}_n^{(1)}(\beta_0,t) + \frac{u^T \tilde{S}_n^{(1)}(\beta_0,t)\tilde{S}_n^{(0)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)}\right.\\
&\quad\left. - \frac{u^T \tilde{S}_n^{(1)}(\beta_0,t)S_n^{(0)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)}\right\}d\Lambda_0(t)\\
&=\int_0^\tau \left[\left(\frac{n}{c_n}\right)^{1/2} u^T\left\{S_n^{(1)}(\beta_0,t) - \tilde{S}_n^{(1)}(\beta_0,t)\right\}\right.\\
&\quad\left. - \left(\frac{n}{c_n}\right)^{1/2}\left\{S_n^{(0)}(\beta_0,t) - \tilde{S}_n^{(0)}(\beta_0,t)\right\}u^T e_n(\beta_0,t)\right]d\Lambda_0(t) + o_p(1). \qquad (3.8)
\end{aligned}
$$

By Lemma 3.7.4, $(n/c_n)^{1/2}u^T\{S_n^{(1)}(\beta_0,t) - \tilde{S}_n^{(1)}(\beta_0,t)\}$, $n^{1/2}\{S_n^{(0)}(\beta_0,t) - \tilde{S}_n^{(0)}(\beta_0,t)\}$ converge to tight zero mean Gaussian processes. Let $e_{jn}(\beta_0,t)$ be the j^{th} component of $e_n(\beta_0,t)$ $(j = 1,...,d_n)$, and $e_{jn}^*(\beta_0,t) = I(u_j \neq 0)e_{jn}(\beta_0,t)$. Since $e_{jn}(\beta_0,t)$ is a bounded deterministic process, by Cauchy-Schwarz inequality, $c_n^{-1/2}u^T e_n(\beta_0,t) = c_n^{-1/2}u^T e_n^*(\beta_0,t) \leq c_n^{-1/2}\|u\|\|e_n^*(\beta_0,t)\| = c_n^{-1/2}O(c_n^{1/2}) = O(1)$. Hence by Slutsky theorem, $n^{1/2}\{S_n^{(0)}(\beta_0,t) - \tilde{S}_n^{(0)}(\beta_0,t)\}c_n^{-1/2}u^T e_n(\beta_0,t)$ converges to a tight zero mean Gaussian process. It then follows that the integrand of the integration in (3.8) converges to a tight zero mean Gaussian process, say $G(t)$. Therefore, (3.8) $= \int_0^\tau G(t)d\Lambda_0(t) + o_p(1)$. Under Condition (A), $\int_0^\tau G(t)d\Lambda_0(t)$ is a continuous linear function from $\ell^\infty[0,\tau]$ to $\mathbb{R}$. By the tightness of $G(t)$, it follows from Lemma 3.9.8 of van der Vaart and Wellner (1996) (p377) that $\int_0^\tau G(t)d\Lambda_0(t)$ is normally distributed with mean zero. Therefore, (3.8) converges to a zero mean normal distribution. It follows that I_2 converges to a zero mean normal distribution.

Finally, we need to show that I_1 and I_2 are independent of each other. I_2 can be written

as

$$I_2 = (nc_n)^{-1/2}u^T \int_0^\tau \left[\frac{\{\tilde{S}_n^{(0)}(\beta_0,t) - S_n^{(0)}(\beta_0,t)\}\tilde{S}_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)\tilde{S}_n^{(0)}(\beta_0,t)} - \frac{\tilde{S}_n^{(1)}(\beta_0,t) - S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)} \right] d\left\{ \sum_{i=1}^n N_i(t) \right\}.$$

Replace $\tilde{S}_n^{(0)}(\beta_0,t) - S_n^{(0)}(\beta_0,t)$ and $\tilde{S}_n^{(1)}(\beta_0,t) - S_n^{(1)}(\beta_0,t)$ in the above expression with (3.5), and denote $A_n^{(k)}(\beta,t) = n^{-1}\sum_{i=1}^n (1-\Delta_i)\xi_i\alpha^{-1}Y_i(t)e^{\beta^T Z_i(t)}\{c_n^{-1/2}u^T Z_i(t)\}^k$ where $k = 0, 1$. Then I_2 is asymptotically equivalent to

$$\begin{aligned} I_2 &= n^{-1/2}\sum_{i=1}^n (1-\Delta_i)\left(1-\frac{\xi_i}{\alpha}\right)\int_0^\tau Y_i(t)\left(\left[\frac{A_n^{(0)}(\beta_0,t)}{\mathrm{E}\{(1-\Delta)Y(t)\}} - e^{\beta^T Z_i(t)}\right]\times \right. \\ &\quad \frac{c_n^{-1/2}u^T\tilde{S}_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)\tilde{S}_n^{(0)}(\beta_0,t)} - \left[\frac{A_n^{(1)}(\beta_0,t)}{\mathrm{E}\{(1-\Delta)Y(t)\}} - e^{\beta^T Z_i(t)}c_n^{-1/2}u^T Z_i(t)\right]\times \\ &\quad \left.\frac{1}{S_n^{(0)}(\beta_0,t)}\right) d\left\{\frac{1}{n}\sum_{i=1}^n N_i(t)\right\} \\ &= n^{-1/2}\sum_{i=1}^n (1-\Delta_i)\left(1-\frac{\xi_i}{\alpha}\right)\int_0^\tau R_i(\beta_0,t)d\left\{\frac{1}{n}\sum_{j=1}^n N_j(t)\right\}, \end{aligned}$$

where $R_i(\beta_0,t)$ denotes the integrand of the integration in the second last expression above. Define $\mathscr{F}(\tau)$ to be the sigma algebra generated by $Y_i(t)$, $N_i(t)$, and $Z_i(t)$ for $0 \le t \le \tau$ and $i = 1, ..., n$. Thus, conditional on $\mathscr{F}(\tau)$, the only random element is ξ_i, and $\mathrm{E}\{\xi|\mathscr{F}(\tau)\} = \alpha$. Given that $\mathrm{E}(I_1) = 0$ and $\mathrm{E}(I_2) = 0$, we have

$$\begin{aligned} \mathrm{cov}(I_1, I_2) &= n^{-1}c_n^{-1/2}\mathrm{E}\left(\mathrm{E}\left[u^T\sum_{i=1}^n\int_0^\tau\left\{Z_i(t) - \frac{S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}\right\}dN_i(t)\times \right.\right. \\ &\quad \left.\left.\sum_{i=1}^n(1-\Delta_i)\left(1-\frac{\xi_i}{\alpha}\right)\int_0^\tau R_i(\beta_0,t)d\left\{\frac{1}{n}\sum_{j=1}^n N_j(t)\right\}\middle|\mathscr{F}(\tau)\right]\right) \\ &= n^{-1}c_n^{-1/2}\mathrm{E}\left[u^T\sum_{i=1}^n\int_0^\tau\left\{Z_i(t) - \frac{S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}\right\}dN_i(t)\times \right. \\ &\quad \left.\sum_{i=1}^n(1-\Delta_i)\mathrm{E}\left(1-\frac{\xi_i}{\alpha}\middle|\mathscr{F}(\tau)\right)\int_0^\tau R_i(\beta_0,t)d\left\{\frac{1}{n}\sum_{j=1}^n N_j(t)\right\}\right] = 0. \end{aligned}$$

Therefore, $(nc_n)^{-1/2}u^T\tilde{\ell}'_n(\beta_0)$ converges to a zero mean normal distribution. Now define vector $u^* = u^T\Gamma_n^{-1/2}(\beta_0)\|u^T\Gamma_n^{-1/2}(\beta_0)\|^{-1}$. Then $\|u^*\| = 1$. Let $c_n^* = \|u^*\|_0$. The quantity $n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}'_n(\beta_0) = \|u^T\Gamma_n^{-1/2}(\beta_0)\|(c_n^*)^{1/2}(nc_n^*)^{-1/2}(u^*)^T\tilde{\ell}'_n(\beta_0)$, which converges to a zero mean normal distribution up to a scalar by the above result. Its variance $\text{var}\{n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}'_n(\beta_0)\} = u^T\Gamma_n^{-1/2}(\beta_0)\text{var}\{n^{-1/2}\tilde{\ell}'_n(\beta_0)\}\Gamma_n^{-1/2}(\beta_0)u = 1$, since $\Gamma_n(\beta_0) = \text{var}\{n^{-1/2}\tilde{\ell}'_n(\beta_0)\}$ and $\|u\| = 1$. Therefore, $n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}'_n(\beta_0)$ converges to a standard normal distribution. □

Lemma 3.7.6. *Under Conditions (A) to (D), $n^{-1/2}\{\tilde{\ell}''_n(\beta_0)_{jk} + nI_n(\beta_0)_{jk}\}$ is $O_p(1)$ for $j,k = 1,...,d_n$, where $I_n(\beta_0)_{jk}$ is the (j,k) component of $I_n(\beta_0)$ as defined in the Notations and Regularity Conditions section.*

Proof. The (j,k) component of the quadratic variation matrix of the partial score function under full cohort is $\langle\ell'_n(\beta_0)\rangle_{jk} = n\int_0^\tau V_{njk}(\beta_0,t)S_n^{(0)}(\beta_0,t)d\Lambda_0(t)$. We decompose $n^{-1/2}\{\tilde{\ell}''_n(\beta_0)_{jk} + nI_n(\beta_0)_{jk}\}$ as

$$
\begin{aligned}
&-n^{-1/2}\left\{\sum_{i=1}^n\int_0^\tau \tilde{V}_{njk}(\beta_0,t)dN_i(t) - \langle\ell'_n(\beta_0)\rangle_{jk}\right\} - n^{-1/2}\left\{\langle\ell'_n(\beta_0)\rangle_{jk} - nI_n(\beta_0)_{jk}\right\}\\
&= -n^{1/2}\int_0^\tau\left\{\tilde{V}_{njk}(\beta_0,t) - V_{njk}(\beta_0,t)\right\}\frac{1}{n}\sum_{i=1}^n dM_i(t) - n^{1/2}\int_0^\tau V_{njk}(\beta_0,t)\frac{1}{n}\sum_{i=1}^n dM_i(t)\\
&\quad - n^{1/2}\int_0^\tau\left\{\tilde{V}_{njk}(\beta_0,t) - V_{njk}(\beta_0,t)\right\}S_n^{(0)}(\beta_0,t)d\Lambda_0(t) - n^{1/2}\left\{\frac{1}{n}\langle\ell'_n(\beta_0)\rangle_{jk} - I_n(\beta_0)_{jk}\right\}\\
&= -I_1 - I_2 - I_3 - I_4.
\end{aligned}
$$

The integrand of I_1 can be further written as

$$
\begin{aligned}
&n^{1/2}\left\{\tilde{V}_{njk}(\beta_0,t) - V_{njk}(\beta_0,t)\right\}\\
&= n^{1/2}\left\{\frac{\tilde{S}_{njk}^{(2)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)} - \frac{S_{njk}^{(2)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}\right\} - n^{1/2}\left\{\frac{\tilde{S}_{nj}^{(1)}(\beta_0,t)\tilde{S}_{nk}^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)^2} - \frac{S_{nj}^{(1)}(\beta_0,t)S_{nk}^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)^2}\right\}\\
&= \frac{n^{1/2}\left\{\tilde{S}_{njk}^{(2)}(\beta_0,t) - S_{njk}^{(2)}(\beta_0,t)\right\}}{\tilde{S}_n^{(0)}(\beta_0,t)} - \frac{n^{1/2}\left\{\tilde{S}_n^{(0)}(\beta_0,t) - S_n^{(0)}(\beta_0,t)\right\}S_{njk}^{(2)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)S_n^{(0)}(\beta_0,t)}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{n^{1/2}\left\{\tilde{S}_{nj}^{(1)}(\beta_0,t) - S_{nj}^{(1)}(\beta_0,t)\right\}\tilde{S}_{nk}^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)^2} - \frac{n^{1/2}\left\{\tilde{S}_{nk}^{(1)}(\beta_0,t) - S_{nk}^{(1)}(\beta_0,t)\right\}S_{nj}^{(1)}(\beta_0,t)}{\tilde{S}_n^{(0)}(\beta_0,t)^2} \\
&+ \frac{n^{1/2}\left\{\tilde{S}_n^{(0)}(\beta_0,t) - S_n^{(0)}(\beta_0,t)\right\}\left\{\tilde{S}_n^{(0)}(\beta_0,t) + S_n^{(0)}(\beta_0,t)\right\}S_{nj}^{(1)}(\beta_0,t)S_{nk}^{(1)}(\beta_0,t)}{\left\{\tilde{S}_n^{(0)}(\beta_0,t)S_n^{(0)}(\beta_0,t)\right\}^2} \\
&= J_1 - J_2 - J_3 - J_4 + J_5.
\end{aligned}
$$

By Lemma 3.7.4 and Slutsky's theorem together with Condition (D) we have that J_1, J_2, J_3, J_4, and J_5 all converge to tight zero mean Gaussian processes. Therefore, $n^{1/2}\left\{\tilde{V}_{njk}(\beta_0,t) - V_{njk}(\beta_0,t)\right\}$ converges to a tight zero mean Gaussian process. It implies that $\tilde{V}_{njk}(\beta_0,t) - V_{njk}(\beta_0,t)$ converges to 0 in probability uniformly in t. It can also be shown that $\tilde{V}_{njk}(\beta_0,t)$ is of bounded variation. As shown in the proof of Lemma 3.7.5, $n^{-1/2}\sum_{i=1}^n dM_i(t)$ converges weakly to a zero mean Gaussian process with continuous sample paths and has bounded variation almost surely. It follows by Lemma 3.7.2 that I_1 converges to 0 in probability.

Since $V_{njk}(\beta_0,t)$ is a predictable process, I_2 is a locally square integrable martingale. To use martingale central limit theorem, we verify the two required conditions. Its quadratic variation process is

$$
\langle I_2\rangle = \int_0^\tau n^{-1}V_{njk}^2(\beta_0,t)\sum_i^n Y_i(t)e^{\beta_0^T Z_i}d\Lambda_0(t) = \int_0^\tau V_{njk}^2(\beta_0,t)S_n^{(0)}(\beta_0,t)d\Lambda_0(t). \tag{3.9}
$$

By Conditions (A) and (D), Lemma 3.7.4, and Slutsky's theorem, (3.9) converges to a finite value as n goes to infinity. Thus, the first condition is satisified. Next we verify the Lindeberg condition. For any $\epsilon > 0$,

$$
\begin{aligned}
&\int_0^\tau n^{-1}V_{njk}^2(\beta_0,t)I\left\{\left|n^{-1/2}V_{njk}^2(\beta_0,t)\right| > \epsilon\right\}\sum_i^n Y_i(t)e^{\beta_0^T Z_i}d\Lambda_0(t) \\
&= \int_0^\tau V_{njk}^2(\beta_0,t)I\left\{\left|n^{-1/2}V_{njk}^2(\beta_0,t)\right| > \epsilon\right\}S_n^{(0)}(\beta_0,t)d\Lambda_0(t).
\end{aligned} \tag{3.10}
$$

By Condition (D), Lemma 3.7.4, and Slutsky's theorem, $V_{njk}^2(\beta_0,t)$ converges uniformly to a bounded process, thus $I\left\{\left|n^{-1/2}V_{njk}^2(\beta_0,t)\right|>\epsilon\right\}$ converges to 0. Then by Conditions (A) and (D) and Lemma 3.7.4, (3.10) converges to 0 as n goes to infinity. Therefore, the Lindeberg condition is satisfied. By martingale central limit theorem, I_2 converges to a zero mean normal distribution.

$$\begin{aligned}
I_3 &= \int_0^\tau n^{1/2}\left\{\tilde{V}_{njk}(\beta_0,t)-V_{njk}(\beta_0,t)\right\}S_n^{(0)}(\beta_0,t)d\Lambda_0(t)\\
&\le \int_0^\tau \sup_{t\in[0,\tau]}\left|n^{1/2}\left\{\tilde{V}_{njk}(\beta_0,t)-V_{njk}(\beta_0,t)\right\}S_n^{(0)}(\beta_0,t)\right|d\Lambda_0(t)\\
&= \sup_{t\in[0,\tau]}\left|n^{1/2}\left\{\tilde{V}_{njk}(\beta_0,t)-V_{njk}(\beta_0,t)\right\}S_n^{(0)}(\beta_0,t)\right|\{\Lambda_0(\tau)-\Lambda_0(0)\}=O_p(1).
\end{aligned}$$

The last equality holds because $n^{1/2}\{\tilde{V}_{njk}(\beta_0,t)-V_{njk}(\beta_0,t)\}$ converges to a tight zero mean Gaussian process and $S_n^{(0)}(\beta_0,t)$ converges uniformly to $s_n^{(0)}(\beta_0,t)$ which is bounded away from 0.

We now consider I_4. By Chebyshev inequality, for any $\varepsilon>0$ and any sequence $\gamma_n\to\infty$,

$$\begin{aligned}
\mathrm{pr}\left\{\left|\frac{1}{n}\langle\ell_n'(\beta_0)\rangle_{jk}-I_n(\beta_0)_{jk}\right|\ge\varepsilon\gamma_n n^{-1/2}\right\} &\le \frac{n\mathrm{E}\left\{n^{-1}\langle\ell_n'(\beta_0)\rangle_{jk}-I_n(\beta_0)_{jk}\right\}^2}{\varepsilon^2\gamma_n^2}\\
&= \frac{\mathrm{var}\{\langle\ell_n'(\beta_0)\rangle_{jk}\}}{n\varepsilon^2\gamma_n^2}. \qquad (3.11)
\end{aligned}$$

Let $\nu_{njk}(\beta_0,t)=\left\{s_{njk}^{(2)}(\beta_0,t)s_n^{(0)}(\beta_0,t)-s_{nj}^{(1)}(\beta_0,t)s_{nk}^{(1)}(\beta_0,t)\right\}/s_n^{(0)}(\beta_0,t)^2$. Then

$$\begin{aligned}
&n^{1/2}\left\{\frac{1}{n}\langle\ell_n'(\beta_0)\rangle_{jk}-\int_0^\tau\nu_{njk}(\beta_0,t)s_n^{(0)}(\beta_0,t)d\Lambda_0(t)\right\}\\
&=\int_0^\tau\left[n^{1/2}\left\{S_{njk}^{(2)}(\beta_0,t)-s_{njk}^{(2)}(\beta_0,t)\right\}-\frac{S_{nj}^{(1)}(\beta_0,t)n^{1/2}\{S_{nk}^{(1)}(\beta_0,t)-s_{nk}^{(1)}(\beta_0,t)\}}{S_n^{(0)}(\beta_0,t)}\right.\\
&\quad-\frac{s_{nk}^{(1)}(\beta_0,t)n^{1/2}\{S_{nj}^{(1)}(\beta_0,t)-s_{nj}^{(1)}(\beta_0,t)\}}{S_n^{(0)}(\beta_0,t)}\\
&\quad\left.+\frac{s_{nj}^{(1)}(\beta_0,t)s_{nk}^{(1)}(\beta_0,t)n^{1/2}\{S_n^{(0)}(\beta_0,t)-s_n^{(0)}(\beta_0,t)\}}{S_n^{(0)}(\beta_0,t)s_n^{(0)}(\beta_0,t)}\right]d\Lambda_0(t). \qquad (3.12)
\end{aligned}$$

Denote the integrand of (3.12) as $H_{njk}(\beta_0, t)$. By the weak convergence of $S_n^{(0)}(\beta_0, t)$, $S_{nj}^{(1)}(\beta_0, t)$, and $S_{njk}^{(2)}(\beta_0, t)$ to respectively $s_n^{(0)}(\beta_0, t)$, $s_{nj}^{(1)}(\beta_0, t)$, and $s_{njk}^{(2)}(\beta_0, t)$ $(j, k = 1, ..., d_n)$ and Slutsky's theorem, $H_{njk}(\beta_0, t)$ converges to a tight zero mean Gaussian process. Therefore,

$$\big|(3.12)\big| \leq \int_0^\tau \sup_{t\in[0,\tau]} \big|H_{njk}(\beta_0, t)\big| d\Lambda_0(t) = \sup_{t\in[0,\tau]} \big|H_{njk}(\beta_0, t)\big| \{\Lambda_0(\tau) - \Lambda_0(0)\} = O_p(1).$$

By Conditions (A) to (D), the variable $n^{-1}\langle \ell_n'(\beta_0)\rangle_{jk} = \int_0^\tau V_{njk}(\beta_0, t) S_n^{(0)}(\beta_0, t) d\Lambda_0(t)$ is bounded, and therefore its first and second moment exist. Using the fact that $|(3.12)| = O_p(1)$, it follows that $\mathrm{var}\left[n^{1/2}\left\{n^{-1}\langle \ell_n'(\beta_0)\rangle_{jk} - \int_0^\tau \nu_{njk}(\beta_0, t) s_n^{(0)}(\beta_0, t) d\Lambda_0(t)\right\}\right] = O(1)$. With some algebra and the fact that $\int_0^\tau \nu_{njk}(\beta_0, t) s_n^{(0)}(\beta_0, t) d\Lambda_0(t)$ is a constant, we have that $\mathrm{var}\{\langle \ell_n'(\beta_0)\rangle_{jk}\} = O(n)$. It follows that (3.11) is $o(1)$. Therefore, $n^{-1}\langle \ell_n'(\beta_0)\rangle_{jk} - I_n(\beta_0)_{jk} = O_p(n^{-1/2})$ and $I_4 = O_p(1)$.

Taking all results together, we have shown that $n^{-1/2}\{\tilde{\ell}_n''(\beta_0)_{jk} + nI_n(\beta_0)_{jk}\}$ is $O_p(1)$ for $j, k = 1, ..., d_n$. □

Proof of Theorem 3.3.1. Let β_0 be the true parameters, and $\alpha_n = d_n^{1/2}(n^{-1/2} + a_n)$. It suffices to show that, for any $\varepsilon > 0$ and any constant vector u with $\|u\| = C$, there exists a large enough C such that $\mathrm{pr}\{\sup_{\|u\|=C} \tilde{Q}_n(\beta_0 + \alpha_n u) < \tilde{Q}_n(\beta_0)\} \geq 1 - \varepsilon$. This implies that there exists a local maximizer $\hat{\beta}$ such that $\|\hat{\beta} - \beta_0\| = O_p(\alpha_n)$. Since $P_{\lambda_{jn}}(0) = 0$ and $P_{\lambda_{jn}}(\cdot) \geq 0$, we have

$$\begin{aligned}\tilde{Q}_n(\beta_0 + \alpha_n u) - \tilde{Q}_n(\beta_0) &\leq \{\tilde{\ell}_n(\beta_0 + \alpha_n u) - \tilde{\ell}_n(\beta_0)\} - n\sum_{j=1}^{k_n}\{P_{\lambda_{jn}}(|\beta_{j0} + \alpha_n u_j|) - P_{\lambda_{jn}}(|\beta_{j0}|)\} \\ &= I_1 + I_2.\end{aligned}$$

We first consider I_1. By Taylor expansion we have

$$I_1 = \alpha_n u^T \tilde{\ell}_n'(\beta_0) + \frac{1}{2}\alpha_n^2 u^T \tilde{\ell}_n''(\beta_0)u + \frac{1}{6}\alpha_n^3 \sum_{i=1}^{n}\sum_{j,k,l=1}^{d_n} \tilde{\ell}_i'''(\beta^*)_{jkl}u_j u_k u_l = I_{11} + I_{12} + I_{13},$$

where β^* lies between β_0 and $\beta_0 + \alpha_n u$.

From Lemma 3.7.5 we have $\tilde{\ell}_n'(\beta_0)_j = O_p(n^{1/2})$ for $j = 1, ..., d_n$. Therefore,

$$|I_{11}| \le \alpha_n \|u\| \|\tilde{\ell}_n'(\beta_0)\| = \alpha_n \|u\| O_p\{(d_n n)^{1/2}\} = \|u\| O_p(d_n^{1/2} n^{-1/2} \alpha_n n) = \|u\| O_p(\alpha_n^2 n).$$

The term I_{12} can be written as $\alpha_n^2 u^T\{\tilde{\ell}_n''(\beta_0) + nI_n(\beta_0)\}u/2 - \alpha_n^2 u^T n I_n(\beta_0)u/2 = J_1 - J_2$. By Cauchy-Schwarz inequality and $\{\tilde{\ell}_n''(\beta_0)_{jk} + nI_n(\beta_0)_{jk}\} = O_p(n^{1/2})$ for $j, k = 1, ..., d_n$, and Lemma 3.7.6, we have $|J_1| \le \alpha_n^2 \|u\|^2 \|\tilde{\ell}_n''(\beta_0) + nI_n(\beta_0)\|/2 = \|u\|^2 O_p(\alpha_n^2 n^{1/2} d_n) = \|u\|^2 o_p(\alpha_n^2 n)$. By spectral decomposition of $I_n(\beta_0)$ and Condition (F) we have that $|J_2| \ge \alpha_n^2 \|u\|^2 n \text{eigen}_{\min}\{I_n(\beta_0)\}/2 \ge \|u\|^2(\alpha_n^2 n)C_3/2$. Under Conditions (A) to (D), $\partial \tilde{V}_{njk}(\beta^*, t)/\partial\beta_l$ is of bounded variation in t for $i = 1, ..., n$, $j, k, l = 1, ..., d_n$. Therefore $\tilde{\ell}_i'''(\beta^*)_{jkl} = -\int_0^\tau \partial \tilde{V}_{njk}(\beta^*, t)/\partial\beta_l dN_i(t)$ is $O_p(1)$. Along 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 have $|I_{13}| = O_p(d_n^{3/2}) n\alpha_n^3 \|u\|^3 = O_p\{d_n^2(n^{-1/2} + a_n)\} n\alpha_n^2 \|u\|^3 = O_p(d_n^2 n^{-1/2} + d_n^2 a_n) n\alpha_n^2 \|u\|^3 = \|u\|^3 o_p(\alpha_n^2 n)$. Therefore, for large enough $\|u\|$, $|J_2|$ dominates $|I_{11}|$, $|J_1|$, and $|I_{13}|$.

We now consider I_2. By Taylor expansion and Cauchy-Schwarz inequality

$$\begin{aligned}
|I_2| &= \left| n\sum_{j=1}^{k_n} P'_{\lambda_{jn}}(|\beta_{j0}|)\text{sgn}(\beta_{j0})\alpha_n u_j + \frac{1}{2} n \sum_{j=1}^{k_n} P''_{\lambda_{jn}}(|\beta_{j0}|)\alpha_n^2 u_j^2\{1 + o(1)\} \right| \\
&\le n\left| \sum_{j=1}^{k_n} P'_{\lambda_{jn}}(|\beta_{j0}|)\alpha_n u_j \right| + \frac{1}{2} n \left| \sum_{j=1}^{k_n} P''_{\lambda_{jn}}(|\beta_{j0}|)\alpha_n^2 u_j^2\{1 + o(1)\} \right| \\
&\le n\alpha_n a_n k_n^{1/2} \|u\| + \frac{1}{2} n\alpha_n^2 b_n \|u\|^2\{1 + o(1)\} \\
&= \|u\| O_p(\alpha_n^2 n).
\end{aligned}$$

The last equality holds because $a_n = O_p(\alpha_n d_n^{-1/2})$ and $b_n \to 0$ under Condition (G). Therefore, $|J_2|$ dominates $|I_2|$ for large enough C. Since J_2 is negative, it follows that for large enough C, $\tilde{Q}_n(\beta_0 + \alpha_n u) - \tilde{Q}_n(\beta_0)$ is negative with probability tending to one as $n \to \infty$. □

Lemma 3.7.7. *Under Conditions (A) to (G), if $d_n^4/n \to 0, \lambda_{jn} \to 0$, and $\lambda_{jn} n^{1/2} d_n^{-1/2} \to \infty$, then with probability tending to one, for any β_I satisfying $\|\beta_I - \beta_{I0}\| = O(d_n^{1/2} n^{-1/2})$ and any constant C, we have $\tilde{Q}_n\{(\beta_I^T, 0^T)^T\} = \max_{\|\beta_{II}\| \le C d_n^{1/2} n^{-1/2}} \tilde{Q}_n\{(\beta_I^T, \beta_{II}^T)^T\}$.*

Proof. It suffices to show that with probability tending to one, for any β_I satisfying $\|\beta_I - \beta_{I0}\| = O(d_n^{1/2} n^{-1/2})$ and $\|\beta_{II}\| \le C d_n^{1/2} n^{-1/2}$, $\partial \tilde{Q}_n(\beta)/\partial \beta_j$ and β_j have different signs for $j = (k_n + 1), ..., d_n$. By Taylor expansion,

$$\begin{aligned}\frac{\partial \tilde{Q}_n(\beta)}{\partial \beta_j} &= \tilde{\ell}_n'(\beta_0)_j + \sum_{k=1}^{d_n} \tilde{\ell}_n''(\beta_0)_{jk}(\beta_k - \beta_{0k}) + \sum_{k,l=1}^{d_n} \tilde{\ell}_n'''(\beta^*)_{jkl}(\beta_k - \beta_{0k})(\beta_l - \beta_{0l}) \\ &\quad - n P_{\lambda_{jn}}'(|\beta_j|)\mathrm{sgn}(\beta_j) \\ &= I_1 + I_2 + I_3 + I_4,\end{aligned}$$

where β^* lies between β_0 and β. From Lemma 3.7.5 we have $I_1 = O_p(n^{1/2}) = o_p(d_n^{1/2} n^{1/2})$.

$$I_2 = \sum_{k=1}^{d_n} \left\{\tilde{\ell}_n''(\beta_0)_{jk} + n I_n(\beta_0)_{jk}\right\}(\beta_k - \beta_{0k}) - \sum_{k=1}^{d_n} n I_n(\beta_0)_{jk}(\beta_k - \beta_{0k}) = I_{21} - I_{22}.$$

From Lemma 3.7.6 we have $\tilde{\ell}_n''(\beta_0)_{jk} + n I_n(\beta_0)_{jk} = O_p(n^{1/2})$ for $j, k = 1, ..., d_n$. Using Cauchy-Schwarz inequality along with $\|\beta - \beta_0\| = O_p(d_n^{1/2} n^{-1/2})$,

$$|I_{21}| \le \|\beta - \beta_0\| \left\{\sum_{k=1}^{d_n} \{\tilde{\ell}_n''(\beta_0)_{jk} + n I_n(\beta_0)_{jk}\}^2\right\}^{1/2} = O_p(d_n) = o_p(d_n^{1/2} n^{1/2}).$$

As $\mathrm{eigen}_{\max}\{I_n(\beta_0)\}$ is bounded by Condition (F), it follows that

$$|I_{22}| \le n\|\beta - \beta_0\| \left\{\sum_{k=1}^{d_n} I_n^2(\beta_0)_{jk}\right\}^{1/2} = n O_p(d_n^{1/2} n^{-1/2}) O(1) = O_p(d_n^{1/2} n^{1/2}).$$

It follows that $|I_2| = O_p(d_n^{1/2} n^{1/2})$. By Cauchy-Schwarz inequality,

$$|I_3| = \left| \sum_{i=1}^{n} \sum_{k,l=1}^{d_n} \tilde{\ell}_i'''(\beta^*)_{jkl} (\beta_k - \beta_{0k})(\beta_l - \beta_{0l}) \right| \leq \|\beta - \beta_0\|^2 \sum_{i=1}^{n} \Big(\sum_{k,l=1}^{d_n} \tilde{\ell}_i'''(\beta^*)_{jkl}^2 \Big)^{1/2}.$$

As shown in the proof of Theorem 3.3.1, $\tilde{\ell}_i'''(\beta^*)_{jkl} = O_p(1)$. Therefore, we have that $\{\sum_{k,l=1}^{d_n} \tilde{\ell}_i'''(\beta^*)_{jkl}^2\}^{1/2} = O_p(d_n)$ and $|I_3| = O_p\{(d_n/n) n d_n\} = O_p(d_n^2) = O_p(d_n^{1/2} n^{1/2})$, and therefore $I_1 + I_2 + I_3 = O_p(d_n^{1/2} n^{1/2})$. Hence,

$$\begin{aligned} \frac{\partial \tilde{Q}_n(\beta)}{\partial \beta_j} &= -n P'_{\lambda_{jn}}(|\beta_j|) \mathrm{sgn}(\beta_j) + O_p(d_n^{1/2} n^{1/2}) \\ &= n \lambda_{jn} \left\{ -\frac{P'_{\lambda_{jn}}(|\beta_j|)}{\lambda_{jn}} \mathrm{sgn}(\beta_j) + O_p\left(\frac{d_n^{1/2} n^{-1/2}}{\lambda_{jn}} \right) \right\}. \end{aligned}$$

For $j = (k_n + 1), ..., d_n$, since $|\beta_j| = O\{(d_n/n)^{1/2}\}$ and $\lambda_{jn}(n/d_n)^{1/2} \to \infty$, the quantity $P'_{\lambda_{jn}}(|\beta_j|)/\lambda_{jn}$ is positive under Condition (H) for all sufficiently large n. Therefore, the quantity in the curly brackets is negative with probability tending to one. Thus, $\partial \tilde{Q}_n(\beta)/\partial \beta_j$ and β_j have different signs with probability tending to one as $n \to \infty$. □

Proof of Theorem 3.3.2. The assertion that $\hat{\beta}_{II}^T = 0$ with probability tending to one as $n \to \infty$ follows directly from Lemma 3.7.7. To prove the second assertion, we first show that

$$\begin{aligned} & n^{1/2} u^T \Gamma_{n11}^{-1/2} (I_{n11} + \Sigma_n)(\hat{\beta}_I - \beta_{I0})(1 + o_p(1)) + n^{1/2} u^T \Gamma_{n11}^{-1/2} B_n \\ &= n^{-1/2} u^T \Gamma_{n11}^{-1/2} \tilde{\ell}'_{n1}(\beta_0) + o_p(1), \end{aligned} \tag{3.13}$$

where $\tilde{\ell}'_{n1}(\beta_0)$ consists of the first k_n components of $\tilde{\ell}'_n(\beta_0)$. Since $\hat{\beta}_I$ is the maximum penalized pseudo-partial likelihood estimator, $\partial \tilde{Q}_n(\hat{\beta})/\partial \beta_I = 0$. By Taylor expansion of $\partial \tilde{Q}_n(\hat{\beta})/\partial \beta_I$ at β_{I0} and the fact that $\hat{\beta}_{II} - \beta_{II0} = 0$ with probability tending to one, we have $\tilde{\ell}'_{n1}(\beta_0) + \tilde{\ell}''_{n1}(\beta_0)(\hat{\beta}_I - \beta_{I0}) + (\hat{\beta}_I - \beta_{I0})^T \tilde{\ell}'''_{n1}(\beta^*)(\hat{\beta}_I - \beta_{I0})/2 - n B_n - n \Sigma_n^{**}(\hat{\beta}_I - \beta_{I0}) = 0$ with

probability tending to one, where $\tilde{\ell}''_{n1}(\beta_0)$ consists of the first $k_n \times k_n$ components of $\tilde{\ell}''_n(\beta_0)$, $\tilde{\ell}'''_{n1}(\beta^*)$ consists of the first $k_n \times k_n \times k_n$ components of $\tilde{\ell}'''_n(\beta^*)$, β^* lies between $\hat{\beta}$ and β_0, $\Sigma_n^{**} = \Sigma_n(\beta^{**})$, β^{**} lies between $\hat{\beta}$ and β_0. After rearranging the above equation we have for all large n,

$$\{\tilde{\ell}''_{n1}(\beta_0) - n\Sigma_n^{**}\}(\hat{\beta}_I - \beta_{I0}) - nB_n = -\tilde{\ell}'_{n1}(\beta_0) - \frac{1}{2}(\hat{\beta}_I - \beta_{I0})^T \tilde{\ell}'''_{n1}(\beta^*)(\hat{\beta}_I - \beta_{I0}). \tag{3.14}$$

Denote $\nu_n = (\hat{\beta}_I - \beta_{I0})^T \tilde{\ell}'''_{n1}(\beta^*)(\hat{\beta}_I - \beta_{I0})$. Multiple both sides of (3.14) by $n^{-1/2}u^T\Gamma_{n11}^{-1/2}$,

$$\begin{aligned} & n^{1/2}u^T\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde{\ell}''_{n1}(\beta_0) - \Sigma_n^{**}\right\}(\hat{\beta}_I - \beta_{I0}) - n^{1/2}u^T\Gamma_{n11}^{-1/2}B_n \\ & = -n^{-1/2}u^T\Gamma_{n11}^{-1/2}\tilde{\ell}'_{n1}(\beta_0) - n^{-1/2}u^T\Gamma_{n11}^{-1/2}\nu_n/2. \end{aligned} \tag{3.15}$$

By Cauchy-Schwarz inequality, $\|\nu_n\| \le \|\hat{\beta}_I - \beta_{I0}\|^2 \sum_{i=1}^n \{\sum_{j,k,l=1}^{k_n} \tilde{\ell}'''_{i1}(\beta^*)_{jkl}^2\}^{1/2}$. As shown in the proof of Theorem 1, $\tilde{\ell}'''_{i1}(\beta^*)_{jkl} = O_p(1)$. Therefore, $\|\nu_n\| = O_p\{(d_n/n)nk_n^{3/2}\} = O_p(d_n^{5/2})$. By spectral decomposition of $\Gamma_{n11}^{-1/2}$, $d_n^5/n \to 0$, and Condition 6,

$$\begin{aligned} \frac{1}{2}n^{-1/2}u^T\Gamma_{n11}^{-1/2}\nu_n &\le \frac{\|u\|\|\nu_n\|}{2}n^{-1/2}\text{eigen}_{\max}(\Gamma_{n11}^{-1/2}) \le \frac{\|u\|\|\nu_n\|}{2}n^{-1/2}\text{eigen}_{\max}(\Gamma_n^{-1/2}) \\ &= O_p(d_n^{5/2}n^{-1/2}) = o_p(1). \end{aligned} \tag{3.16}$$

The second inequality in (3.16) holds by interlacing inequality of symmetric matrix. Meanwhile, $u^T\Gamma_{n11}^{-1/2}n^{-1}\tilde{\ell}''_{n1}(\beta_0)(\hat{\beta}_I - \beta_{I0}) = u^T\Gamma_{n11}^{-1/2}\{n^{-1}\tilde{\ell}''_{n1}(\beta_0) + I_{n11}(\beta_0)\}(\hat{\beta}_I - \beta_{I0}) - u^T\Gamma_{n11}^{-1/2}I_{n11}(\beta_0)(\hat{\beta}_I - \beta_{I0}) = J_1 - J_2$. By Cauchy-Schwarz inequality and Lemma 3.7.6, we have $|J_1| \le \|u^T\Gamma_{n11}^{-1/2}\|\|n^{-1}\tilde{\ell}''_{n1}(\beta_0) + I_{n11}(\beta_0)\|\|\hat{\beta}_I - \beta_{I0}\| = \|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|O_p(d_n n^{-1/2})$. By spectral decomposition of I_{n11}, we have $|J_2| \ge \|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|\text{eigen}_{\min}(I_{n11}) \ge \|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|\text{eigen}_{\min}(I_n)$. Therefore, by Condition 6 we have

$$\left|\frac{J_1}{J_2}\right| \le \frac{\|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|O_p(d_n n^{-1/2})}{\|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|\text{eigen}_{\min}(I_n)} = O_p(d_n n^{-1/2}) = o_p(1).$$

Therefore, $J_1 = o_p(J_2)$, and $u^T\Gamma_{n11}^{-1/2} n^{-1}\tilde{\ell}''_{n1}(\beta_0)(\hat{\beta}_I - \beta_{I0}) = -u^T\Gamma_{n11}^{-1/2} I_{n11}(\beta_0)(\hat{\beta}_I - \beta_{I0})\{1 + o_p(1)\}$. Since $\hat{\beta}$ converges to β_0 in probability, it follows that

$$u^T\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde{\ell}''_{n1}(\beta_0) - \Sigma_n^{**}\right\}(\hat{\beta}_I - \beta_{I0}) = -u^T\Gamma_{n11}^{-1/2}\left\{I_{n11}(\beta_0) + \Sigma_n\right\}(\hat{\beta}_I - \beta_{I0})\{1 + o_p(1)\}. \tag{3.17}$$

By (3.15), (3.16), (3.17), we know (3.13) holds. By Lemma 3.7.5, $n^{-1/2}u^T\Gamma_{n11}^{-1/2}\tilde{\ell}'_{n1}(\beta_0)$ converges to the standard normal distribution. Therefore,

$$n^{1/2}u^T\Gamma_{n11}^{-1/2}(I_{n11} + \Sigma_n)\{\hat{\beta}_I - \beta_{I0} + (I_{n11} + \Sigma_n)^{-1}B_n\} \to N(0,1)$$

in distribution. □

Derivation of $\hat{\Gamma}_n(\hat{\beta})$. As defined in Section 3.3.2, $\Gamma_n(\beta_0) = n^{-1}\mathrm{var}\{\tilde{\ell}'_n(\beta_0)\}$. We first derive its asymptotic expression. Since the dimension of $\tilde{\ell}'_n(\beta_0)$ goes to infinity, it is only meaningful to consider the variance of its linear combination $(nc_n)^{-1/2}u^T\tilde{\ell}'_n(\beta_0)$, where u is an arbitrary constant vector with $\|u\| = 1$, and $\|u\|_0 = c_n$. Under this setting, $\mathrm{var}\{(nc_n)^{-1/2}u^T\tilde{\ell}'_n(\beta_0)\} = c_n^{-1}u^T\Gamma_n(\beta_0)u$. Let the limit of $c_n^{-1}u^T\Gamma_n(\beta_0)u$ be $\Gamma(\beta_0)$.

As shown in the proof of Lemma 3.7.5, $(nc_n)^{-1/2}u^T\tilde{\ell}'_n(\beta_0)$ is asymptotically equivalent to

$$\begin{aligned}
&(nc_n)^{-1/2}u^T\sum_{i=1}^n\int_0^\tau\left\{Z_i(t) - \frac{S_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)}\right\}dN_i(t) + n^{-1/2}\sum_{i=1}^n\left(1 - \frac{\xi_i}{\alpha}\right)\times\\
&\int_0^\tau(1-\Delta_i)Y_i(t)\left(\left[\frac{A_n^{(0)}(\beta_0,t)}{\mathrm{E}\{(1-\Delta)Y(t)\}} - e^{\beta_0^TZ_i(t)}\right]\frac{c_n^{-1/2}u^T\tilde{S}_n^{(1)}(\beta_0,t)}{S_n^{(0)}(\beta_0,t)\tilde{S}_n^{(0)}(\beta_0,t)}\right.\\
&\left.-\left[\frac{A_n^{(1)}(\beta_0,t)}{\mathrm{E}\{(1-\Delta)Y(t)\}} - e^{\beta_0^TZ_i(t)}c_n^{-1/2}u^TZ_i(t)\right]\frac{1}{S_n^{(0)}(\beta_0,t)}\right)d\left\{\frac{1}{n}\sum_{i=1}^nN_i(t)\right\},
\end{aligned} \tag{3.18}$$

where $A_n^{(k)}(\beta_0,t) = n^{-1}\sum_{i=1}^n(1-\Delta_i)\xi_i\alpha^{-1}Y_i(t)e^{\beta_0^TZ_i(t)}\{c_n^{-1/2}u^TZ_i(t)\}^k$ $(k = 0,1)$. Consider the quantity $c_n^{-1/2}u^TZ_i(t)$ in $A_n^{(k)}(\beta_0,t)$. By Condition (C) and $\|u\| = 1$, it is a bounded

deterministic process for each sample path $Z_i(t)$ and all n. Assume $c_n^{-1/2}u^T Z_i(t)$ converges to $L\{u, Z_i(t)\}$ as $n \to \infty$. Then by Lemma 3.7.1, $A_n^{(k)}(\beta_0, t)$ is asymptotically equivalent to $\mathrm{E}[(1-\Delta)Y(t)e^{\beta_0^T Z_i(t)}L\{u, Z_i(t)\}^k]$. Therefore, (3.18) is asymptotically equivalent to

$$
\begin{aligned}
&(nc_n)^{-1/2}u^T \sum_{i=1}^{n} \int_0^\tau \left\{ Z_i(t) - \frac{S_n^{(1)}(\beta_0, t)}{S_n^{(0)}(\beta_0, t)} \right\} dN_i(t) \\
&+ n^{-1/2} \sum_{i=1}^{n} \left(1 - \frac{\xi_i}{\alpha}\right) \int_0^\tau (1-\Delta_i) Y_i(t) \left\{ \left[\frac{\mathrm{E}\{(1-\Delta)Y(t)e^{\beta_0^T Z_i(t)}\}}{\mathrm{E}\{(1-\Delta)Y(t)\}} \right.\right. \\
&\left.\left. - e^{\beta_0^T Z_i(t)} \right] \frac{c_n^{-1/2} u^T \tilde{S}_n^{(1)}(\beta_0, t)}{S_n^{(0)}(\beta_0, t)\tilde{S}_n^{(0)}(\beta_0, t)} - \left(\frac{\mathrm{E}[(1-\Delta)Y(t)e^{\beta_0^T Z_i(t)}L\{u, Z_i(t)\}]}{\mathrm{E}\{(1-\Delta)Y(t)\}} \right.\right. \\
&\left.\left. - e^{\beta_0^T Z_i(t)} c_n^{-1/2} u^T Z_i(t) \right) \frac{1}{S_n^{(0)}(\beta_0, t)} \right\} d\left\{ \frac{1}{n} \sum_{i=1}^{n} N_i(t) \right\} \\
&= I_1 + I_2.
\end{aligned}
$$

The quantity I_1 is a linear combination of the partial likelihood score vector of the full cohort data. Let $s^{(0)}(\beta_0, t)$, $s^{(1)}(\beta_0, t)$, and $s^{(2)}(\beta_0, t)$ be the limit of $S_n^{(0)}(\beta_0, t)$, $c_n^{-1/2}u^T S_n^{(1)}(\beta_0, t)$, and $c_n^{-1}u^T S_n^{(2)}(\beta_0, t)u$ respectively as $n \to \infty$. By Andersen and Gill (1982), the asymptotic variance of I_1 is

$$
\mathscr{I}_1(\beta_0) = \int_0^\tau \frac{s^{(2)}(\beta_0, t)s^{(0)}(\beta_0, t) - \{s^{(1)}(\beta_0, t)\}^2}{s^{(0)}(\beta_0, t)} d\Lambda_0(t).
$$

Let $W_i(\beta_0)$ be the integration in I_2, which equals $n^{-1/2}\sum_{i=1}^n (1 - \xi_i/\alpha) W_i(\beta_0)$.

Define $\mathscr{F}(\tau)$ as the sigma algebra generated by $Y_i(t)$, $N_i(t)$, and $Z_i(t)$ for $0 \le t \le \tau$ and $i = 1, ..., n$. Conditional on $\mathscr{F}(\tau)$, the only random element in I_2 is ξ. Since $\mathrm{E}\{\xi | \mathscr{F}(\tau)\} = \alpha$, the asymptotic variance of I_2, denoted by $\mathscr{I}_2(\beta_0)$, can be derived as

$$
\begin{aligned}
\mathscr{I}_2(\beta_0) &= \frac{1}{n}\mathrm{E}\left[\mathrm{var}\left\{ \sum_{i=1}^{n} \left(1 - \frac{\xi_i}{\alpha}\right) W_i(\beta_0) \middle| \mathscr{F}(\tau) \right\} \right] + \frac{1}{n}\mathrm{var}\left[\mathrm{E}\left\{ \sum_{i=1}^{n} \left(1 - \frac{\xi_i}{\alpha}\right) W_i(\beta_0) \middle| \mathscr{F}(\tau) \right\} \right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} \frac{\mathrm{var}\{\xi_i | \mathscr{F}(\tau)\}}{\alpha^2} W_i^2(\beta_0) \right] + \frac{1}{n}\mathrm{var}\left(\sum_{i=1}^{n} \left[1 - \frac{\mathrm{E}\{\xi_i | \mathscr{F}(\tau)\}}{\alpha} \right] W_i(\beta_0) \right)
\end{aligned}
$$

$$= \frac{1-\alpha}{\alpha}\mathrm{E}\{W^2(\beta_0)\}.$$

Finally, since I_1 and I_2 are independent as shown in the proof of Lemma 3.7.5, the asymptotic variance of $(nc_n)^{-1/2}u^T\tilde{\ell}_n'(\beta_0)$ is

$$\begin{aligned}\Gamma(\beta_0) &= \mathscr{I}_1(\beta_0) + \mathscr{I}_2(\beta_0) \\ &= \int_0^\tau \frac{s^{(2)}(\beta_0,t)s^{(0)}(\beta_0,t) - \{s^{(1)}(\beta_0,t)\}^2}{s^{(0)}(\beta_0,t)} d\Lambda_0(t) + \frac{1-\alpha}{\alpha}\mathrm{E}\{W^2(\beta_0)\}.\end{aligned}$$

Under finite sample, the matrix $\Gamma_n(\beta_0)$ has finite dimension and is therefore well defined. Then it can be estimated by estimating $\mathscr{I}_1(\beta_0)$ and $\mathscr{I}_2(\beta_0)$ without linear combination. Thus,

$$\begin{aligned}\hat{\Gamma}_n(\hat{\beta}) &= \hat{\mathscr{I}}_{n1}(\hat{\beta}) + \hat{\mathscr{I}}_{n2}(\hat{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\tilde{S}_n^{(2)}(\hat{\beta},t_i)\tilde{S}_n^{(0)}(\hat{\beta},t_i) - \{\tilde{S}_n^{(1)}(\hat{\beta},t_i)\}^{\otimes 2}}{\{\tilde{S}_n^{(0)}(\hat{\beta},t_i)\}^2}\Delta_i + \\ &\quad \frac{1-\alpha}{\alpha}\frac{1}{n}\sum_{i=1}^n \frac{\xi_i}{\alpha}\left\{\frac{1}{n}(1-\Delta_i)\sum_{j=1}^n \Delta_j Y_i(t_j)\left(\left[\frac{\hat{\mathrm{E}}\{(1-\Delta)Y(t_j)e^{\hat{\beta}^T Z_i(t_j)}\}}{\hat{\mathrm{E}}\{(1-\Delta)Y(t_j)\}} - e^{\hat{\beta}^T Z_i(t_j)}\right]\times\right.\right. \\ &\quad \left.\left.\frac{\tilde{S}_n^{(1)}(\hat{\beta},t_j)}{\{\tilde{S}_n^{(0)}(\hat{\beta},t_j)\}^2} - \left[\frac{\hat{\mathrm{E}}\{(1-\Delta)Y(t_j)e^{\hat{\beta}^T Z_i(t_j)}Z_i(t_j)\}}{\hat{\mathrm{E}}\{(1-\Delta)Y(t_j)\}} - e^{\hat{\beta}^T Z_i(t_j)}Z_i(t_j)\right]\frac{1}{\tilde{S}_n^{(0)}(\hat{\beta},t_j)}\right)\right\}^2,\end{aligned}$$

where

$$\begin{aligned}\hat{\mathrm{E}}\{(1-\Delta)Y(t)\} &= \frac{1}{n}\sum_{i=1}^n (1-\Delta_i)Y_i(t), \\ \hat{\mathrm{E}}\left\{(1-\Delta)Y(t_j)e^{\hat{\beta}^T Z_i(t_j)}\right\} &= \frac{1}{n}\sum_{i=1}^n \frac{\xi_i}{\hat{\alpha}(t)}(1-\Delta_i)Y_i(t)e^{\hat{\beta}^T Z_i(t_j)}, \\ \hat{\mathrm{E}}\left\{(1-\Delta)Y(t_j)e^{\hat{\beta}^T Z_i(t_j)}Z_i(t_j)\right\} &= \frac{1}{n}\sum_{i=1}^n \frac{\xi_i}{\hat{\alpha}(t)}(1-\Delta_i)Y_i(t)e^{\hat{\beta}^T Z_i(t_j)}Z_i(t_j).\end{aligned}$$

CHAPTER 4: REGULARIZED VARIABLE SELECTION FOR ADDITIVE HAZARDS MODEL WITH A STRATIFIED CASE-COHORT DESIGN

4.1 Introduction

In modern large-scale epidemiological cohort studies, investigators are usually interested in assessing the association between a large number of risk factors and the outcome. Collecting information on risk factors often requires expensive bioassays and precious biological specimens such as serum and genetic material. When the outcome is time-to-event data, Prentice (1986) proposed a case-cohort design to reduce the cost and effort in measuring expensive covariates without decreasing much efficiency in the estimation. In a case-cohort design, the complete covariate information is only obtained from a randomly sampled subset of the full cohort plus all subjects who developed the outcome. In practice, some covariates that are correlated with the more expensive exposure variables may be readily available for the entire cohort. Borgan et al. (2000) proposed a stratified case-cohort design based on the correlated covariates to gain efficiency in the estimation. For example, in the Atherosclerosis Risk in Communities (ARIC) study (Ballantyne et al. 2004) a large cohort of 15,792 individuals aged 45 to 64 years old were sampled from four U.S. communities and were followed for ten years for the development of Coronary Heart Disease (CHD). The primary interest was to assess the association between the protein hs-CRP level and risk of incident CHD. To preserve stored plasma and reduce costs, a stratified case-cohort design was implemented, where a random subset was selected from each stratum defined by sex, race, and baseline age. The hs-CRP level was measured only on these subsets plus all incident CHD cases.

Perhaps the most popular model for the analysis of time-to-event data is the Cox proportional hazards model (Cox 1972), where the effect of covariates on the risk of event is assumed multiplicative. The popularity of the Cox proportional hazards model is largely due to its desirable theoretical properties and wide availability of its implementation in computer programs. However, the critical assumption of proportional hazards may fail to hold in many situations, making the Cox model invalid. For example, in the ARIC study there is evidence that the risk of CHD does not satisfy the proportionality assumption (Kang et al. 2013). Moreover, investigators are sometimes more interested in the risk difference attributed to the covariates. The risk difference is more relevant to public health because it translates directly into the number of disease cases that would be avoided by eliminating a particular exposure (Kulich and Lin 2000). The risk difference is also easier to interpret and communicate to medical practitioners. Therefore, the additive hazards model is often used as an important alternative to the Cox proportional hazards model to analyze time-to-event outcome. As its name suggests, the additive hazards model assumes that the effect of covariates on the risk of event is additive. Since Aalen (1980) first introduced the additive hazards model, many authors have investigated its estimation procedure and the properties of the estimator. Lin and Ying (1994) proposed a semiparametric estimating equation for a special case of additive hazards model where the regression coefficients are time-independent. The authors derived the limiting distribution of the estimator and studied its semiparametric efficiency. Kulich and Lin (2000) extended this estimation method to case-cohort design and assessed its asymptotic relative efficiency with respect to the full cohort analysis.

In case-cohort studies where a large number of covariates are collected, researchers are often interested in selecting a subset of the covariates that are related to the event of interest. With the inclusion of interaction terms and polynomial terms, the number of candidate covariates can be very large. In the ARIC study, there are a number of potential

confounders or effect modifiers that need to be considered in the modeling process. With the pairwise interactions between hs-CRP level and all the other covariates as well as the squared continuous covariates, the total number of candidate covariates is quite large in comparison to the number of events. As Huber (1973) argued, in the context of variable selection the number of parameters should be considered as increasing with sample size, and goes to infinity as sample size goes to infinity. Therefore, an efficient variable selection procedure that allows a diverging number of parameters is needed for an additive hazards model with a case-cohort design. Here we allow the number of parameters to increase at a slower rate than the sample size. Thus, the model dimension is still less than the sample size even though it diverges to infinity.

Regularized variable selection procedures have been developed over the last few decades. Under certain regularity conditions, these procedures can simultaneously select variables and estimate their coefficients. Among various penalty functions used in these procedures, the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) and a few other have been shown to identify the true model with probability tending to one as sample size goes to infinity and estimate the non-zero parameters with full efficiency as if the true model is known *a priori*. The SCAD variable selection procedure has been successfully applied to linear, generalized linear, Cox proportional hazards, and additive hazards model. However, to our knowledge, its properties have not been studied under additive hazards model with stratified case-cohort design where covariates are not observed for all subjects. The diverging number of parameters adds to the complexity of the theoretical derivation.

In this chapter of the dissertation, we investigate the asymptotic properties and finite sample performance of the SCAD-penalized variable selection procedure in additive hazards model with a stratified case-cohort design. We focus on Lin and Ying (1994) estimation method assuming time-independent parameters and simple random sampling in

the case-cohort design. We first establish the rate of convergence of the maximum penalized pseudo-partial likelihood estimator. We then prove the model selection consistency of the procedure and derive the limiting distribution of the estimator. As tuning parameter selection is critical for the performance of regularized variable selection procedure, we propose a new cross-validation based tuning parameter selection strategy, and empirically evaluate its performance under large cohort size but fairly high censoring percentage settings, which are two typical features of case-cohort studies. The aim of this chapter is to provide theoretical foundation as well as practical guidance for variable selection in additive hazards model under stratified case-cohort design and a diverging dimension, and thereby facilitates large-scale studies on public health issues.

4.2 Additive Hazards Model with A Stratified Case-Cohort Design

Suppose the full cohort of size n is divided into H mutually exclusive strata based on some categorical variables that are available for all subjects. For subject i in stratum h, let T and C be respectively the time to the outcome of interest and the censoring time, and $Z(t)$ be the $d_n \times 1$ possibly time-dependent covariate vector. T and C are assumed to be independent conditional on Z. Let $\beta = (\beta_1, ..., \beta_{d_n})^T$ be a vector of unknown regression coefficients. Let $X = \min(T, C)$ be the observed time and $\Delta = I(T \leq C)$ be the censoring indicator, where $I(\cdot)$ is an indicator function. Let τ be the time at the end of study. Define for subject i in stratum h the counting process $N_{hi}(t) = I(X_{hi} \leq t, \Delta_{hi} = 1)$, and the at risk process $Y_{hi}(t) = I(X_{hi} \geq t)$. Let $\lambda_{hi}(t)$ denote the hazard function for subject i in stratum h. The additive hazards model assumes

$$\lambda_{hi}(t|Z_{hi}(t)) = \lambda_0(t) + \beta^T Z_{hi}(t), \tag{4.1}$$

where $\lambda_0(t)$ is an unspecified common baseline hazard function for all strata, and β is constant over time. Under the stratified case-cohort design, we randomly select a subcohort

of fixed size from each stratum. We assume that the selection of subcohort is independent across the strata. Let $\tilde{n}_h$ denote the size of subcohort in stratum h, n_h denote the size of stratum h, and ξ_{hi} be the indicator of subject i being selected into the subcohort in stratum h. Then for subject in stratum $h = 1, ..., H$, the selection probability $\text{pr}(\xi_{hi} = 1) = \tilde{n}_h/n_h = \alpha_h$. Under the simple random sampling $(\xi_{h1},...,\xi_{hn_h})$ are correlated. Assuming the complete covariate histories are available for the cases outside the subcohort throughout their at-risk periods, we proposed the following estimating equation for the regression coefficients β,

$$U(\beta) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \rho_{hi}(t)\left\{Z_{hi}(t) - \bar{Z}(t)\right\}\left\{dN_{hi}(t) - Y_{hi}(t)\beta^T Z_{hi}(t)dt\right\},$$

where $\bar{Z}(t) = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\rho_{hj}(t)Y_{hj}(t)Z_{hj}(t)/\sum_{h=1}^{H}\sum_{j=1}^{n_h}\rho_{hj}(t)Y_{hj}(t)$, $\rho_{hi}(t) = \Delta_{hi}+(1-\Delta_{hi})\xi_{hi}\hat{\alpha}_h^{-1}(t)$, and $\hat{\alpha}_h(t) = \sum_{i=1}^{n_h}\xi_{hi}(1-\Delta_{hi})Y_{hi}(t)/\sum_{i=1}^{n_h}(1-\Delta_{hi})Y_{hi}(t)$. This estimating equation is based on Kulich and Lin (2000) with the selection probability α_h replaced by its time-dependent sample estimate $\hat{\alpha}_h(t)$. The estimator $\hat{\beta}$ solves $U(\beta)$ and takes on a closed form

$$\hat{\beta} = \left[\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \rho_{hi}(t)\left\{Z_{hi}(t) - \bar{Z}(t)\right\}^{\otimes 2} Y_{hi}(t)dt\right]^{-1}\left[\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \left\{Z_{hi}(t) - \bar{Z}(t)\right\} dN_{hi}(t)\right], \tag{4.2}$$

where $a^{\otimes 2} = aa^T$ for a vector a.

4.3 Variable Selection in Additive Hazards Model with A Stratified Case-Cohort Design

4.3.1 Penalized loss function

Unlike the Cox proportional hazards model where the log-partial likelihood function is a natural choice of loss function for variable selection, under additive hazards model the likelihood function is difficult to work with due to the nonparametric baseline hazard function and the additive structure. Motivated by the similarity between the Lin-Ying

estimator for additive hazards model (Lin and Ying 1994) and the least square estimator, Martinussen and Scheike (2009) proposed the loss function that is the integral of the Lin-Ying estimating equation with respect to β. We propose a loss function under stratified case-cohort design

$$\tilde{L}_n(\beta) = \frac{1}{2}(\beta^T \tilde{A}_n \beta - 2\beta^T \tilde{b}_n),$$

where

$$\tilde{A}_n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau \rho_{hi}(t) \left\{ Z_{hi} - \bar{Z}(t) \right\}^{\otimes 2} Y_{hi}(t) dt,$$

$$\tilde{b}_n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau \left\{ Z_{hi} - \bar{Z}(t) \right\} dN_{hi}(t).$$

We then propose the following objective function for variable selection,

$$\tilde{Q}_n(\beta) = \tilde{L}_n(\beta) + n \sum_{j=1}^{d_n} P_{\lambda_{jn}}(|\beta_j|), \tag{4.3}$$

where $P_{\lambda_{jn}}(|\beta_j|)$ is a nonnegative penalty function with λ_{jn} as the tuning parameter controlling the model complexity. We use SCAD penalty proposed by Fan and Li (2001) with the modification that the tuning parameter λ_n is covariate-specific, which allows different regression coefficients to have different penalty functions. When $\lambda_{jn} = 0$, no penalty is applied to β_j. The first derivative of the SCAD penalty is given by

$$P'_{\lambda_n}(\theta) = \lambda_n I(\theta \le \lambda_n) + \frac{(a\lambda_n - \theta)_+}{a-1} I(\theta > \lambda_n), \tag{4.4}$$

for some $a > 2$ and $\theta > 0$, with $P_{\lambda_n}(0) = 0$.

4.3.2 Notations and Regularity Conditions

We denote by $\hat{\beta}$ the penalized estimator that minimizes (4.3). We denote by β_0 the true value of β. Let $\beta_0 = (\beta_{I0}^T, \beta_{II0}^T)^T$, where β_{I0} and β_{II0} are the nonzero and zero components

of β_0, respectively. Let $\hat{\beta} = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$, where $\hat{\beta}_I$ and $\hat{\beta}_{II}$ are the penalized pseudo-partial likelihood estimators of β_{I0} and β_{II0}, respectively. Denote by k_n the dimension of β_{I0} with k_n/d_n converging to a constant $c \in [0,1]$. We define the following notations.

$$S^{(k)}(t) = n^{-1}\sum_{h=1}^{H}\sum_{i=1}^{n_h} Y_{hi}(t)Z_{hi}(t)^{\otimes k} \quad \tilde{S}^{(k)}(t) = n^{-1}\sum_{h=1}^{H}\sum_{i=1}^{n_h} \rho_{hi}(t)Y_{hi}(t)Z_{hi}(t)^{\otimes k}, \;\; k = 0,1,2$$

$$s^{(k)}(t) = \mathrm{E}\{S^{(k)}(t)\}, k = 0,1,2, \qquad e(t) = \frac{s^{(1)}(t)}{s^{(0)}(t)}$$

$$\mathscr{A}_n(\beta) = \mathrm{E}\left[\int_0^\tau \{Z(t) - e(t)\}^{\otimes 2} Y(t)dt\right] \qquad \Gamma_n(\beta) = \frac{1}{n}\mathrm{var}\{\tilde{L}_n'(\beta)\}$$

$$\phi_n = \max_{1\le j\le k_n}\{|P'_{\lambda_{jn}}(|\beta_{j0}|)|\}, \qquad \psi_n = \max_{1\le j\le k_n}\{|P''_{\lambda_{jn}}(|\beta_{j0}|)|\}$$

$$\Psi_n = \mathrm{diag}\{P''_{\lambda_{1n}}(|\beta_{10}|), ..., P''_{\lambda_{k_n n}}(|\beta_{k_n 0}|)\}$$

$$\Phi_n = \left(P'_{\lambda_{1n}}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), ..., P'_{\lambda_{s_n n}}(|\beta_{k_n 0}|)\mathrm{sgn}(\beta_{k_n 0})\right)^T$$

We require the following regularity conditions. The conditions on the higher-order moment of the loss function is necessary due to the diverging number of parameters.

(A) $\int_0^\tau \lambda_0(t)dt < \infty$ and $\mathrm{E}\{Y(\tau)\} > 0$.

(B) $|Z_{hij}(0)| + \int_0^\tau |dZ_{hij}(t)| < C_1 < \infty$ almost surely for some constant C_1 and $h = 1, ..., H$, $i = 1, ..., n$, and $j = 1, ..., d_n$, i.e. $Z_{hij}(t)$ has bounded variation almost surely.

(C) There exists a neighborhood $\mathscr{B}$ of β_0 such that for all $\beta \in \mathscr{B}$ and $t \in [0,\tau]$, $\partial s^{(0)}(\beta,t)/\partial\beta = s^{(1)}(\beta,t)$, and $\partial^2 s^{(0)}(\beta,t)/\partial\beta\partial\beta^T = s^{(2)}(\beta,t)$. The functions $s^{(k)}(\beta,t)$ $(k = 0,1,2)$ are continuous and bounded and $s^{(0)}(\beta,t)$ is bounded away from 0 on $\mathscr{B} \times [0,\tau]$.

(D) $\alpha_h = \tilde{n}_h/n_h$ converges to a constant $C_{2h} \in (0,1)$ for $h = 1, ..., H$ as $n \to \infty$.

(E) For each n, there exist positive constants C_3, C_4, C_5, and C_6 such that

$$0 < C_3 < \mathrm{eigen}_{\min}\{\mathscr{A}_n(\beta_0)\} \le \mathrm{eigen}_{\max}\{\mathscr{A}_n(\beta_0)\} < C_4 < \infty$$

$$0 < C_5 < \text{eigen}_{\min}\{\Gamma_n(\beta_0)\} \leq \text{eigen}_{\max}\{\Gamma_n(\beta_0)\} < C_6 < \infty$$

where $\text{eigen}_{\min}\{\cdot\}$ and $\text{eigen}_{\max}\{\cdot\}$ are the minimum and maximum of the eigenvalues of a matrix, respectively.

(F) $\liminf_{n\to+\infty} \liminf_{\theta\to 0+} P'_{\lambda_{jn}}(\theta)/\lambda_{jn} > 0$.

(G) $\min_{1\leq j\leq k_n} |\beta_{0j}|/\lambda_{jn} \to \infty$ as $n \to \infty$.

4.3.3 Asymptotic Properties of Penalized Estimator

We first prove the existence of a penalized estimator and establish its convergence rate. Only main results are presented here. The outline of the proofs are provided in Section 4.7.

Theorem 4.3.1. *Under Conditions (A) to (E), if $\psi_n \to 0$ and $d_n^2/n \to 0$ as $n \to \infty$, then with probability tending to one there exists a local minimizer $\hat{\beta}$ of $\tilde{Q}_n(\beta)$, as defined in (4.3), such that $\|\hat{\beta} - \beta_0\| = O_p\{d_n^{1/2}(n^{-1/2} + \phi_n)\}$.*

From Theorem 4.3.1 one can obtain a $n^{1/2}d_n^{-1/2}$-consistent penalized estimator, provided that $\phi_n = O(n^{-1/2})$, which is the case for SCAD penalty.

Theorem 4.3.2. *Under Conditions (A) to (H), as $n \to \infty$, if $\psi_n \to 0, d_n^2/n \to 0, \lambda_{jn} \to 0, \lambda_{jn}n^{1/2}d_n^{-1/2} \to \infty$, and $\phi_n = O(n^{-1/2})$, then the $n^{1/2}d_n^{-1/2}$-consistent local minimizer $\hat{\beta} = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ must satisfy*

(i) *$\hat{\beta}_{II} = 0$ with probability tending to one;*

(ii) *for any nonzero $k_n \times 1$ constant vector u with $u^T u = 1$,*

$$n^{1/2}u^T\Gamma_{n11}^{-1/2}(\beta_0)\{\mathscr{A}_{n11}(\beta_0) + \Psi_n\}\{\hat{\beta}_I - \beta_{I0} + (\mathscr{A}_{n11}(\beta_0) + \Psi_n)^{-1}\Phi_n\} \to N(0,1)$$

in distribution, where $\mathscr{A}_{n11}(\beta_0)$ consists of the first $k_n \times k_n$ components of $\mathscr{A}_n(\beta_0)$, and $\Gamma_{n11}(\beta_0)$ consists of the first $k_n \times k_n$ components of $\Gamma_n(\beta_0)$.

For the SCAD penalty, $\phi_n = 0$, $\Psi_n = 0$, and $\Phi_n = 0$ for large enough n under Condition (G). Therefore, the result of Theorem 4.3.2 reduces to

$$n^{1/2} u^T \Gamma_{n11}^{-1/2}(\beta_0) \mathscr{A}_{n11}(\beta_0)(\hat{\beta}_I - \beta_{I0}) \to N(0,1)$$

in distribution.

4.4 Considerations in Practical Implementation

4.4.1 Local Quadratic Approximation and Variance Estimation

Since the SCAD penalty function is singular at the origin, in practical implementation the penalized estimator cannot be directly obtained by solving the first derivative of (4.3). Instead, we follow Fan and Li (2001) to use a local quadratic approximation (LQA) to the penalty function. The unpenalized loss function $\tilde{L}_n(\beta)$ is a special case of (4.3) with $P_{\lambda_{jn}}(|\beta_j|) = 0$ for all $j = 1, ..., d_n$. Applying Theorem 4.3.1 with $a_n = 0$, we know there exists a $n^{1/2} d_n^{-1/2}$-consistent minimizer of (4.3). We use this minimizer as the initial value $\beta^{(0)}$ for the LQA algorithm. If $|\beta_j^{(0)}|$ is less than a pre-specified small positive constant c_j, then set $\hat{\beta}_j = 0$. In practice c_j is set to equal λ_{jn}. Otherwise, the penalty function is locally approximated by a quadratic function as

$$P_{\lambda_{jn}}(|\beta_j|) \approx P_{\lambda_{jn}}(|\beta_j^{(0)}|) + \frac{1}{2} \frac{P'_{\lambda_{jn}}(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}(\beta_j^2 - \beta_j^{(0)2}),$$

and therefore $P'_{\lambda_{jn}}(|\beta_j|) \approx \{P'_{\lambda_{jn}}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j$. With the approximated quadratic penalty function, a closed-form maximizer can be computed by solving the first derivative of the approximated objective function. The absolute value of each component of the minimizer is again compared to the pre-specified constant c_j and set to 0 if it is smaller than c_j. The remaining nonzero updated parameter estimate is used as the new initial value. This

process is iterated until convergence or no nonzero parameter estimate is left.

The sandwich estimate of the covariance matrix for $\hat{\beta}^*$, the nonzero components of $\hat{\beta}$, can be directly obtained from the last iteration of the above LQA algorithm as

$$\begin{aligned}\hat{\text{cov}}(\hat{\beta}^*) =& \{\tilde{L}_n^{*\prime\prime} + n\Phi_n(\hat{\beta}^*)\}^{-1}\hat{\text{var}}\{\tilde{L}_n^{*\prime}(\hat{\beta}^*)\}\{\tilde{L}_n^{*\prime\prime} + n\Phi_n(\hat{\beta}^*)\}^{-1} \\ =& \{\tilde{A}_n^* + n\Phi_n(\hat{\beta}^*)\}^{-1} n\hat{\Gamma}_n(\hat{\beta}^*)\}\{\tilde{A}_n^* + n\Phi_n(\hat{\beta}^*)\}^{-1},\end{aligned}$$

where $\tilde{A}_n^*$ is the sub-matrix of $\tilde{A}_n$ corresponding to $\hat{\beta}^*$, $\Phi_n(\beta^*) = \text{diag}\{P'_{\lambda_{1n}}(|\hat{\beta}_1^*|)/|\hat{\beta}_1^*|, ..., P'_{\lambda_{k_n^* n}}(|\hat{\beta}_{k_n^*}^*|)/|\hat{\beta}_{k_n^*}^*|\}$, $\hat{\Gamma}_n(\hat{\beta}^*)$ is the estimate of $\Gamma_n(\hat{\beta}^*)$, and k_n^* is the dimension of $\hat{\beta}^*$. Note that the sandwich estimate of the covariance matrix does not apply to the zero estimate of the parameters.

4.4.2 Selection of Tuning Parameters

The tuning parameters λ's involved in the SCAD penalty function $P_\lambda(\cdot)$ control the magnitude of the penalty on each regression coefficient and thereby control the complexity of the selected model. In practical implementation, the attractive properties of the penalized estimator heavily depend on the choice of the appropriate tuning parameters. The typical methods of selecting the tuning parameters are automatic data-driven procedures such as K-fold cross-validation and generalized cross-validation (GCV) (Craven and Wahba 1979). The d_n-dimensional optimization problem is difficult to solve in practice. We follow Cai et al. (2005) to take $\lambda_{jn} = \lambda_n \text{se}(\beta_j^{(0)})$, where $\text{se}(\beta_j^{(0)})$ is the estimated standard error of the unpenalized estimator. Then the optimization problem reduces to 1-dimensional and a grid-search can be performed to identify the optimal λ_n. In the literature of variable selection in Cox's proportional hazards model the GCV is predominantly used due to the availability of the partial likelihood function. Under additive hazards model, however, no such likelihood function is available. Therefore, authors have been exclusively using K-fold cross-validation with $\tilde{L}_n(\beta)$ as the natural choice of loss function. In this study we take

$K = 5$. Denote the full dataset by D and the training and validation dataset by $D - D^{\nu}$ and D^{ν}, respectively, for $\nu = 1, ..., 5$. For each λ, compute $\tilde{L}_{n^{\nu}}(\hat{\beta}^{-\nu}(\lambda))$ based on the validation dataset, where $\hat{\beta}^{-\nu}(\lambda)$ is the penalized estimate based on the training dataset and λ. The conventional cross-validation statistics is defined as

$$\mathrm{CV}(\lambda) = \sum_{\nu=1}^{5} \tilde{L}_{n^{\nu}}(\hat{\beta}^{-\nu}(\lambda)), \tag{4.5}$$

and λ is chosen by minimizing (4.5). However, cross-validation method is based on minimizing the prediction error rather than model selection consistency. In fact, it is asymptotically equivalent to the Akaike information criterion (AIC) (Akaike 1973), which has a positive probability of overfitting the model as sample size goes to infinity. Case-cohort studies usually bare the distinctive property of large sample size. Therefore, the overfitting effect of cross-validation may become more prominent in case-cohort studies. In this study we propose a modified cross-validation method that incorporates an additional penalty term in the cross-validation statistics. The penalized statistic is defined as

$$\mathrm{CV}^{\mathrm{P}}(\lambda) = \sum_{\nu=1}^{5} \{\tilde{L}_{n^{\nu}}(\hat{\beta}^{-\nu}(\lambda)) + k^{-\nu}\}, \tag{4.6}$$

where $k^{-\nu}$ is the number of nonzero components of $\hat{\beta}^{-\nu}$. We denote the minimizer of (4.5) and (4.6) as λ_n^{CV} and λ_n^{CVP}, respectively. In the simulation section that follows, we empirically investigate the model selection performance of these two tuning parameter selection criteria. According to Fan and Li (2001), the second tuning parameter a in the SCAD penalty is set to 3.7 in our study.

4.5 Numerical Study and Application

4.5.1 Simulation Study

Independent failure times are generated by the additive hazards model $\lambda_i(t|Z_i(t)) = \lambda_0(t) + \beta^T Z_i(t)$. We set $\lambda_0(t) = 2$ and the dimension of β to be $d_n = [0.3 * n_c^{1/2}]$ to reflect its dependence on sample size, where n_c is the number of cases and $[x]$ rounds x to the nearest integer. We use n_c instead of n to determine model size because the former better represents the amount of information in the dataset. The smallest nonzero parameter in terms of the absolute value is set to 0.70 or 0.43, which represents 35% and 22% increase from the baseline hazard for one standard deviation increase in the covariate. The remaining nonzero parameters recycling from values -0.8 and 1. There is one nonzero parameter for every two zero parameters. To generate the design matrix and strata, we first generate a $(d_n + 1)$-dimensional multivariate standard normal variable Z^* with the correlation coefficient between Z_i^* and Z_j^* being $0.5^{|i-j|}$. The first component is then dichotomized with a cutoff value of 0 and used to define two strata. For the remaining d_n components, we dichotomize half of them with a cutoff value of 0. As a result, the design matrix consists of a mixture of correlated binary and continuous covariates that are correlated with the stratification variable. A simple random sample is selected independently for each stratum. Censoring times C_i are generated from a uniform distribution $\mathrm{U}(0, c)$ where c is adjusted to achieve desired censoring percentage.

Two sample sizes, two censoring rates, and two sampling proportions of the random subcohort are considered for each minimum effect size ($\beta_{\min}$=0.70 and 0.43). Comparisons are made on the performance of penalized variable selection procedures with tuning parameter λ_n^{CV} and λ_n^{CVP}. As a benchmark, we include the hard threshold variable selection procedure, where the component of the minimizer of the unpenalized loss function $\tilde{L}_n(\beta)$ is set to 0 if its p-value from the Wald test is larger than 0.05. We also include as another benchmark the Oracle procedure where the correct subset of covariates is used to fit the model. As

the censoring rate in case-cohort studies is typically high, we set it to 80% and 90% in our simulation to better mimic real-world studies. For each setting 1000 replications are conducted.

The performance of the model selection procedure is evaluated by model error defined as $\mathrm{ME}(\hat{\mu}) = \mathrm{E}\{\mathrm{E}(Y|Z) - \hat{\mu}(Z)\}^2$. Under the additive hazards model with constant baseline hazard λ_0, it can be shown that $\mathrm{E}(Y|Z) = (\lambda_0 + \beta_0^T Z)^{-1}$ and $\hat{\mu}(Z) = (\lambda_0 + \hat{\beta}^T Z)^{-1}$. Therefore, $\mathrm{ME}(\hat{\mu}) = \mathrm{E}\{(\lambda_0 + \hat{\beta}^T Z)^{-1} - (\lambda_0 + \beta_0^T Z)^{-1}\}^2$. We further define the relative model error (RME) of a model selection procedure as the ratio of its model error to that of the unpenalized estimates from the full model. Following Tibshirani (1996), we use the median and the median absolute deviation (MAD) of the relative model error to compare the performance of different model selection procedures. We also calculate the average number of parameters correctly estimated as 0, the average number of parameters erroneously estimated as 0, and the overall rate of identifying the true model (RITM). In addition, point estimates, empirical and model-based standard errors, and the empirical 95% confidence interval coverage are calculated for $\hat{\beta}_{\min}$ using replications with nonzero $\hat{\beta}_{\min}$.

Table 4.1 summarizes the model selection performance when $\beta_{\min} = 0.70$. The CVP tuning parameter selection method outperforms the CV tuning parameter selection method in all settings in terms of relative model error (RME) and the rate of identifying the true model (RITM). It also outperforms the hard threshold method except for the scenarios with $n = 5000$ and 90% censoring rate. Further, higher sampling proportion of the random subcohort is associated with better model selection performance of the CVP method but seems to have no effect on the performance of CV and hard threshold methods. The relatively low RITM for the CV method is apparently due to its overfitting effect as shown by the low average number of correctly identified zero parameters. Table 4.2 summarizes the estimation result of $\beta_{\min}$ under settings in Table 4.1. Given that $\beta_{\min}$ is correctly identified as nonzero, all procedures produce approximately unbiased point estimates. The estimates

are slightly smaller than the true value when the model dimension is the largest (d_n = 13). The model-based standard error estimates are very close to the empirical standard errors and the 95% confidence interval coverage is close to the nominal level.

Table 4.3 summarizes the model selection performance when $\beta_{\min}$ = 0.43. Under the same setting, there is a decrease in the model selection performance for all three procedures in comparison to that with larger $\beta_{\min}$. This is expected as smaller effect is more difficult to detect. Nevertheless, similar to Table 4.1, the procedure with λ_n^{CVP} outperforms the other procedures in all settings. Higher sampling proportion of the random sub-cohort is again associated with better performance of the CVP method but not the other two methods. Table 4.4 shows the estimation result of $\beta_{\min}$ under settings in Table 4.3. Conditional on correctly identifying $\beta_{\min}$ all procedures produce fairly unbiased estimation in the parameter and its standard error and the 95% confidence interval coverage is close to the nominal level.

4.5.2 Analysis of ARIC Study

We use the model selection procedures investigated in Section 4.5.1 to analyze the ARIC study data (Ballantyne et al. 2004). As mentioned in Section 4.1, a cohort of 15,792 individuals were sampled from four U.S. communities and followed for ten years for the development of CHD. After excluding subjects for missing data and other reasons, a total of 12,351 subjects comprised the potential full cohort. Those who were alive or free of disease by the end of 1998 or lost to follow-up in the middle of the study periods were treated as censored. A random subcohort of size 890 was selected by stratified random sampling from strata defined by sex, race (black versus white), and age at baseline ($\leq$ 55 versus >55). After including all CHD cases, the case-cohort size is 1567. There is a total of 735 CHD cases, corresponding to a censoring rate of 94.1%. In this analysis we are primarily interested in identifying risk factors for incidence CHD. In particular, the

Table 4.1: Model selection performance with $\beta_{\min} = 0.70$

	80% Censored				90% Censored			
	RME	Zero Parm.		RITM	RME	Zero Parm.		RITM
Method	median (MAD)	C	I	(%)	median (MAD)	C	I	(%)
	$n = 5000$, $\alpha = 0.3$, $d_n = 9$ for 80% censored, $d_n = 7$ for 90% censored							
HT	0.76 (0.33)	5.7	0	74.2	0.81 (0.33)	3.78	0.06	76.2
CV	0.79 (0.21)	5.09	0.02	63.6	0.92 (0.19)	3.23	0.08	56.6
CVP	0.63 (0.29)	5.92	0.08	86.6	0.84 (0.46)	3.93	0.31	66.6
Oracle	0.58 (0.28)	6	0	100	0.59 (0.29)	4	0	100
	$n = 5000$, $\alpha = 0.5$, $d_n = 9$ for 80% censored, $d_n = 7$ for 90% censored							
HT	0.79 (0.32)	5.68	0	71.6	0.78 (0.31)	3.79	0.03	78.4
CV	0.78 (0.22)	5.08	0.01	65	0.93 (0.18)	3.16	0.09	55.6
CVP	0.58 (0.29)	5.96	0.05	91.2	0.76 (0.43)	3.95	0.26	71.6
Oracle	0.54 (0.26)	6	0	100	0.59 (0.28)	4	0	100
	$n = 10000$, $\alpha = 0.3$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored							
HT	0.77 (0.25)	7.57	0	65.2	0.72 (0.31)	5.69	0	73.8
CV	0.77 (0.28)	7.09	0.01	65.6	0.74 (0.26)	5.13	0.02	67.8
CVP	0.6 (0.32)	7.88	0.04	86.4	0.54 (0.26)	5.96	0.05	91.6
Oracle	0.51 (0.27)	8	0	100	0.5 (0.25)	6	0	100
	$n = 10000$, $\alpha = 0.5$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored							
HT	0.76 (0.24)	7.59	0	67.4	0.79 (0.32)	5.69	0	73.8
CV	0.79 (0.26)	7.07	0	66	0.76 (0.24)	5.09	0.01	65.4
CVP	0.59 (0.31)	7.96	0.02	94.6	0.54 (0.27)	5.96	0.04	92.8
Oracle	0.57 (0.31)	8	0	100	0.51 (0.26)	6	0	100

n: sample size; α: sampling proportion of random sub-cohort for both strata; d_n: number of parameters; RME: relative model error; MAD: median absolute deviation; C: average number of 0 parameters correctly identified as 0; I: average number of nonzero parameters incorrectly identified as 0; RITM: rate of identifying true model; HT: hard threshold method; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection.

Table 4.2: Estimation result for $\beta_{\min} = 0.70$

	80% Censored				90% Censored			
Method	$\hat{\beta}_{\min}$	se_e	se_m	95% CI_e	$\hat{\beta}_{\min}$	se_e	se_m	95% CI_e
$n = 5000$, $\alpha = 0.3$, $d_n = 9$ for 80% censored, $d_n = 7$ for 90% censored								
HT	0.69	0.14	0.13	94.4	0.68	0.14	0.15	97.4
CV	0.69	0.12	0.11	92.4	0.68	0.13	0.14	96.4
CVP	0.69	0.1	0.1	94	0.69	0.13	0.13	96.8
Oracle	0.69	0.1	0.1	94.6	0.68	0.13	0.13	95.8
$n = 5000$, $\alpha = 0.5$, $d_n = 9$ for 80% censored, $d_n = 7$ for 90% censored								
HT	0.69	0.12	0.11	92.8	0.67	0.13	0.14	96.4
CV	0.68	0.1	0.09	92.4	0.68	0.12	0.12	96
CVP	0.69	0.09	0.09	94.2	0.68	0.12	0.12	97.1
Oracle	0.69	0.09	0.09	94.2	0.68	0.12	0.12	96
$n = 10000$, $\alpha = 0.3$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored								
HT	0.67	0.1	0.1	93.8	0.69	0.11	0.12	95.6
CV	0.67	0.09	0.09	93.6	0.69	0.1	0.1	94.4
CVP	0.67	0.09	0.09	94.2	0.68	0.09	0.09	94.4
Oracle	0.67	0.09	0.09	94.8	0.68	0.09	0.09	94.4
$n = 10000$, $\alpha = 0.5$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored								
HT	0.67	0.09	0.09	92.6	0.69	0.11	0.11	95.6
CV	0.67	0.08	0.08	90.8	0.68	0.09	0.09	93
CVP	0.67	0.08	0.08	92.2	0.68	0.09	0.09	93.6
Oracle	0.67	0.08	0.08	92.4	0.68	0.09	0.09	93.4

n: sample size; α: sampling proportion of random sub-cohort for both strata; d_n: number of parameters; se_e: empirical standard error; se_m: model-based standard error; 95% CI_e: empirical 95% confidence interval coverage; HT: hard threshold method; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_{\min}$.

Table 4.3: Model selection performance with $\beta_{\min} = 0.43$

	80% Censored				90% Censored			
	RME	Zero Parm.		RITM	RME	Zero Parm.		RITM
Method	median (MAD)	C	I	(%)	median (MAD)	C	I	(%)
	$n = 10000$, $\alpha = 0.3$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored							
HT	0.7 (0.26)	7.61	0.02	66.8	0.73 (0.33)	5.74	0.03	75.2
CV	0.72 (0.28)	6.83	0.03	61	0.76 (0.24)	5.06	0.06	59
CVP	0.54 (0.28)	7.89	0.09	83.8	0.6 (0.34)	5.95	0.22	76.8
Oracle	0.47 (0.24)	8	0	100	0.47 (0.27)	6	0	100
	$n = 10000$, $\alpha = 0.5$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored							
HT	0.7 (0.25)	7.63	0	68.6	0.69 (0.32)	5.73	0.02	76
CV	0.69 (0.31)	7.07	0.01	67.2	0.75 (0.25)	5	0.04	61.8
CVP	0.53 (0.28)	7.92	0.05	88.6	0.58 (0.3)	5.95	0.17	80.8
Oracle	0.49 (0.25)	8	0	100	0.49 (0.26)	6	0	100
	$n = 15000$, $\alpha = 0.3$, $d_n = 16$ for 80% censored, $d_n = 11$ for 90% censored							
HT	0.73 (0.28)	9.46	0.05	54.9	0.72 (0.32)	6.67	0.03	70.5
CV	0.7 (0.3)	8.81	0.08	53	0.78 (0.28)	5.82	0.05	57.4
CVP	0.58 (0.31)	9.78	0.25	66.5	0.61 (0.35)	6.86	0.22	74
Oracle	0.44 (0.23)	10	0	100	0.48 (0.26)	7	0	100
	$n = 15000$, $\alpha = 0.5$, $d_n = 16$ for 80% censored, $d_n = 11$ for 90% censored							
HT	0.73 (0.27)	9.46	0.01	57.1	0.72 (0.35)	6.7	0.01	73.2
CV	0.73 (0.28)	8.84	0.03	57.6	0.8 (0.27)	5.91	0.04	61.3
CVP	0.59 (0.32)	9.85	0.18	75.6	0.58 (0.32)	6.92	0.13	83.4
Oracle	0.47 (0.26)	10	0	100	0.51 (0.3)	7	0	100

n: sample size; α: sampling proportion of random sub-cohort for both strata; d_n: number of parameters; RME: relative model error; MAD: median absolute deviation; C: average number of 0 parameters correctly identified as 0; I: average number of nonzero parameters incorrectly identified as 0; RITM: rate of identifying true model; HT: hard threshold method; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection.

Table 4.4: Estimation result for $\beta_{\min} = 0.43$

	80% Censored				90% Censored			
Method	$\hat{\beta}_{\min}$	se_e	se_m	95% CI_e	$\hat{\beta}_{\min}$	se_e	se_m	95% CI_e
$n = 10000$, $\alpha = 0.3$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored								
HT	0.43	0.09	0.1	95.9	0.44	0.1	0.12	98.6
CV	0.43	0.09	0.09	95.7	0.44	0.09	0.1	96.8
CVP	0.43	0.08	0.09	97.2	0.45	0.08	0.09	98
Oracle	0.42	0.08	0.09	95.6	0.43	0.09	0.09	96.2
$n = 10000$, $\alpha = 0.5$, $d_n = 13$ for 80% censored, $d_n = 9$ for 90% censored								
HT	0.42	0.09	0.09	93.6	0.43	0.1	0.11	97
CV	0.42	0.08	0.08	93.5	0.43	0.09	0.09	95.4
CVP	0.43	0.07	0.08	96	0.44	0.08	0.09	97.9
Oracle	0.42	0.08	0.08	94.6	0.43	0.08	0.08	96.2
$n = 15000$, $\alpha = 0.3$, $d_n = 16$ for 80% censored, $d_n = 11$ for 90% censored								
HT	0.43	0.1	0.11	98.5	0.44	0.11	0.11	96.2
CV	0.43	0.1	0.1	95.9	0.43	0.1	0.09	94.8
CVP	0.44	0.09	0.1	97.2	0.45	0.09	0.09	96.9
Oracle	0.42	0.1	0.1	95	0.43	0.09	0.09	94.4
$n = 15000$, $\alpha = 0.5$, $d_n = 16$ for 80% censored, $d_n = 11$ for 90% censored								
HT	0.42	0.1	0.1	96	0.43	0.1	0.1	96.5
CV	0.42	0.1	0.09	92.9	0.43	0.09	0.09	93.9
CVP	0.43	0.09	0.09	96.6	0.44	0.08	0.08	96.4
Oracle	0.42	0.09	0.09	93.8	0.43	0.08	0.08	95.7

n: sample size; α: sampling proportion of random sub-cohort for both strata; d_n: number of parameters; se_e: empirical standard error; se_m: model-based standard error; 95% CI_e: empirical 95% confidence interval coverage; HT: hard threshold method; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_{\min}$.

Table 4.5: Baseline characteristics of the cohort of ARIC study

Variables	Full cohort (n=12,351) Mean (SD) or %	Subcohort ($\tilde{n}$=890) Mean (SD) or %
Age (yrs)	58.4 (5.5)	58.2 (5.6)
BMI	28.4 (5.4)	28.1 (5.5)
Systolic blood pressure (mmHg)	126.6 (20.2)	123.5 (18.9)
LDL (mmol/L)	139.4 (38.3)	133.3 (36.6)
HDL (mmol/L)	46.9 (15.6)	49.9 (17.0)
Diabetes (%)	22.9	18.0
Current Smoker (%)	25.5	20.9
CRP level	–	3.12 (3.30)
CRP category (%)		
Low (<1.0mg/L)	–	35.7
Middle (1.0 - 3.0mg/L)	–	33.6
High (>3.0mg/L)	–	25.1

main risk factor of interest is the protein hs-CRP level, which is modeled as a categorical variable of low (<1.0mg/L), middle (1.0 - 3.0mg/L), and high (>3.0mg/L) levels due to its nonlinear effect on the risk of CHD. Since CRP level is the main exposure variable, we do not penalize its regression coefficients and therefore set their tuning parameters to 0. Similarly, we keep the CRP terms in the model for the hard threshold method regardless of their p values. We also consider several other factors in the model selection process: age (years), BMI, systolic blood pressure (mmHg), LDL (mmol/L), HDL (mmol/L), diabetes (yes/no), and current smoker (yes/no). As shown in Kang et al. (2013), the empirical cumulative hazards functions for the different CRP groups increase approximately in a linear fashion. Therefore, the additive hazards model is a reasonable choice.

Table 4.5 summarizes the baseline characteristics of the full cohort and the subcohort. Note that the CRP level is not available for the full cohort due to the case-cohort design. It seems that the distribution of the covariates are similar between the full cohort and sub-cohort, so the subcohort is representative of the full cohort.

We apply the Hard threshold, SCAD penalized variable selection procedures with tuning parameter λ_n^{CV} or λ_n^{CVP} to the ARIC study data to identify important risk factors for

CHD. We include all covariates in Table 4.5 in the initial model. To ensure we do not miss any higher order effect of continuous variables and interactions between CRP and other variables, we include quadratic terms of all continuous variables as well as pairwise interaction between CRP and all other variables in the initial model. All continuous variables are standardized. The tuning parameter selector identified $\lambda_n^{\mathrm{CV}} = 1.577$ and $\lambda_n^{\mathrm{CVP}} = 2.467$. Table 4.6 shows the selected covariates and their estimated coefficients and standard errors by the three methods. The SCAD with λ_n^{CV} selected the largest model and SCAD with λ_n^{CVP} selected the smallest model. This is consistent with the observation in the simulation study that λ_n^{CV} tends to over-select variables compared to λ_n^{CVP}. Besides CRP levels, all three methods identified current smoker, age, LDL, HDL, HDL^2, systolic blood pressure, and interaction between CRP2 and BMI as significant risk factors for CHD. The SCAD with λ_n^{CV} additionally included diabetes, age^2, SBP^2, interaction between CRP3 and BMI, and interaction between CRP2 and SBP in the model.

Based on the model selection result from SCAD penalty with λ_n^{CVP}, the risk of CHD for subjects who are current smoker is 1.099×10^{-5} per-day, or 4.01 per 1,000 person years, higher than those who are not current smoker. Increased age, LDL level, and systolic blood pressure are associated with higher risk of CHD. The effect of HDL level on risk of CHD follows a quadratic form with the minimum risk achieved at an HDL level of 4.4 standard deviations above population mean. This point is so far away from the mean that vast majority of the population lie below this level. Hence there is a negative association between HDL level and risk of CHD, and the magnitude of the association decreases as HDL level increases. This result is consistent with the common knowledge that HDL is the "good" cholesterol. The interaction between CRP2 and BMI means that the effect of BMI on risk of CHD is different in the middle CRP group than the other two CRP groups.

Table 4.6: Estimated coefficients and standard errors from ARIC study data

Variable	Hard Threshold $\hat{\beta}$ (ŝe) ($\times 10^{-5}$)	SCAD (λ_n^{CV}) $\hat{\beta}$ (ŝe) ($\times 10^{-5}$)	SCAD (λ_n^{CVP}) $\hat{\beta}$ (ŝe) ($\times 10^{-5}$)
CRP2 (middle (1.0 - 3.0mg/L))	-0.550(0.282)	-0.4(0.731)	-0.351(0.73)
CRP3 (high (>3.0mg/L))	0.251(0.306)	0.27(0.787)	0.319(0.738)
Current Smoker	1.062(0.362)	1.045(0.738)	1.099(0.733)
Diabetes	0(–)	1.861(0.879)	0(–)
Age	0.457(0.141)	0.401(0.327)	0.469(0.32)
Age^2	0(–)	0.209(0.34)	0(–)
BMI	0(–)	0(–)	0(–)
BMI^2	0(–)	0(–)	0(–)
LDL (mmol/L)	0.57(0.184)	0.615(0.316)	0.587(0.315)
LDL^2	0(–)	0(–)	0(–)
HDL (mmol/L)	-1.328(0.187)	-1.407(0.37)	-1.46(0.366)
HDL^2	0.301(0.058)	0.319(0.173)	0.331(0.172)
Systolic blood pressure (mmHg)	0.745(0.21)	0.967(0.432)	0.858(0.318)
SBP^2	0(–)	0.152(0.193)	0(–)
CRP2*age	0(–)	0(–)	0(–)
CRP3*age	0(–)	0(–)	0(–)
CRP2*BMI	-0.906(0.339)	-0.569(0.647)	-0.515(0.633)
CRP3*BMI	0(–)	-0.466(0.441)	0(–)
CRP2*LDL	0(–)	0(–)	0(–)
CRP3*LDL	0(–)	0(–)	0(–)
CRP2*HDL	0(–)	0(–)	0(–)
CRP3*HDL	0(–)	0(–)	0(–)
CRP2*SBP	-0.546(0.266)	-0.746(0.642)	0(–)
CRP3*SBP	0(–)	0(–)	0(–)
CRP2*current smoker	0(–)	0(–)	0(–)
CRP3*current smoker	0(–)	0(–)	0(–)
CRP2*diabetes	0(–)	0(–)	0(–)
CRP3*diabetes	0(–)	0(–)	0(–)

All continuous covariates are standardized.

4.6 Discussion

In this chapter of the dissertation, we proposed a variable selection procedure based on SCAD penalty in additive hazards model with a stratified case-cohort design and a diverging number of parameters. We investigated its asymptotic and finite sample properties. We showed that, under certain regularity conditions, the variable selection procedure identifies the true model with probability one as samples size goes to infinity, and the penalized estimates from this procedure is consistent and asymptotically normally distributed.

In the simulation study we compared the model selection performance of the conventional cross-validation tuning parameter selection method and the proposed AIC-penalized cross-validation method. We found that the proposed tuning parameter selection method outperforms the conventional cross-validation method in identifying the true model under all simulation scenarios. The cross-validation method focuses on minimizing the prediction error, and have been shown to yield overfitted models (Hastie et al. 2009). Our proposed tuning parameter selection method incorporate an additional penalty term to compensate for the overfitting effect of cross-validation, and therefore gives better result in terms of identifying the true model. In many epidemiological studies, one typical purpose of model fitting is to investigate risk factors and underlying biological mechanisms of diseases on the population level. Under such situation, the emphasis is on identifying the true model rather than predicting the risk of a new individual. In light of this argument, we recommend the AIC-penalized cross-validation method for tuning parameter selection when performing SCAD-penalized model selection in additive hazards model with case-cohort design. Although we have provided empirical evidence for the superiority of AIC-penalized cross-validation method, a theoretical proof is yet to be established.

It is interesting to observe from the simulation study that the variable selection performance of the hard threshold method is closely related to the number of parameters. More parameters in the model leads to decreased performance even if the censoring rate is lower

and sample size is larger. In contrast, the penalized variable selection method identifies the true model with higher rate with increased number of cases and sample size despite the associated larger number of parameters. Therefore, when the model size becomes larger, one can expect the penalized variable selection method to be far more useful than the hard threshold method.

The proposed variable selection method does not have any mechanism to ensure the hierarchical structure of the candidate covariates such as polynomial terms and interactions. As a result, the selected models from the ARIC study does not maintain the hierarchical structure. For example, the model identified by SCAD with λ_n^{CVP} contains an interaction between CRP2 and BMI but not the main effect of BMI. Although this issue does not pose any theoretical difficulties and one can argue that the final model is still a special case of hierarchical model with the coefficients of lower order terms being exactly 0, it poses some difficulties in interpretation. Therefore, a future research topic would be to consider the hierarchical structure of the candidate covariates in model selection of additive hazards model with case-cohort design by using group variable selection techniques.

4.7 Proof of Theorems

Throughout the proofs, denote $\tilde{\ell}_n'(\beta_0)_j = \partial \tilde{L}_n(\beta_0)/\partial\beta_j$, $\tilde{\ell}_n''(\beta_0)_{jk} = \partial^2 \tilde{L}_n(\beta_0)/\partial\beta_j\partial\beta_k$. For a matrix $A = \{a_{ij}\}, i, j = 1, ..., n$, the norm is defined as $\|A\| = (\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2)^{1/2}$.

Lemma 4.7.1. *Given that ξ is independent of Δ and $Y(t)$, for stratum $h = 1, ..., H$, $n_h^{1/2}\{\hat{\alpha}_h^{-1}(t) - \alpha_h^{-1}\}$ converges to independent zero-mean Gaussian processes.*

Proof. By Taylor expansion of $\hat{\alpha_h}(t)$ around α_h,

$$
\begin{aligned}
n_h^{1/2}\{\hat{\alpha}_h^{-1}(t) - \alpha_h^{-1}\} &= -\frac{n_h^{1/2}}{\alpha_h^*(t)^2}\left(\frac{\sum_{i=1}^{n_h}(1-\Delta_{hi})\xi_{hi}Y_{hi}(t)}{\sum_{i=1}^{n_h}(1-\Delta_{hi})Y_{hi}(t)} - \alpha_h\right) \\
&= -\frac{n_h^{1/2}\left\{\sum_{i=1}^{n_h}(1-\Delta_{hi})\xi_{hi}Y_{hi}(t) - \alpha_h\sum_{i=1}^{n_h}(1-\Delta_{hi})Y_{hi}(t)\right\}}{\alpha_h^*(t)^2\sum_{i=1}^{n_h}(1-\Delta_{hi})Y_{hi}(t)}
\end{aligned}
$$

$$= \frac{\alpha_h}{\alpha_h^*(t)^2} \frac{n_h}{\sum_{i=1}^{n_h}(1-\Delta_{hi})Y_{hi}(t)} n_h^{-1/2} \sum_{i=1}^{n_h} \left(1 - \frac{\xi_{hi}}{\alpha_h}\right)(1-\Delta_{hi})Y_{hi}(t),$$

where $\alpha_h^*(t)$ lies between $\hat{\alpha}_h(t)$ and α_h. Since $(1-\Delta)Y(t)$ is of bounded variation almost surely, and $\text{var}\{(1-\Delta)Y(0)\} < \infty$ and $\text{var}\{(1-\Delta)Y(\tau)\} < \infty$, by Lemma 3.7.1, $n_h^{-1/2}\sum_{i=1}^{n_h}(\xi_{hi}/\alpha_h - 1)(1-\Delta_{hi})Y_{hi}(t)$ converges weakly to a tight zero mean Gaussian process. This implies that $n^{-1}\sum_{i=1}^{n_h}(\xi_{hi}/\alpha_h - 1)(1-\Delta_{hi})Y_{hi}(t)$ converges to 0 in probability uniformly in $t \in [0,\tau]$. Since $n_h^{-1/2}\sum_{i=1}^{n_h}[(1-\Delta_{hi})Y_{hi}(t) - \text{E}\{(1-\Delta_h)Y_h(t)\}]$ is a special case of $n_h^{-1/2}\sum_{i=1}^{n_h}\xi_{hi}[(1-\Delta_{hi})Y_{hi}(t) - \text{E}\{(1-\Delta_h)Y_h(t)\}]$ with $\xi_{hi} = 1$ for all i, by Lemma 3.7.1 it converges weakly to a zero mean Gaussian process. This implies that $n_h^{-1}\sum_{i=1}^{n_h}(1-\Delta_{hi})Y_{hi}(t)$ converges to $\text{E}\{(1-\Delta_h)Y_h(t)\}$ in probability uniformly in t. By Condition (D), $\hat{\alpha_h}(t)$ and α_h converge to the same constant limit C_{2h} uniformly in t. Therefore, $\alpha_h^*(t)$ and α_h also converge to the limit. By Slutsky's theorem,

$$n_h^{1/2}\{\hat{\alpha}_h^{-1}(t) - \alpha_h^{-1}\} = \frac{1}{\alpha_h \text{E}\{(1-\Delta_h)Y_h(t)\}} n_h^{-1/2} \sum_{i=1}^{n_h} \left(1 - \frac{\xi_{hi}}{\alpha_h}\right)(1-\Delta_{hi})Y_{hi}(t) + o_p(1), \quad (4.7)$$

which converges to a zero mean Gaussian process. Since the sampling process is independent across the H strata, $n_h^{1/2}\{\hat{\alpha}_h^{-1}(t) - \alpha_h^{-1}\}$ converges to independent zero-mean Gaussian processes for $h = 1, ..., H$. □

Lemma 4.7.2. *Under Conditions (B) and (C), for any nonzero $d_n \times 1$ constant vector u with $\|u\| = C < \infty$ and $\|u\|_0 = c_n > 0$ where $\|\cdot\|_0$ denotes the number of nonzero components of a vector, $n^{1/2}\{\tilde{S}^{(0)}(\beta,t) - S^{(0)}(\beta,t)\}$ and $(n/c_n)^{1/2}u^T\{\tilde{S}^{(1)}(\beta,t) - S^{(1)}(\beta,t)\}$ converge to tight zero mean Gaussian processes.*

Proof. The two processes can be written in a unified form as the following ($k = 0, 1$),

$$n^{1/2}\left\{n^{-1}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\rho_{hi}(t)Y_{hi}(t)\{c_n^{-1/2}u^T Z_{hi}(t)\}^k - n^{-1}\sum_{h=1}^{H}\sum_{i=1}^{n_h}Y_{hi}(t)\{c_n^{-1/2}u^T Z_{hi}(t)\}^k\right\}$$
$$= n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\{\Delta_{hi} + (1-\Delta_{hi})\xi_{hi}\hat{\alpha}_h(t)^{-1}\}Y_{hi}(t)\{c_n^{-1/2}u^T Z_{hi}(t)\}^k$$

$$
\begin{aligned}
&- n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} (1-\Delta_{hi}) \xi_{hi} \alpha_h^{-1} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \\
&+ n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \Big((1-\Delta_{hi}) \xi_{hi} \alpha_h^{-1} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k - Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \Big) \\
= \; & n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} (1-\Delta_{hi}) \xi_{hi} \hat{\alpha}_h(t)^{-1} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \\
&- n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} (1-\Delta_{hi}) \xi_{hi} \alpha_h^{-1} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \\
&+ n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \Big((1-\Delta_{hi}) \xi_{hi} \alpha_h^{-1} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \\
&- (1-\Delta_{hi}) Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \Big) \\
= \; & n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \{\hat{\alpha}_h(t)^{-1} - \alpha_h^{-1}\} (1-\Delta_{hi}) \xi_{hi} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \\
&- n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left(1 - \frac{\xi_{hi}}{\alpha_h}\right) (1-\Delta_{hi}) Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \\
= \; & \left\{ n^{-1/2} \sum_{h=1}^{H} \frac{1}{\mathrm{E}\{(1-\Delta_h) Y_h(t)\}} \sum_{i=1}^{n_h} \left(1 - \frac{\xi_{hi}}{\alpha_h}\right) (1-\Delta_{hi}) Y_{hi}(t) + o_p(1) \right\} \times \\
&\left\{ \frac{1}{n_h} \sum_{i=1}^{n_h} (1-\Delta_{hi}) \frac{\xi_{hi}}{\alpha_h} Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \right\} \\
&- n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left(1 - \frac{\xi_{hi}}{\alpha_h}\right) (1-\Delta_{hi}) Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k. \qquad (4.8)
\end{aligned}
$$

The last equality holds by (4.7). By Cauchy-Schwarz inequality, $u^T Z_{hi}(t) \le \|u\| \|Z_{hi}(t)\| = C\{\sum_{j=1}^{d_n} Z_{hij}^2(t)\}^{1/2}$. Under Condition (B), $Z_{hij}^2(t)$ has bounded variation, and therefore $c_n^{-1/2} u^T Z_{hi}(t)$ has bounded variation. This along with Condition (C) gives that $(1-\Delta_{hi}) Y_{hi}(t) \{c_n^{-1/2} u^T Z_{hi}(t)\}^k$ is of bounded variation. By Lemma 3.7.1, $n_h^{-1} \sum_{i=1}^{n_h} (1-\Delta_{hi})(\xi_{hi}/\alpha_h) Y_{hi}(t) \{c_n^{-1/2} u$ converges to a deterministic process $L_h(t)$ in probability uniformly in $[0, \tau]$ for $h = 1, ..., H$. Therefore,

$$
\begin{aligned}
(4.8) = \; & n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} (1-\Delta_{hi}) \left(1 - \frac{\xi_{hi}}{\alpha_h}\right) Y_{hi}(t) \left\{ \frac{L_h(t)}{\mathrm{E}\{(1-\Delta_h) Y_h(t)\}} - \{c_n^{-1/2} u^T Z_{hi}(t)\}^k \right\} \\
&+ o_p(1). \qquad (4.9)
\end{aligned}
$$

Under Conditions (B) and (C) the term in the curly braces of (4.9) is of bounded variation. It follows by Lemma 3.7.1 that (4.9) converges weakly to a tight zero mean Gaussian process. Therefore, $n^{1/2}\{\tilde{S}^{(0)}(\beta,t)-S^{(0)}(\beta,t)\}$ and $(n/c_n)^{1/2}u^T\{\tilde{S}^{(1)}(\beta,t)-S^{(1)}(\beta,t)\}$ converge weakly to tight zero mean Gaussian processes. □

Lemma 4.7.3. *Under Conditions (A), (B), and (C), for any nonzero $d_n\times 1$ constant vector u with $\|u\|=1$, $n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}_n'(\beta_0)$ converges to a standard normal distribution, where $\Gamma_n(\beta_0)$ is the covariance matrix of $n^{-1/2}\tilde{\ell}_n'(\beta_0)$.*

Proof. Let $c_n=\|u\|_0$, the number of nonzero components of u. We first consider the quantity $(nc_n)^{-1/2}u^T\tilde{\ell}_n'(\beta_0)$, which can be written as

$$
\begin{aligned}
&(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\left[\{Z_{hi}(t)-\bar{Z}(t)\}dN_{hi}(t)-\{Z_{hi}(t)-\bar{Z}(t)\}\rho_{hi}(t)Y_{hi}(t)\beta_0^TZ_{hi}(t)dt\right]\\
&=(nc_n)^{-1/2}u^T\left\{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}(t)\}dM_{hi}(t)\right.\\
&\quad+\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}(t)\}Y_{hi}(t)\{\lambda_0(t)+\beta_0^TZ_{hi}(t)\}dt\\
&\quad\left.-\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}(t)\}\rho_{hi}(t)Y_{hi}(t)\beta_0^TZ_{hi}(t)dt\right\}\\
&=(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}(t)\}dM_{hi}(t)\\
&\quad+(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}(t)\}\{1-\rho_{hi}(t)\}Y_{hi}(t)\{\lambda_0(t)+\beta_0^TZ_{hi}(t)\}dt\\
&\quad-(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}(t)\}\rho_{hi}(t)Y_{hi}(t)\lambda_0(t)dt\\
&=I_1+I_2+I_3.
\end{aligned}
$$

Let $\bar{Z}_0(t)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}Y_{hi}(t)Z_{hi}(t)/\sum_{h=1}^{H}\sum_{i=1}^{n_h}Y_{hi}(t)$. I_1 can be decomposed as

$$
\begin{aligned}
&(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}_0(t)\}dM_{hi}(t)\\
&\quad+\int_0^\tau c_n^{-1/2}u^T\{\bar{Z}_0(t)-\bar{Z}(t)\}n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}dM_{hi}(t)
\end{aligned}
$$

$$= I_{11} + I_{12}.$$

Let $\mathscr{F}(t^-)$ be the filtration generated by $Y_i(s)$, $N_i(s)$, and $Z_i(s)$ for $s \in [0,t)$. Since $\mathrm{E}\{dN_{hi}(t)|\mathscr{F}(t^-)\} = Y_{hi}(t^-)\{\lambda_0(t^-) + \beta_0^T Z_{hi}(t^-)\}dt$, we have that $dM_{hi}(t)$ is a martingale and therefore $n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h} dM_{hi}(t)$ converges to a tight zero mean Gaussian process, say $G_M(t)$. It follows that I_{11} is a linear combination of a multivariate martingale. By standard martingale theorem (Andersen and Gill 1982) I_{11} converges to a zero mean normal distribution with variance $\Sigma_1(\beta_0) = c_n^{-1}\mathrm{E}\left\{\int_0^\tau u^T\{Z(t) - \bar{Z}_0(t)\}^{\otimes 2} u dN(t)\right\}$.

It can be shown that $\mathrm{E}\{G_M(t) - G_M(s)\}^4 \le C_M(t-s)^2$ for all $0 \le s < t \le \tau$ and some constant C_M. Therefore, by Kolmogorov-Centsov Theorem (Karatzas and Shereve, 1988, p53), $G_M(t)$ has continuous sample path almost surely. $G_M(t)$ is also of bounded variation almost surely. On the other hand, $c_n^{-1/2}u^T\{\bar{Z}_0(t) - \bar{Z}(t)\} = c_n^{-1/2}u^T\{\bar{Z}_0(t) - e(t)\} - c_n^{-1/2}u^T\{\bar{Z}(t) - e(t)\}$. By Lemma 4.7.2 and Slutsky's theorem, both $c_n^{-1/2}u^T\bar{Z}_0(t)$ and $c_n^{-1/2}u^T\bar{Z}(t)$ converge to $c_n^{-1/2}u^T e(t)$ in probability uniformly in t. Moreover, $c_n^{-1/2}u^T\bar{Z}_0(t)$ and $c_n^{-1/2}u^T\bar{Z}(t)$ are of bounded variation almost surely and $c_n^{-1/2}u^T e(t)$ has bounded variation. It then follows from Lemma 3.7.2 that I_{12} converges to 0 in probability. Thus, I_1 converges in distribution to a zero mean normal distribution with variance $\Sigma_1(\beta_0)$.

I_2 can be further decomposed as

$$\begin{aligned}
&(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \{Z_{hi}(t) - \bar{Z}(t)\}\left(1 - \frac{\xi_{hi}}{\alpha_h}\right)(1-\Delta_{hi})Y_{hi}(t)\{\lambda_0(t) + \beta_0^T Z_{hi}(t)\}dt \\
&\quad - (nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \{Z_{hi}(t) - \bar{Z}(t)\}\{\hat{\alpha}_h^{-1}(t) - \alpha_h^{-1}\}\times \\
&\quad \xi_{hi}(1-\Delta_{hi})Y_{hi}(t)\{\lambda_0(t) + \beta_0^T Z_{hi}(t)\}dt \\
&= I_{21} - I_{22}.
\end{aligned}$$

By Lemma 4.7.1, I_{22} can be written as

$$
\begin{aligned}
&(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\left\{\frac{1}{\alpha_h\mathrm{E}\{(1-\Delta_h)Y_h(t)\}}n_h^{-1}\sum_{j=1}^{n_h}\left(1-\frac{\xi_{hj}}{\alpha_h}\right)(1-\Delta_{hj})Y_{hj}(t)+o_p(1)\right\}\times\\
&\quad\{Z_{hi}(t)-\bar{Z}(t)\}\xi_{hi}(1-\Delta_{hi})Y_{hi}(t)\{\lambda_0(t)+\beta_0^TZ_{hi}(t)\}dt\\
&=(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(1-\frac{\xi_{hj}}{\alpha_h}\right)(1-\Delta_{hj})\int_0^\tau\frac{Y_{hj}(t)}{\mathrm{E}\{(1-\Delta_h)Y_h(t)\}}\times\\
&\quad\left\{\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\xi_{hi}}{\alpha_h}(1-\Delta_{hi})\{Z_{hi}(t)-\bar{Z}(t)\}Y_{hi}(t)\{\lambda_0(t)+\beta_0^TZ_{hi}(t)\}\right\}dt+o_p(1)\\
&=(nc_n)^{-1/2}u^T\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(1-\frac{\xi_{hj}}{\alpha_h}\right)(1-\Delta_{hj})\int_0^\tau\frac{Y_{hj}(t)}{\mathrm{E}\{(1-\Delta_h)Y_h(t)\}}\times\\
&\quad\mathrm{E}\left[(1-\Delta_h)\{Z_h(t)-e(t)\}Y_h(t)\{\lambda_0(t)+\beta_0^TZ_h(t)\}\right]dt+o_p(1).
\end{aligned}
$$

The last equality holds by Lemma 3.7.1. Therefore, I_2 is asymptotically equivalent to

$$
n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{\xi_{hi}}{\alpha_h}\right)(1-\Delta_{hi})\int_0^\tau c_n^{-1/2}u^T\left\{R_{hi}(\beta_0,t)-\frac{Y_{hi}(t)\mathrm{E}\{(1-\Delta_h)R_h(\beta_0,t)\}}{\mathrm{E}\{(1-\Delta_h)Y_h(t)\}}\right\}dt, \tag{4.10}
$$

where $R_{hi}(\beta_0,t)=\{Z_{hi}(t)-e(t)\}Y_{hi}(t)\{\lambda_0(t)+\beta_0^TZ_{hi}(t)\}$. Under Condition (A), (B), and (C), the integration in (4.10) is bounded in probability. By Lemma 3.7.1, I_2 converges in distribution to a zero mean normal distribution. Let

$$
W_{hi}(\beta_0)=(1-\Delta_{hi})\int_0^\tau c_n^{-1/2}u^T\left\{R_{hi}(\beta_0,t)-\frac{Y_{hi}(t)\mathrm{E}\{(1-\Delta_h)R_h(\beta_0,t)\}}{\mathrm{E}\{(1-\Delta_h)Y_h(t)\}}\right\}dt.
$$

For a given stratum h, define $\mathscr{F}_h(\tau)$ to be the sigma algebra generated by $Y_{hi}(t)$, $N_{hi}(t)$, and $Z_{hi}(t)$ for $0\le t\le\tau$ and $i=1,...,n$. Conditional on $\mathscr{F}_h(\tau)$, the only random element in I_2 is ξ and $\mathrm{E}\{\xi_h|\mathscr{F}_h(\tau)\}=\alpha_h$. Furthermore, $\mathscr{F}_h(\tau)$ are independent of each other for

$h = 1, ..., H$. Then the asymptotic variance of I_2, denoted by $\Sigma_2(\beta_0)$, can be derived as

$$\begin{aligned}
\Sigma_2(\beta_0) &= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{var}\left\{\left(1-\frac{\xi_{hi}}{\alpha_h}\right)W_{hi}(\beta_0)\right\} \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{E}\left[\mathrm{var}\left\{\left(1-\frac{\xi_{hi}}{\alpha_h}\right)W_{hi}(\beta_0)\middle|\mathscr{F}_h(\tau)\right\}\right] \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{var}\left[\mathrm{E}\left\{\left(1-\frac{\xi_{hi}}{\alpha_h}\right)W_{hi}(\beta_0)\middle|\mathscr{F}_h(\tau)\right\}\right] \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{E}\left[\frac{\mathrm{var}\{\xi_{hi}|\mathscr{F}_h(\tau)\}}{\alpha_h^2}W_{hi}^2(\beta_0)\right] + \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{var}\left[\left(1-\frac{\mathrm{E}\{\xi_{hi}|\mathscr{F}_h(\tau)\}}{\alpha_h}\right)W_{hi}(\beta_0)\right] \\
&= \frac{1}{n}\sum_{h=1}^{H} n_h\mathrm{E}\left[\frac{\alpha_h(1-\alpha_h)}{\alpha_h^2}W_h^2(\beta_0)\right] + 0 \\
&= \sum_{h=1}^{H}\frac{n_h}{n}\frac{1-\alpha_h}{\alpha_h}\mathrm{E}\{W_h^2(\beta_0)\}.
\end{aligned}$$

It is easy to see that $I_3 = 0$. Furthermore, I_1 and I_2 are asymptotically independent of each other since their asymptotic covariance $\Sigma_{12} = 0$. To show this, notice that $\mathrm{E}(I_1) = 0$ and $\mathrm{E}(I_2) = 0$. Define $\mathscr{F}(\tau)$ to be the sigma algebra generated by $Y_{hi}(t)$, $N_{hi}(t)$, and $Z_{hi}(t)$ for $0 \le t \le \tau$, $i = 1, ..., n$, and $h = 1, ..., H$. Conditional on $\mathscr{F}(\tau)$, the only random element is ξ and $\mathrm{E}\{\xi_h|\mathscr{F}(\tau)\} = \alpha_h$.Then

$$\begin{aligned}
\Sigma_{12} &= \mathrm{E}\left[\mathrm{E}\left\{\frac{u^T}{nc_n^{1/2}}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}_0(t)\}dM_{hi}(t)\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{\xi_{hi}}{\alpha_h}\right)W_{hi}(\beta_0)\middle|\mathscr{F}(\tau)\right\}\right] \\
&= \mathrm{E}\left[\frac{u^T}{nc_n^{1/2}}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau\{Z_{hi}(t)-\bar{Z}_0(t)\}dM_{hi}(t)\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(1-\frac{\mathrm{E}\{\xi_{hi}|\mathscr{F}(\tau)\}}{\alpha_h}\right)W_{hi}(\beta_0)\right] \\
&= 0.
\end{aligned}$$

Taken the above results together, $(nc_n)^{-1/2}u^T\tilde{\ell}_n'(\beta_0)$ converges to a zero mean normal distribution with variance $\Sigma(\beta_0) = \Sigma_1(\beta_0) + \Sigma_2(\beta_0)$. Now define a vector $u^* := u^T\Gamma_n^{-1/2}(\beta_0)\|u^T\Gamma_n^{-1/2}(\beta_0)\|^{-1}$. Let $c_n^* = \|u^*\|_0$. Then $n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}_n'(\beta_0) = \|u^T\Gamma_n^{-1/2}(\beta_0)\|(c_n^*)^{1/2}(nc_n^*)^{-1/2}($ Since $\|u^*\| = 1$, the above quantity converges to a zero mean normal distribution up to a

scalar by previous derivation. Since $\Gamma_n(\beta_0) = \text{var}\{n^{-1/2}\tilde{\ell}'_n(\beta_0)\}$ and $\|u\| = 1$, we have

$$\text{var}\{n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}'_n(\beta_0)\} = u^T\Gamma_n^{-1/2}(\beta_0)\text{var}\{n^{-1/2}\tilde{\ell}'_n(\beta_0)\}\Gamma_n^{-1/2}(\beta_0)u = 1.$$

Therefore, $n^{-1/2}u^T\Gamma_n^{-1/2}(\beta_0)\tilde{\ell}'_n(\beta_0)$ converges to a standard normal distribution. □

Lemma 4.7.4. *Under Conditions (A), (B), and (C), $\tilde{\ell}''_n(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk}$ is $O_p(n^{1/2})$ for $j,k = 1,...,d_n$, where $\mathscr{A}_n(\beta_0)_{jk}$ is the (j,k) component of $\mathscr{A}_n(\beta_0)$ as defined in the Notations and Regularity Conditions section.*

Proof. Let $Z_{hi}(t)_j$, $\tilde{S}^{(1)}(t)_j$, $\bar{Z}(t)_j$, $s^{(1)}(t)_j$, and $e(t)_j$ be the j^{th} component of the corresponding vectors. Define $\bar{\mathbb{Z}}_{hi}(t)_{jk} = \{Z_{hi}(t)_j - \bar{Z}(t)_j\}\{Z_{hi}(t)_k - \bar{Z}(t)_k\}$ and $\mathbb{E}(t)_{jk} = \{Z_{hi}(t)_j - e(t)_j\}\{Z_{hi}(t)_k - e(t)_k\}$. Then $n^{-1/2}\{\tilde{\ell}''_n(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk}\}$ can be written as

$$\begin{aligned}
&n^{-1/2}\{\tilde{\ell}''_n(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk}\} \\
&= n^{-1/2}\left\{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \rho_{hi}(t)\bar{\mathbb{Z}}_{hi}(t)_{jk}Y_{hi}(t)dt - n\text{E}\left(\int_0^\tau \mathbb{E}(t)_{jk}Y_{hi}(t)dt\right)\right\} \\
&= n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \left\{\rho_{hi}(t)\bar{\mathbb{Z}}_{hi}(t)_{jk} - \mathbb{E}(t)_{jk}\right\}Y_{hi}(t)dt \\
&\quad + n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left\{\int_0^\tau \mathbb{E}(t)_{jk}Y_{hi}(t)dt - \text{E}\left(\int_0^\tau \mathbb{E}(t)_{jk}Y_{hi}(t)dt\right)\right\} \\
&= I_1 + I_2.
\end{aligned}$$

We further decompose I_1 as

$$\begin{aligned}
I_1 &= n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \{\rho_{hi}(t) - 1\}\bar{\mathbb{Z}}_{hi}(t)_{jk}Y_{hi}(t)dt \\
&\quad + n^{-1/2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\int_0^\tau \left\{\bar{\mathbb{Z}}_{hi}(t)_{jk} - \mathbb{E}(t)_{jk}\right\}Y_{hi}(t)dt \\
&= I_{11} + I_{12}.
\end{aligned}$$

The term I_{11} can be written as

$$\begin{aligned} I_{11} &= n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau (1-\Delta_{hi}) \left(\frac{\xi_{hi}}{\hat{\alpha}_h(t)} - 1 \right) \bar{\mathbb{Z}}_{hi}(t)_{jk} Y_{hi}(t) dt \\ &= n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau (1-\Delta_{hi}) \left(\frac{\xi_{hi}}{\alpha_h} - 1 \right) \bar{\mathbb{Z}}_{hi}(t)_{jk} Y_{hi}(t) dt \\ &\quad + n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \int_0^\tau (1-\Delta_{hi}) \{ \hat{\alpha}_h^{-1}(t) - \alpha_h^{-1} \} \xi_{hi} \bar{\mathbb{Z}}_{hi}(t)_{jk} Y_{hi}(t) dt. \end{aligned}$$

By Lemma 4.7.1 and following similar derivation as for I_2 in the proof of Lemma 4.7.3, we have

$$\begin{aligned} I_{11} = n^{-1/2} \sum_{h=1}^{H} \sum_{i=1}^{n_h} (1-\Delta_{hi}) \left(\frac{\xi_{hi}}{\alpha_h} - 1 \right) \times \\ \int_0^\tau \left\{ \bar{\mathbb{Z}}_{hi}(t)_{jk} Y_{hi}(t) - \frac{\mathrm{E}\left\{ (1-\Delta_h) \bar{\mathbb{Z}}_h(t)_{jk} Y_h(t) \right\} Y_{hi}(t)}{\mathrm{E}\{(1-\Delta_h) Y_h(t)\}} \right\} dt. \end{aligned}$$

Since the integration of the above expression is bounded in probability under Conditions (A) and (B), by Lemma 3.7.1 I_{11} converges to a zero mean normal distribution. Thus, $I_{11} = O_p(1)$.

Now we consider I_{12}. We first show that $\bar{Z}(t)_j - e(t)_j$ is $O_p(n^{-1/2})$ for $j = 1, ..., d_n$.

$$\begin{aligned} \bar{Z}(t)_j - e(t)_j &= \frac{\tilde{S}^{(1)}(t)_j}{\tilde{S}^{(0)}(t)_j} - \frac{s^{(1)}(t)_j}{s^{(0)}(t)_j} \\ &= \frac{\{\tilde{S}^{(1)}(t)_j - s^{(1)}(t)_j\} s^{(0)}(t) - \{\tilde{S}^{(0)}(t) - s^{(0)}(t)\} s^{(1)}(t)_j}{\tilde{S}^{(0)}(t) s^{(0)}(t)}. \end{aligned}$$

By Lemma 4.7.2 with $u_k = I(k = j)$ for $k = 1, ..., d_n$, we have that $\tilde{S}^{(1)}(t)_j - s^{(1)}(t)_j$ and $\tilde{S}^{(0)}(t) - s^{(0)}(t)$ are both $O_p(n^{-1/2})$. Under Condition (C), it follows that $\bar{Z}(t)_j - e(t)_j$ is $O_p(n^{-1/2})$. Therefore,

$$\bar{\mathbb{Z}}_{hi}(t)_{jk} - \mathbb{E}(t)_{jk}$$

$$
\begin{aligned}
&= \{Z_{hi}(t)_j - \bar{Z}(t)_j\}\{Z_{hi}(t)_k - \bar{Z}(t)_k\} - \{Z_{hi}(t)_j - e(t)_j\}\{Z_{hi}(t)_k - e(t)_k\} \\
&= \{Z_{hi}(t)_j Z_{hi}(t)_k - Z_{hi}(t)_j \bar{Z}(t)_k - \bar{Z}(t)_j Z_{hi}(t)_k + \bar{Z}(t)_j \bar{Z}(t)_k\} \\
&\quad - \{Z_{hi}(t)_j Z_{hi}(t)_k - Z_{hi}(t)_j e(t)_k - e(t)_j Z_{hi}(t)_k + e(t)_j e(t)_k\} \\
&= -Z_{hi}(t)_j\{\bar{Z}(t)_k - e(t)_k\} - Z_{hi}(t)_k\{\bar{Z}(t)_j - e(t)_j\} + \bar{Z}(t)_j\{\bar{Z}(t)_k - e(t)_k\} \\
&\quad + e(t)_k\{\bar{Z}(t)_j - e(t)_j\} \\
&= O_p(n^{-1/2}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\int_0^\tau \left\{\bar{\mathbb{Z}}_{hi}(t)_{jk} - \mathbb{E}(t)_{jk}\right\} Y_{hi}(t)dt \le \int_0^\tau \sup_{t\epsilon[0,\tau]} \left|\bar{\mathbb{Z}}_{hi}(t)_{jk} - \mathbb{E}(t)_{jk}\right| dt \\
&= \sup_{t\epsilon[0,\tau]} \left|\bar{\mathbb{Z}}_{hi}(t)_{jk} - \mathbb{E}(t)_{jk}\right| \tau = O_p(n^{-1/2}).
\end{aligned}
$$

It follows that $I_{12} = O_p(1)$, and therefore $I_1 = O_p(1)$.

By central limit theorem, we have that I_2 is $O_p(1)$. Taken the above results together, we conclude that $n^{-1/2}\{\tilde{\ell}_n''(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk}\} = O_p(1)$, which implies $\tilde{\ell}_n''(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk} = O_p(n^{1/2})$. □

Proof of Theorem 4.3.1. Let β_0 be the true parameters, and $\alpha_n = d_n^{1/2}(n^{-1/2} + a_n)$. It suffices to show that, for any given $\varepsilon > 0$, there exists a constant vector u and a large enough constant C such that $\mathrm{pr}\{\inf_{\|u\|=C} \tilde{Q}_n(\beta_0 + \alpha_n u) > \tilde{Q}_n(\beta_0)\} \ge 1 - \varepsilon$. This implies that there exists a local minimizer $\hat{\beta}$ such that $\|\hat{\beta} - \beta_0\| = O_p(\alpha_n)$. Since $P_{\lambda_{jn}}(0) = 0$ and $P_{\lambda_{jn}}(\cdot) \ge 0$,

$$
\begin{aligned}
\tilde{Q}_n(\beta_0 + \alpha_n u) - \tilde{Q}_n(\beta_0) &\ge \{\tilde{L}_n(\beta_0 + \alpha_n u) - \tilde{L}_n(\beta_0)\} + n\sum_{j=1}^{k_n}\{P_{\lambda_{jn}}(|\beta_{j0} + \alpha_n u|) - P_{\lambda_{jn}}(|\beta_{j0}|)\} \\
&= I_1 + I_2.
\end{aligned}
$$

By Taylor expansion,

$$I_1 = \alpha_n u^T \tilde{\ell}'_n(\beta_0) + \frac{1}{2}\alpha_n^2 u^T \tilde{\ell}''_n(\beta_0) u = I_{11} + I_{12}.$$

By Lemma 4.7.3 we have $\tilde{\ell}'_n(\beta_0)_j = O_p(n^{1/2})$ for $j = 1, ..., d_n$. Therefore,

$$|I_{11}| = |\alpha_n u^T \tilde{\ell}'_n(\beta_0)| \leq \alpha_n \|u\| \|\tilde{\ell}'_n(\beta_0)\| = \alpha_n \|u\| O_p(d_n^{1/2} n^{1/2}) = \|u\| O_p(d_n^{1/2} n^{-1/2} \alpha_n n)$$
$$= \|u\| O_p(\alpha_n^2 n).$$

The term I_{12} can be written as

$$I_{12} = \frac{1}{2}\alpha_n^2 u^T \{\tilde{\ell}''_n(\beta_0) - n\mathscr{A}_n(\beta_0)\} u + \frac{1}{2}\alpha_n^2 u^T n\mathscr{A}_n(\beta_0) u = J_1 - J_2.$$

By Lemma 4.7.4, Cauchy-Schwarz inequality, and the fact that $d_n^2/n \to 0$,

$$|J_1| \leq \frac{1}{2}\alpha_n^2 \|u\|^2 \|\tilde{\ell}''_n(\beta_0) - n\mathscr{A}_n(\beta_0)\| = \|u\|^2 O_p(\alpha_n^2 n^{1/2} d_n) = \|u\|^2 o_p(\alpha_n^2 n).$$

By spectral decomposition of $\mathscr{A}_n(\beta_0)$ and Condition (E)

$$|J_2| \geq \frac{1}{2}\alpha_n^2 \|u\|^2 n \text{eigen}_{\min}\{\mathscr{A}_n(\beta_0)\} \geq \|u\|^2 (\alpha_n^2 n) \frac{C_3}{2}.$$

Then $|I_{12}| \geq |J_2| - |J_1| \geq \|u\|^2 (\alpha_n^2 n) C_3/2 - \|u\|^2 o_p(\alpha_n^2 n)$ as $n \to \infty$. Therefore, for large enough $\|u\|$, $|I_{12}|$ dominates $|I_{11}|$.

We now consider I_2. By Taylor expansion and Cauchy-Schwarz inequality

$$|I_2| = \left| n \sum_{j=1}^{k_n} P'_{\lambda_{jn}}(|\beta_{j0}|) \text{sgn}(\beta_{j0}) \alpha_n u_j + \frac{1}{2} n \sum_{j=1}^{k_n} P''_{\lambda_{jn}}(|\beta_{j0}|) \alpha_n^2 u_j^2 \{1 + o(1)\} \right|$$
$$\leq n \left| \sum_{j=1}^{k_n} P'_{\lambda_{jn}}(|\beta_{j0}|) \alpha_n u_j \right| + \frac{1}{2} n \left| \sum_{j=1}^{k_n} P''_{\lambda_{jn}}(|\beta_{j0}|) \alpha_n^2 u_j^2 \{1 + o(1)\} \right|$$

$$\leq n\left|\sum_{j=1}^{k_n}\phi_n\alpha_n u_j\right| + \frac{1}{2}n\left|\sum_{j=1}^{k_n}\psi_n\alpha_n^2 u_j^2\{1+o(1)\}\right|$$
$$\leq n\alpha_n\phi_n k_n^{1/2}\|u\| + \frac{1}{2}n\alpha_n^2\psi_n\|u\|^2\{1+o(1)\}$$
$$= \|u\|O_p(\alpha_n^2 n).$$

The last equality holds because $\phi_n = O_p(\alpha_n d_n^{-1/2})$ and $\psi_n \to 0$ under Condition (G). Therefore, $|I_{12}|$ also dominates $|I_2|$ for large enough C. By Condition (E) I_{12} is positive as $n \to \infty$, it follows that for large enough C, $\tilde{Q}_n(\beta_0 + \alpha_n u) - \tilde{Q}_n(\beta_0)$ is positive with probability tending to one as $n \to \infty$. □

The following Lemma proves that the SCAD-penalized estimator must possess the sparsity property $\hat{\beta}_{II} = 0$ with probability tending to one.

Lemma 4.7.5. *Under conditions (A)-(H), as $n \to \infty$, if $d_n^2/n \to 0, \lambda_{jn} \to 0$, and $\lambda_{jn} n^{1/2} d_n^{-1/2} \to \infty$, with probability tending to one, for any given β_I satisfying $\|\beta_I - \beta_{I0}\| = O_p(d_n^{1/2} n^{-1/2})$ and any constant C,*

$$\tilde{Q}_n\{(\beta_I^T, 0)^T\} = \min_{\|\beta_{II}\| \leq C d_n^{1/2} n^{-1/2}} \tilde{Q}_n\{(\beta_I^T, \beta_{II}^T)^T\}.$$

Proof. It suffices to show that with probability tending to one as $n \to \infty$, for any β_I satisfying $\|\beta_I - \beta_{I0}\| = O_p(d_n^{1/2} n^{-1/2})$ and $\|\beta_{II}\| \leq C d_n^{1/2} n^{-1/2}$, $\partial \tilde{Q}_n(\beta)/\partial\beta_j$ and β_j have the same signs for $j = k_n + 1, ..., d_n$. By Taylor expansion,

$$\frac{\partial \tilde{Q}_n(\beta)}{\partial \beta_j} = \tilde{\ell}_n'(\beta_0)_j + \sum_{k=1}^{d_n}\tilde{\ell}_n''(\beta_0)_{jk}(\beta_k - \beta_{k0}) + nP'_{\lambda_{jn}}(|\beta_j|)\mathrm{sgn}(\beta_j) = I_1 + I_2 + I_3.$$

From Lemma 4.7.3 we have $I_1 = O_p(n^{1/2})$. The term I_2 can be written as

$$I_2 = \sum_{k=1}^{d_n}\left\{\tilde{\ell}_n''(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk}\right\}(\beta_k - \beta_{0k}) + \sum_{k=1}^{d_n} n\mathscr{A}_n(\beta_0)_{jk}(\beta_k - \beta_{0k}) = I_{21} - I_{22}.$$

From Lemma 4.7.4 we have $\tilde{\ell}_n''(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk} = O_p(n^{1/2})$ for $j, k = 1, ..., d_n$. Using

Cauchy-Schwarz inequality along with $\|\beta - \beta_0\| = O_p(d_n^{1/2} n^{-1/2})$,

$$|I_{21}| \leq \|\beta - \beta_0\| \left[\sum_{k=1}^{d_n} \left\{ \tilde{\ell}_n''(\beta_0)_{jk} - n\mathscr{A}_n(\beta_0)_{jk} \right\}^2 \right]^{1/2} = O_p(d_n) = o_p(d_n^{1/2} n^{1/2}).$$

As $\text{eigen}_{\max}\{\mathscr{A}_n(\beta_0)\}$ is bounded by Condition (E), it follows that

$$|I_{22}| \leq n\|\beta - \beta_0\| \left\{ \sum_{k=1}^{d_n} \mathscr{A}_n^2(\beta_0)_{jk} \right\}^{1/2} = nO_p(d_n^{1/2} n^{-1/2}) O(1) = O_p(d_n^{1/2} n^{1/2}).$$

It follows that $|I_2| = O_p(d_n^{1/2} n^{1/2})$. Therefore, $I_1 + I_2 = O_p(d_n^{1/2} n^{1/2})$. Hence,

$$\frac{\partial \tilde{Q}_n(\beta)}{\partial \beta_j} = nP'_{\lambda_{jn}}(|\beta_j|)\text{sgn}(\beta_j) + O_p(\sqrt{d_n n}) = n\lambda_{jn} \left\{ \frac{P'_{\lambda_{jn}}(|\beta_j|)}{\lambda_{jn}} \text{sgn}(\beta_j) + O_p\left(\frac{d_n^{1/2} n^{-1/2}}{\lambda_{jn}} \right) \right\}.$$

For $j = (k_n + 1), ..., d_n$, since $|\beta_j| = O\{d_n^{1/2} n^{-1/2}\}$ and $\lambda_{jn} d_n^{-1/2} n^{1/2} \to \infty$, the quantity $P'_{\lambda_{jn}}(|\beta_j|)/\lambda_{jn}$ is positive under Condition (F) for all sufficiently large n. Therefore, the quantity in the curly brackets is positive with probability tending to one. Thus, $\partial \tilde{Q}_n(\beta)/\partial \beta_j$ and β_j have the same signs with probability tending to one as $n \to \infty$. □

Proof of Theorem 4.3.2. Part (i) follows directly from Lemma 4.7.5. To prove assertion (ii), we first show that

$$\begin{aligned} & n^{1/2} u^T \Gamma_{n11}^{-1/2} (\mathscr{A}_{n11} + \Psi_n)(\hat{\beta}_I - \beta_{I0})(1 + o_p(1)) + n^{1/2} u^T \Gamma_{n11}^{-1/2} \Phi_n \\ & = -n^{-1/2} u^T \Gamma_{n11}^{-1/2} \tilde{\ell}'_{n1}(\beta_0) + o_p(1), \end{aligned} \tag{4.11}$$

where $\tilde{\ell}'_{n1}(\beta_0)$ consists of the first k_n components of $\tilde{\ell}'_n(\beta_0)$ and $\mathscr{A}_{n11}(\beta_0)$ is the first $k_n \times k_n$ components of $\mathscr{A}_n(\beta_0)$. Since $\hat{\beta}_I$ is the minimizer of $\tilde{Q}_n(\beta)$, we have $\partial \tilde{Q}_n(\hat{\beta})/\partial \beta_I = 0$. By Taylor expansion of $\partial \tilde{Q}_n(\hat{\beta})/\partial \beta_I$ at β_{I0} and the fact that $\hat{\beta}_{II} - \beta_{II0} = 0$,

$$\tilde{\ell}'_{n1}(\beta_0) + \tilde{\ell}''_{n1}(\beta_0)(\hat{\beta}_I - \beta_{I0}) + n\Phi_n + n\Psi_n^*(\hat{\beta}_I - \beta_{I0}) = 0,$$

where $\tilde{\ell}''_{n1}(\beta_0)$ consists of the first $k_n \times k_n$ components of $\tilde{\ell}''_n(\beta_0)$, β^* lies between $\hat{\beta}$ and β_0, $\Psi_n^* = \Psi_n(\beta^*)$, β^* lies between $\hat{\beta}$ and β_0. Rearrange the above equation we have,

$$\{\tilde{\ell}''_{n1}(\beta_0) + n\Psi_n^*\}(\hat{\beta}_I - \beta_{I0}) + n\Phi_n = -\tilde{\ell}'_{n1}(\beta_0). \tag{4.12}$$

Multiply both sides of (4.12) by $n^{-1/2}u^T\Gamma_{n11}^{-1/2}$,

$$n^{1/2}u^T\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde{\ell}''_{n1}(\beta_0) + \Psi_n^*\right\}(\hat{\beta}_I - \beta_{I0}) + n^{1/2}u^T\Gamma_{n11}^{-1/2}\Phi_n = -n^{-1/2}u^T\Gamma_{n11}^{-1/2}\tilde{\ell}'_{n1}(\beta_0). \tag{4.13}$$

The quantity $u^T\Gamma_{n11}^{-1/2}n^{-1}\tilde{\ell}''_{n1}(\beta_0)(\hat{\beta}_I - \beta_{I0})$ can be written as,

$$u^T\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde{\ell}''_{n1}(\beta_0) - \mathscr{A}_{n11}(\beta_0)\right\}(\hat{\beta}_I - \beta_{I0}) + u^T\Gamma_{n11}^{-1/2}\mathscr{A}_{n11}(\beta_0)(\hat{\beta}_I - \beta_{I0}) = I_1 + I_2.$$

By Cauchy-Schwarz inequality and Lemma 4.7.4,

$$|I_1| \le \|u^T\Gamma_{n11}^{-1/2}\|\left\|\frac{1}{n}\tilde{\ell}''_{n1}(\beta_0) - \mathscr{A}_{n11}(\beta_0)\right\|\|\hat{\beta}_I - \beta_{I0}\| = \|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|O_p(d_n n^{-1/2}).$$

By spectral decomposition of $\mathscr{A}_{n11}$,

$$I_2 \ge \|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|\text{eigen}_{\min}(\mathscr{A}_{n11}) \ge \|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|\text{eigen}_{\min}(\mathscr{A}_n).$$

Therefore, by Condition (E) and $d_n^2/n \to 0$ we have

$$\left|\frac{I_1}{I_2}\right| \le \frac{\|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|O_p(d_n n^{-1/2})}{\|u^T\Gamma_{n11}^{-1/2}\|\|\hat{\beta}_I - \beta_{I0}\|\text{eigen}_{\min}(\mathscr{A}_n)} = O_p(d_n n^{-1/2}) = o_p(1).$$

Therefore, $I_1 = o_p(I_2)$, and $u^T\Gamma_{n11}^{-1/2}n^{-1}\tilde{\ell}''_{n1}(\beta_0)(\hat{\beta}_I - \beta_{I0}) = u^T\Gamma_{n11}^{-1/2}\mathscr{A}_{n11}(\beta_0)(\hat{\beta}_I - \beta_{I0})\{1 +$

$o_p(1)\}$. Since $\hat{\beta}$ converges to β_0 in probability, Ψ_n^* converges to Ψ_n in probability. Therefore

$$u^T\Gamma_{n11}^{-1/2}\left\{\frac{1}{n}\tilde{\ell}''_{n1}(\beta_0)+\Psi_n^*\right\}(\hat{\beta}_I-\beta_{I0})=u^T\Gamma_{n11}^{-1/2}\{\mathscr{A}_{n11}(\beta_0)+\Psi_n\}(\hat{\beta}_I-\beta_{I0})\{1+o_p(1)\}. \tag{4.14}$$

By (4.13) and (4.14), we have that (4.11) holds.

By Lemma 4.7.3, $n^{-1/2}u^T\Gamma_{n11}^{-1/2}\tilde{\ell}'_{n1}(\beta_0)$ converges to the standard normal distribution. Thus,

$$n^{1/2}u^T\Gamma_{n11}^{-1/2}(\beta_0)\{\mathscr{A}_{n11}(\beta_0)+\Psi_n\}\{\hat{\beta}_I-\beta_{I0}+(\mathscr{A}_{n11}(\beta_0)+\Psi_n)^{-1}\Phi_n\}\to N(0,1)$$

in distribution. □

CHAPTER 5: TUNING PARAMETER SELECTION FOR REGULARIZED VARIABLE SELECTION UNDER COX PROPORTIONAL HAZARDS MODEL

5.1 Introduction

In the first two topics of the dissertation, we have shown that the SCAD-penalized variable selection procedure can identify the true model with probability tending to one as the sample size goes to infinity under Cox proportional hazards model and additive hazards model with a case-cohort design. This result implies that with probability approaching one the true model is contained in the solution path of the tuning parameter λ. If one can select the correct tuning parameter λ_0, then one will be able to identify the true model. However, the theorems developed in the first two topics do not offer any theoretical insight to the tuning parameter selection methods used there (AIC- and BIC-based method). Wang et al. (2007) studied the asymptotic properties of the two tuning parameter selection methods in linear models. Zhang et al. (2010) proposed a new tuning parameter selection criterion in generalized linear models. Wang et al. (2009) extended the investigation on tuning parameter selection to linear models with a diverging number of parameters. More recently, Fan and Tang (2013) studied tuning parameter selection in generalized linear model with ultra-high dimension. To the best of our knowledge, a consistent tuning parameter selection method for regularized variable selection has not been established for Cox proportional hazards model with a diverging number of parameters. In this chapter of the dissertation we focus on regular Cox model without the case-cohort design and propose a tuning parameter selection criterion that consistently identifies the true model. We theoretically prove its asymptotic properties and empirically demonstrate

its finite sample performance via simulation. We then apply the proposed method to the Framingham Heart Study (Dawber 1980).

5.2 Tuning Parameter Selection Criterion under Cox Proportional Hazards Model

Suppose there are n subjects in the dataset. 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 $\Delta = I(T \leq C)$ be the censoring indicator, where $I(\cdot)$ is an indicator function. Let $Z_i(t)$ be the $d_n \times 1$ possibly time-dependent covariate vector for subject i at time t, where d_n goes to infinity with the sample size n. T and C are assumed to be independent conditional on Z. Let $\beta = (\beta_1, ..., \beta_{d_n})^T$ be a vector of unknown regression coefficients and $\beta_0 = (\beta_{01}, ..., \beta_{0d_n})^T$ be its true value. Without loss of generality, assume the first k_n components of β_0 is nonzero and the other components of β_0 are zero. Hence, k_n is the size of the true model, which is allowed to go to infinity with sample size and k_n/d_n converges to a constant $c \in [0, 1]$. Define for subject i the counting process $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, and the at risk process $Y_i(t) = I(X_i \geq t)$. The partial likelihood under Cox proportional hazards model is

$$\ell_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta^T Z_i(t) - \log \sum_{j=1}^{n} Y_j(t) \exp\{\beta^T Z_j(t)\} \right] dN_i(t), \tag{5.1}$$

where τ is the time at the end of study. Let $P_{\lambda_{jn}}(\cdot)$ be the SCAD penalty function with tuning parameter λ_{jn}. For notational simplicity, we suppress the subscript n for λ_{jn} and assume it is the same for all parameters. The SCAD-penalized maximum partial likelihood estimator $\hat{\beta}_\lambda$ is the maximizer of the following objective function,

$$\ell_n(\beta) - n \sum_{j=1}^{d_n} P_\lambda(|\beta_j|). \tag{5.2}$$

Let α_λ be the model that is identified by the tuning parameter λ . Let $|\alpha_\lambda|$ be the size of model α_λ. We propose the generalized information criterion

$$\mathrm{GIC}(\lambda) = \frac{1}{n}\{-\ell_n(\hat{\beta}_\lambda) + a_n|\alpha_\lambda|\}, \tag{5.3}$$

where a_n is a positive sequence depending on n. When $a_n = 2$, GIC becomes the AIC statistic. When $a_n = \log(n)$, GIC becomes the BIC statistic. The selected tuning parameter $\hat{\lambda}$ is the minimizer of (5.3). We have shown in the previous chapters that there exists one or a range of λ that gives rise to the true model α_0. Note that $|\alpha_0| = k_n$. Our goal in this chapter is to determine the characteristic of the sequence a_n in (5.3) so that the λ that gives the true model is identified with probability tending to one as sample size goes to infinity.

5.3 Notations and Regularity Conditions

Denote for any model α_λ the penalized maximum partial likelihood estimator and the unpenalized maximum partial likelihood estimator as $\hat{\beta}_{\alpha_\lambda}$ and $\hat{\beta}_{\alpha_\lambda}$, respectively. Define $\beta^0_{\alpha_0}$ as the true parameter under the true model. Similar to Fan and Tang (2013), for any model α_λ, we define its "population parameter" $\beta^0_{\alpha_\lambda}$ to be the minimizer of the Kullback-Leibler distance $D_{KL}(\beta_{\alpha_\lambda}) := n^{-1}\mathrm{E}_{\beta^0_{\alpha_0}}\{\ell_n(\beta^0_{\alpha_0}) - \ell_n(\beta_{\alpha_\lambda})\}$, where $\ell_n(\beta^0_{\alpha_0})$ and $\ell_n(\beta_{\alpha_\lambda})$ are the partial likelihood defined in (5.1) under model α_0 and α_λ, respectively. The expectation is taken under the true model with respect to all random variables.

We define the following notations for each n:

$$S_n^{(k)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)Z_i(t)^{\otimes k}e^{\beta' Z_i(t)}, \quad s_n^{(k)}(\beta, t) = \mathrm{E}\{S_n^{(k)}(\beta, t)\}, \quad k = 0, 1, 2,$$

$$e_n(\beta, t) = \frac{s_n^{(1)}(\beta, t)}{s_n^{(0)}(\beta, t)}, \quad V_n(\beta, t) = \frac{S_n^{(2)}(\beta, t)S_n^{(0)}(\beta, t) - S_n^{(1)}(\beta, t)^{\otimes 2}}{S_n^{(0)}(\beta, t)^2},$$

$$I_n(\beta) = -\frac{1}{n}\mathrm{E}\left\{\frac{\partial^2 \ell_n(\beta)}{\partial \beta^2}\right\} = \mathrm{E}\left\{\int_0^\tau V_n(\beta,t) S_n^{(0)}(\beta,t) d\Lambda_0(t)\right\}.$$

We require the following regularity conditions for theoretical derivations in this chapter:

(A) $\int_0^\tau \lambda_0(t)dt < \infty$.

(B) $\mathrm{E}\{Y(\tau)\} > 0$.

(C) $| Z_{ij}(0) | + \int_0^\tau |dZ_{ij}(t)| < C_1 < \infty$ almost surely for some constant C_1 and $i = 1, ..., n$ and $j = 1, ..., d_n$. That is, $Z_{ij}(t)$ has bounded variation almost surely. This implies that $|Z_{ij}(t)|$ is bounded almost surely. Define $K_n := \max_{1\le j\le d_n, 1\le i\le n} \|Z_{ij}(t)\|_\infty < \infty$.

(D) For any model α_λ, there exists a neighborhood $\mathscr{B}_{\alpha_\lambda}$ of $\beta^0_{\alpha_\lambda}$ such that for all $\beta_{\alpha_\lambda} \in \mathscr{B}_{\alpha_\lambda}$ and $t \in [0,\tau]$, $\partial s_n^{(0)}(\beta_{\alpha_\lambda}, t)/\partial \beta_{\alpha_\lambda} = s_n^{(1)}(\beta_{\alpha_\lambda}, t)$, and $\partial^2 s_n^{(0)}(\beta_{\alpha_\lambda}, t)/\partial \beta_{\alpha_\lambda} \partial \beta_{\alpha_\lambda}^T = s_n^{(2)}(\beta_{\alpha_\lambda}, t)$. The functions $s_n^{(k)}(\beta_{\alpha_\lambda}, t)$ $(k = 0, 1, 2)$ are continuous and bounded and $s_n^{(0)}(\beta_{\alpha_\lambda}, t)$ is bounded away from 0 on $\mathscr{B}_{\alpha_\lambda} \times [0,\tau]$.

(E) For any model α_λ, there exists a neighborhood $\mathscr{B}_{\alpha_\lambda}$ of $\beta^0_{\alpha_\lambda}$ such that for all $\beta_{\alpha_\lambda} \in \mathscr{B}_{\alpha_\lambda}$, there exist positive constants C_3, C_4 such that

$$0 < C_3 < \mathrm{eigen}_{\min}\{I_n(\beta_{\alpha_\lambda})\} \le \mathrm{eigen}_{\max}\{I_n(\beta_{\alpha_\lambda})\} < C_4 < \infty,$$

where $\mathrm{eigen}_{\min}\{\cdot\}$ and $\mathrm{eigen}_{\max}\{\cdot\}$ are the minimum and maximum eigenvalues of a matrix.

(F) $\min_{1\le j\le k_n} |\beta_{0j}|/\lambda_0 \to \infty$ as $n \to \infty$.

(G) $\liminf_{n\to+\infty} \liminf_{\theta\to 0+} P'_{\lambda_0}(\theta)/\lambda_0 > 0$ for $j = 1, ..., d_n$.

(H) $d_n^5/n \to 0$ and $k_n/d_n \to c \in [0, 1)$ as $n \to \infty$.

(I) $L_n := \|\beta_0\|_1 < \infty$, where $\|\cdot\|_1$ denotes the L_1 norm. As a consequence of this condition and Condition (C), we can define $\exp(|\beta_0^T Z_i(t)|) \leq \exp(K_n L_n) := U_n < \infty$ for $i = 1, ..., n$.

5.4 Asymptotic Properties of the Generalized Information Criterion

Let $\lambda_{\max}$ be the smallest λ that results in an empty model (i.e. a model with no non-zero parameters). We partition the tuning parameter space $\Omega = [0, \lambda_{\max}]$ into the underfit, true, and overfit subspaces as follows,

$$\Omega_- = \{\lambda : \alpha_\lambda \not\supset \alpha_0\}, \quad \Omega_0 = \{\lambda : \alpha_\lambda = \alpha_0\}, \quad \Omega_+ = \{\lambda : \alpha_\lambda \not\supseteq \alpha_0\},$$

where $a \not\supseteq b$ means a contains b but is not equal to b. Since $\hat{\beta}_\lambda$ is the maximizer of the nonconcave objective function (5.2), the asymptotic property of $\ell_n(\hat{\beta}_\lambda)$ is difficult to study. Instead, we work with the unpenalized version of the likelihood. Define an approximation of $\mathrm{GIC}(\lambda)$ as

$$\mathrm{GIC}^*(\alpha_\lambda) = \frac{1}{n}\left\{-\ell_n(\hat{\beta}_{\alpha_\lambda}) + a_n|\alpha_\lambda|\right\}.$$

Note that $\mathrm{GIC}(\lambda)$ is a function of the tuning parameter whereas $\mathrm{GIC}^*(\alpha_\lambda)$ is a function of the model.

We only present main results in this section. The proof of the lemmas and theorems presented in this section can be found in Section 5.7. The following lemma states that the difference between $\mathrm{GIC}(\lambda)$ and $\mathrm{GIC}(\lambda_0)$ is no less than that between $\mathrm{GIC}^*(\alpha_\lambda)$ and $\mathrm{GIC}^*(\alpha_0)$ for any λ with probability tending to one as sample size goes to infinity.

Lemma 5.4.1. *Under Conditions (A) to (H), for any* $\lambda \in \Omega$, $\mathrm{pr}\{\mathrm{GIC}(\lambda) - \mathrm{GIC}(\lambda_0) \geq \mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)\} \to 1$ *as* $n \to \infty$.

Lemma 5.4.1 allows us to study the asymptotic properties of $\mathrm{GIC}^*(\alpha_\lambda)$ instead of

GIC(λ).

The following theorem describes the uniform stochastic rate of the difference between $\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})$ and the corresponding Kullback-Leibler distance between model α_λ and α_0 over all possible model α_λ, the number of which increases to infinity combinatorially fast with sample size. All expectations are taken under the true model.

Theorem 5.4.2. *Under Conditions (A) to (I), uniformly for all models,*

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| = O_p[n^{1/2}\{\log(d_n)\}^{1/2}].$$

Based on Theorem 5.4.2, for any underfitted model $\alpha_\lambda \not\supset \alpha_0$ we have that,

$$\begin{aligned}
&\inf_{\alpha_\lambda \not\supset \alpha_0} \{\mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)\} \\
&= \inf_{\alpha_\lambda \not\supset \alpha_0} \frac{1}{n} \left\{ \ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) + a_n(|\alpha_\lambda| - |\alpha_0|) \right\} \\
&= \inf_{\alpha_\lambda \not\supset \alpha_0} \frac{1}{n} \left[\ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_0}) - \ell_n(\beta^0_{\alpha_\lambda})\} + \mathrm{E}\{\ell_n(\beta^0_{\alpha_0}) - \ell_n(\beta^0_{\alpha_\lambda})\} \right. \\
&\quad \left. + a_n(|\alpha_\lambda| - |\alpha_0|) \right] \\
&\geq -\frac{1}{n} \sup_{\alpha_\lambda \not\supset \alpha_0} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| - \frac{1}{n} \left| \ell_n(\hat{\beta}_{\alpha_0}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_0})\} \right| + \inf_{\alpha_\lambda \not\supset \alpha_0} D_{KL}(\beta^0_{\alpha_\lambda}) \\
&\quad + \inf_{\alpha_\lambda \not\supset \alpha_0} \frac{1}{n} a_n(|\alpha_\lambda| - |\alpha_0|) \\
&\geq -\frac{1}{n}|\alpha_0| O_p[n^{1/2}\{\log(d_n)\}^{1/2}] - \frac{1}{n}|\alpha_0| O_p[n^{1/2}\{\log(d_n)\}^{1/2}] + \delta_n - \frac{1}{n} a_n k_n \\
&\geq -\frac{2}{n} k_n O_p[n^{1/2}\{\log(d_n)\}^{1/2}] + \delta_n - \frac{1}{n} a_n k_n \\
&= \delta_n - \frac{1}{n} k_n \left(O_p[n^{1/2}\{\log(d_n)\}^{1/2}] + a_n \right) \qquad (5.4)
\end{aligned}$$

where $\delta_n := \inf_{\alpha_\lambda \not\supset \alpha_0} D_{KL}(\beta^0_{\alpha_\lambda})$ defines the smallest Kullback-Leibler distance to the true model among all underfitted models. It can be deemed as the signal strength of the true model. Since δ_n is always positive, when $\delta_n k_n^{-1} n^{1/2}\{\log(d_n)\}^{-1/2} \to \infty$ and $a_n = o(\delta_n n k_n^{-1})$, (5.4) is positive with probability tending to one. By Lemma 5.4.1, we then have that

$\mathrm{pr}\left[\inf_{\lambda\in\Omega_-}\{\mathrm{GIC}(\lambda)-\mathrm{GIC}(\lambda_0)\}>0\right]\to 1$ as $n\to\infty$. This result suggests that as long as the signal strength of the true model does not decay to 0 too fast and the sequence a_n does not go to infinity too fast, then the GIC of any underfitted model is larger than that of the true model with probability tending to one as sample size goes to infinity. Note that $a_n = O_p[n^{1/2}\{\log(d_n)\}^{1/2}]$ always works as long as δ_n satisfies its requirement.

For overfitted models, the Kullback-Leibler distance based method used in Theorem 5.4.2 does not apply anymore. This is because for any overfitted model $\alpha_\lambda \supsetneq \alpha_0$, its Kullback-Leibler distance to the true model is always 0. We instead study the asymptotic property of $\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\hat{\beta}_{\alpha_0})$ directly. If the dimension of the model is finite, then it is well established that 2 times the log-partial likelihood ratio converges to a χ^2 distribution with $|\alpha_\lambda|-|\alpha_0|$ degree of freedom. However, when the model size goes to infinity, we have to consider higher order terms in the linearization of the log-partial likelihood ratio statistic. Moreover, obtaining a uniform stochastic rate of $\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\hat{\beta}_{\alpha_0})$ over all overfitted models is also challenging since the number of overfitted models increases to infinity at an extremely fast rate.

Theorem 5.4.3. *Under Conditions (A) to (I), uniformly for all $\alpha_\lambda \supsetneq \alpha_0$,*

$$\sup_{\alpha_\lambda\supsetneq\alpha_0}\frac{1}{|\alpha_\lambda|-|\alpha_0|}\left\{\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\hat{\beta}_{\alpha_0})\right\}=O_p\{\log(d_n)\}.$$

As a consequence of Theorem 5.4.3, uniformly for all overfitted model we have that

$$\begin{aligned}
&\inf_{\alpha_\lambda\supsetneq\alpha_0}\frac{\mathrm{GIC}^*(\alpha_\lambda)-\mathrm{GIC}^*(\alpha_0)}{|\alpha_\lambda|-|\alpha_0|}\\
&=\inf_{\alpha_\lambda\supsetneq\alpha_0}\frac{1}{n(|\alpha_\lambda|-|\alpha_0|)}\left\{\ell_n(\hat{\beta}_{\alpha_0})-\ell_n(\hat{\beta}_{\alpha_\lambda})+a_n(|\alpha_\lambda|-|\alpha_0|)\right\}\\
&=-\sup_{\alpha_\lambda\supsetneq\alpha_0}\frac{1}{n(|\alpha_\lambda|-|\alpha_0|)}\left\{\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\hat{\beta}_{\alpha_0})\right\}+\frac{a_n}{n}\\
&=-O_p\{n^{-1}\log(d_n)\}+\frac{a_n}{n}. \qquad (5.5)
\end{aligned}$$

Therefore, when $a_n/\log(d_n) \to \infty$, (5.5) is positive with probability tending to one. Since $|\alpha_\lambda| - |\alpha_0|$ is positive for all overfitted model, it follows that $\inf_{\alpha_\lambda \not\supseteq \alpha_0} \mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)$ is positive with probability tending to one. By Lemma 5.4.1 we then have that $\mathrm{pr}\left[\inf_{\lambda \in \Omega_+}\{\mathrm{GIC}(\lambda) - \mathrm{GIC}(\lambda_0)\} > 0\right] \to 1$ as $n \to \infty$.

With Theorem 5.4.2 and 5.4.3, we finally arrive at the following theorem.

Theorem 5.4.4. *Under Conditions (A) to (I), if* $\delta_n k_n^{-1} n^{1/2}\{\log(d_n)\}^{-1/2} \to \infty$, $a_n = o(\delta_n n k_n^{-1})$, *and* $a_n/\log(d_n) \to \infty$, *then as* $n \to \infty$,

$$\mathrm{pr}\left\{\inf_{\lambda \in \Omega_- \cup \Omega_+} \mathrm{GIC}(\lambda) > \mathrm{GIC}(\lambda_0)\right\} \to 1.$$

Theorem 5.4.4 is a direct consequence of Theorem 5.4.2 and 5.4.3. It entails that, if the signal strength of the true model does not decrease to 0 too fast and a_n diverges with sample size within a proper range of rate, then by minimizing GIC we can identify the tuning parameter that leads to the true model with probability tending to one as sample size goes to infinity.

5.5 Numerical Study and Application

5.5.1 Simulation Study

Independent failure times are generated from the exponential hazard model. We set $\lambda_0(t) = 2$ and the dimension of β to be $d_n = [10 n_c^{1/5 - 1/500}]$, where n_c is the number of cases and $[x]$ rounds x to the nearest integer. We relate the model dimension to the number of cases rather than sample size as the former better represents the amount of information carried in the dataset. The first component of β is the smallest nonzero parameter in terms of the absolute value, which is related to δ_n, the signal strength of the true model. As it is not possible to verify the requirement on the convergence rate of δ_n under finite sample size, we consider two different values of the smallest nonzero

parameter in terms of the absolute value: 0.34 (large effect scenario with corresponding hazard ratio of 1.4) and 0.18 (small effect scenario with corresponding hazard ratio of 1.2). There is one nonzero parameter for every two zero parameters, with the other nonzero parameters recycling from values 0.6 and -0.8. For example, when $d_n = 15$, $\beta_{\min} = 0.34$, then $\beta = (0.34, 0, 0, 0.6, 0, 0, -0.8, 0, 0, 0.6, 0, 0, -0.8, 0, 0)$. We generate the design matrix Z as a mixture of correlated binary and continuous variables. First, d_n-dimensional multivariate standard normal variable Z^* are generated with the correlation coefficient between Z_i^* and Z_j^* being $0.5^{|i-j|}$. Then the first three components of Z^* are kept as continuous, and the next three components are dichotomized at 0, and this pattern is repeated for the rest of Z^*. Thus half of the covariates become binary with parameter 0.5. Censoring times C_i are generated from a uniform distribution $\mathrm{U}(0, c)$ where c is adjusted to achieve desired censoring percentage.

Two sample sizes and two censoring rates are considered for each $\beta_{\min}$ value (0.34 or 0.18). Performance of the SCAD-penalized variable selection procedures with the GIC tuning parameter selection criterion is assessed for four different choices of a_n: 2, $\log(n)$, $\log\{\log(n)\}\log(d_n)$, and $\log\{\log(d_n)\}\log(d_n)$. The first two choices correspond to the AIC and BIC statistic, respectively. Obviously $a_n = 2$ does not satisfy the required divergence rate as described in Theorem 5.4.4, whereas the other three choices all meet the requirement on a_n. We will empirically evaluate their performance. As a benchmark, we include the hard threshold variable selection procedure, where the component of the unpenalized maximum partial likelihood estimator from the full model is selected if its p-value from the Wald test is less than 0.05. We also include the result from the oracle procedure where the correct subset of covariates is used to fit the model. For each setting 500 replications are conducted.

We define model error of a variable selection procedure as $\mathrm{ME}(\hat{\mu}) = \mathrm{E}\{\mathrm{E}(T|z) - \hat{\mu}(z)\}^2$, and the relative model error as the ratio of its model error to that of the unpenalized

pseudo-partial likelihood estimates from the full model. We use the median and the median absolute deviation of the relative model error to compare the performance of different variable selection procedures. We also calculate the average number of parameters correctly estimated as 0, the average number of parameters erroneously estimated as 0, and the overall rate of identifying the true model. Point estimates, empirical and model-based standard errors, and the empirical 95% confidence interval coverage are also calculated for $\hat{\beta}_{\min}$ using replications with nonzero $\hat{\beta}_{\min}$.

Table 5.1 summarizes the variable selection performance for different GIC statistics. Overall, GIC 4 with $a_n = \log\{\log(d_n)\}\log(d_n)$ gives the best performance in terms of rate of identifying the true model and the median relative model error. This observation is consistent across different $\beta_{\min}$ sizes, censoring rates, and sample sizes. The only scenarios where the performance of GIC 2 and 3 are similar to or slightly better than that of GIC 4 are when all these GICs have very high rate of identifying the true model (over 90%). Based on the average number of correctly identified zero parameters (column C) and incorrectly identified zero parameters (column I), GIC 1 tends to select more parameters into the final model than does GIC 4, whereas GIC 2 and 3 tend to select less parameters than does GIC 4. This is consistent with the fact that the divergence rate of a_n in GIC 4 lies between that in GIC 1 ($a_n = 2$) and GIC 2 ($a_n = \log(n)$) and 3 ($a_n = \log\{\log(n)\}\log(d_n)$). As a result, the penalty from GIC 4 on the model size lies between that from GIC 1 and GIC 2 and 3. As expected, the variable selection performance of all procedures increases with larger effect size, lower censoring rate, and larger sample size.

Table 5.2 summarizes the parameter estimation of $\beta_{\min}$ for different GIC statistics under the same settings as in Table 5.1. Under large $\beta_{\min}$ (0.34) scenario, given that it is correctly identified as nonzero, GIC 4 produces approximately unbiased point and standard error estimates and the 95% confidence interval coverage is close to the nominal level. Under small $\beta_{\min}$ (0.18) scenario, given that it is correctly identified as nonzero, GIC 4 tends to

overestimates the parameter as the other three GICs. However, the bias decreases as their variable selection performance increases. The overestimation is due to the fact that very small $\hat{\beta}_{\min}$ are set to 0 by the variable selection algorithm and therefore are not accounted for in the computation of average of the point estimates. Eventually, when the rate of identifying the true model is over 90%, all GICs give unbiased point and standard error estimates and correct 95% confidence interval coverage. This observation is consistent with the simulation result of Chapter 3.

5.5.2 Analysis of Framingham Heart Study

We apply the proposed tuning parameter selection method to the Framingham Heart Study (Dawber 1980). This study was initiated in 1948, with 2,336 men and 2,873 women aged between 30 and 62 years at their baseline examination. Participants were followed up to the year 1980, and times to multiple cardiovascular events were observed from each individual. For the analysis in this section, we only include participants who had an examination at age 44 or 45 and were event-free at that time. We use that examination time as the time origin for the survival analysis. We analyze the time to obtain the first evidence of coronary heart disease (CHD). The dataset consists of 1,571 participants, 250 of which developed evidence of CHD, corresponding to a censoring rate of 84.1%. We consider the following risk factors of interest: body mass index (BMI), cholesterol level, systolic blood pressure (SBP), smoking status (1=smoker and 0=otherwise), gender (1=female, 0=male). The risk factors were measured at the time origin of each participant. Since some individuals were in the study for several years prior to their time origin for this analysis, the waiting time from entering the study to the time origin is used as another covariate to account for the potential cohort effect. All continuous covariates are standardized for the analysis. To explore possible quadratic and interaction effects of the risk factors, we include quadratic terms of all continuous covariates and all pairwise interactions in addition to the

Table 5.1: Model selection performance of different choice of a_n in the GIC statistic.

	80% Censored				90% Censored			
	RME	Zero Parm.		RITM	RME	Zero Parm.		RITM
Method	median (MAD)	C	I	(%)	median (MAD)	C	I	(%)
	$n = 1500$, $\beta_{\min} = 0.34$, $d_n = 31$ for 80% censored, $d_n = 27$ for 90% censored							
HT	0.72 (0.17)	18.75	0.05	29.6	0.81 (0.25)	16.87	0.63	16.6
GIC 1	0.46 (0.18)	19.25	0.02	48	0.71 (0.21)	15.31	0.21	8
GIC 2	0.56 (0.39)	19.98	0.73	52.8	3.64 (2.48)	17.99	3.25	2
GIC 3	0.46 (0.29)	19.97	0.53	62.4	2.65 (1.91)	17.98	2.8	6
GIC 4	0.36 (0.18)	19.88	0.12	80.6	0.86 (0.57)	17.77	1.24	24.4
Oracle	0.33 (0.14)	20	0	100	0.29 (0.14)	18	0	100
	$n = 2500$, $\beta_{\min} = 0.34$, $d_n = 34$ for 80% censored, $d_n = 30$ for 90% censored							
HT	0.71 (0.15)	20.74	0	31.2	0.7 (0.19)	18.87	0.1	31.2
GIC 1	0.47 (0.18)	21.45	0	60	0.63 (0.18)	17.55	0.02	9.8
GIC 2	0.36 (0.16)	22	0.03	96.8	1.61 (1.25)	19.98	1.59	20.6
GIC 3	0.36 (0.16)	22	0.03	97.4	1.22 (0.91)	19.97	1.19	28
GIC 4	0.37 (0.16)	21.94	0	93.8	0.44 (0.25)	19.85	0.29	67.2
Oracle	0.36 (0.15)	22	0	100	0.31 (0.13)	20	0	100
	$n = 2500$, $\beta_{\min} = 0.18$, $d_n = 34$ for 80% censored, $d_n = 30$ for 90% censored							
HT	0.71 (0.15)	20.74	0.07	26	0.69 (0.19)	18.87	0.38	21.8
GIC 1	0.49 (0.17)	21.45	0.05	54.6	0.66 (0.18)	17.59	0.16	10.6
GIC 2	0.45 (0.21)	22	0.56	47.6	2.27 (1.74)	19.99	2.48	3
GIC 3	0.44 (0.2)	22	0.5	52.8	1.51 (1.15)	19.98	2.03	6.6
GIC 4	0.4 (0.17)	21.92	0.17	77.2	0.5 (0.28)	19.85	0.79	35.8
Oracle	0.36 (0.15)	22	0	100	0.32 (0.14)	20	0	100
	$n = 5000$, $\beta_{\min} = 0.18$, $d_n = 39$ for 80% censored, $d_n = 34$ for 90% censored							
HT	0.71 (0.18)	24.56	0	24.6	0.67 (0.16)	20.7	0.06	27.4
GIC 1	0.44 (0.16)	25.47	0	59.2	0.66 (0.18)	19.59	0.01	9
GIC 2	0.37 (0.15)	25.99	0.08	91.6	0.47 (0.2)	21.99	0.62	43.8
GIC 3	0.36 (0.15)	25.99	0.05	94	0.44 (0.19)	21.99	0.49	53.4
GIC 4	0.37 (0.15)	25.94	0.01	93.2	0.4 (0.17)	21.93	0.17	79.2
Oracle	0.35 (0.14)	26	0	100	0.37 (0.16)	22	0	100

RME: relative model error; MAD: median absolute deviation; C: average number of 0 parameters correctly identified as 0; I: average number of nonzero parameters incorrectly identified as 0; RITM: rate of identifying true model; HT: hard threshold; GIC 1: $a_n = 2$; GIC 2: $a_n = \log(n)$; GIC 3: $a_n = \log\{\log(n)\}\log(d_n)$; GIC 4: $a_n = \log\{\log(d_n)\}\log(d_n)$.

Table 5.2: Parameter estimation for $\beta_{\min}$ for different choice of a_n in the GIC statistic.

	80% Censored				90% Censored			
Method	$\hat{\beta}_{\min}$	se_e	se_m	95% CI_e	$\hat{\beta}_{\min}$	se_e	se_m	95% CI_e
$n = 1500$, $\beta_{\min} = 0.34$, $d_n = 31$ for 80% censored, $d_n = 27$ for 90% censored								
HT	0.35	0.07	0.07	92.8	0.37	0.09	0.1	96.1
GIC 1	0.35	0.06	0.06	92	0.35	0.09	0.09	93.4
GIC 2	0.35	0.05	0.06	96.7	0.36	0.07	0.08	48.8
GIC 3	0.35	0.06	0.06	96.3	0.38	0.06	0.08	96
GIC 4	0.35	0.06	0.06	94.4	0.36	0.07	0.08	96.8
Oracle	0.35	0.06	0.06	94.8	0.34	0.09	0.08	94.2
$n = 2500$, $\beta_{\min} = 0.34$, $d_n = 34$ for 80% censored, $d_n = 30$ for 90% censored								
HT	0.34	0.05	0.05	94.4	0.35	0.08	0.08	94.9
GIC 1	0.34	0.05	0.05	94.2	0.35	0.07	0.07	93.2
GIC 2	0.34	0.05	0.05	95.2	0.34	0.06	0.06	86.2
GIC 3	0.34	0.05	0.05	95	0.35	0.06	0.06	98.3
GIC 4	0.34	0.05	0.05	94.8	0.34	0.06	0.06	95.5
Oracle	0.34	0.05	0.05	95	0.34	0.06	0.06	94.6
$n = 2500$, $\beta_{\min} = 0.18$, $d_n = 34$ for 80% censored, $d_n = 30$ for 90% censored								
HT	0.19	0.05	0.05	97	0.22	0.05	0.08	95.7
GIC 1	0.19	0.04	0.05	96.4	0.2	0.05	0.07	96.5
GIC 2	0.23	0.02	0.05	92.3	0.24	0.03	0.06	19.7
GIC 3	0.22	0.03	0.05	95.1	0.27	0.03	0.06	91.5
GIC 4	0.2	0.04	0.05	96.9	0.23	0.04	0.06	95.4
Oracle	0.18	0.05	0.05	94.2	0.18	0.06	0.06	95.6
$n = 5000$, $\beta_{\min} = 0.18$, $d_n = 39$ for 80% censored, $d_n = 34$ for 90% censored								
HT	0.18	0.04	0.04	95	0.19	0.05	0.05	96
GIC 1	0.18	0.03	0.03	93.6	0.19	0.05	0.05	95.1
GIC 2	0.19	0.02	0.03	96.9	0.2	0.03	0.05	59.8
GIC 3	0.18	0.03	0.03	97.3	0.22	0.03	0.05	94.7
GIC 4	0.18	0.03	0.03	94.5	0.2	0.04	0.05	96.4
Oracle	0.18	0.03	0.03	93.8	0.18	0.05	0.05	94.2

se_e: empirical standard error; se_m: model-based standard error; 95% CI_e: empirical 95% confidence interval coverage; HT: hard threshold; GIC 1: $a_n = 2$; GIC 2: $a_n = \log(n)$; GIC 3: $a_n = \log\{\log(n)\}\log(d_n)$; GIC 4: $a_n = \log\{\log(d_n)\}\log(d_n)$. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_{\min}$.

main effects. Thus, the full Cox proportional hazards model contains 25 covariates in total.

We analyze the data with the SCAD-penalized variable selection procedure with the four tuning parameter selection criteria assessed in the simulation. The hard threshold method is also used for comparison. The selected tuning parameters are: $\lambda = 0.2560$ for $a_n = 2$, $\lambda = 0.3572$ for $a_n = \log(n)$, $\lambda = 0.3572$ for $a_n = \log\{\log(n)\}\log(d_n)$, and $\lambda = 0.3235$ for $a_n = \log\{\log(d_n)\}\log(d_n)$. The selected models are summarized in Table 5.3. Consistent with the observations in the simulation study, the GIC with $a_n = 2$ identifies a larger model than the other methods. The GIC with $a_n = \log(n)$ and $a_n = \log\{\log(n)\}\log(d_n)$ both identify the same model with only two covariates (gender and BMI*wait time). The hard threshold method also selects a model with only two covariates (SBP and smoking status). In comparison, the GIC with $a_n = \log\{\log(d_n)\}\log(d_n)$ identifies a model that contains the smaller models selected by the GICs with $a_n = \log(n)$ and $a_n = \log\{\log(n)\}\log(d_n)$ and the hard threshold method, yet not as many covariates as the one from the GIC with $a_n = 2$.

Based on the results from GIC 4 model in Table 5.3, with other covariates being equal, higher systolic blood pressure, being a smoker, or being a male is associated with higher risk of developing coronary heart disease. There is also a cohort effect represented by the interaction between BMI and wait time. BMI seems to exhibit a negative association with the risk of CHD in people with longer wait time.

5.6 Discussion

In this chapter of the dissertation, we propose a tuning parameter selection criterion for the SCAD-penalized variable selection procedure under regular Cox proportional hazards model with a random sample and a diverging number of parameters. We prove that the proposed generalized information criterion (GIC) can identify the true model with probability tending to one as sample size goes to infinity, and establish the conditions required on the true model signal strength and divergence rate of the penalty term in the

Table 5.3: Estimated coefficients and standard errors from Framingham Heart Study.

	HT	GIC 1	GIC 2	GIC 3	GIC 4
Variable	$\hat{\beta}$ (ŝe)	$\hat{\beta}$ (ŝe)	$\hat{\beta}$ (ŝe)	$\hat{\beta}$ (ŝe)	$\hat{\beta}$ (ŝe)
BMI	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Cholesterol	0 (–)	0.17 (0.06)	0 (–)	0 (–)	0 (–)
SBP	0.45 (0.18)	0.20 (0.06)	0 (–)	0 (–)	0.24 (0.06)
Smoke (Y vs. N)	0.49 (0.24)	0.27 (0.14)	0 (–)	0 (–)	0.30 (0.14)
Gender (F vs. M)	0 (–)	-0.61 (0.14)	-0.82 (0.13)	-0.82 (0.13)	-0.69 (0.13)
Wait time (years)	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI^2	0 (–)	-0.07 (0.05)	0 (–)	0 (–)	0 (–)
Cholesterol^2	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
SBP^2	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Wait time^2	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI*Cholesterol	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI*SBP	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI*Smoke	0 (–)	0.22 (0.10)	0 (–)	0 (–)	0 (–)
BMI*Gender	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI*Wait time	0 (–)	-0.11 (0.08)	-0.14 (0.07)	-0.14 (0.07)	-0.13 (0.06)
Chol*SBP	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Chol*Smoke	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Chol*Gender	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Chol*Wait time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
SBP*Smoke	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
SBP*Gender	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
SBP*Wait time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Smoke*Gender	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Smoke*Wait time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Gender*Wait time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)

HT: hard threshold; GIC 1: $a_n = 2$; GIC 2: $a_n = \log(n)$; GIC 3: $a_n = \log\{\log(n)\}\log(d_n)$; GIC 4: $a_n = \log\{\log(d_n)\}\log(d_n)$.

GIC for the model selection consistency to hold.

The theorems developed in this chapter specify a range of divergence rates required for the sequence a_n. Any rate within that range leads to model selection consistency asymptotically. However, in real-life applications with finite sample size, different choice of a_n may yield different result. Therefore, we conduct simulation to compare four different choices of a_n, three of which satisfy the asymptotic requirement. The simulation results suggest that when the variable selection performance is close to perfect, there is not much difference among the three choices of a_n. When the setting is such that the variable selection performance is moderate, the choice of $a_n = \log\{\log(d_n)\}\log(d_n)$ works much better than the other choices. Based on this observation, we recommend using $a_n = \log\{\log(d_n)\}\log(d_n)$ in practice.

Some of the parameter estimation and inference results presented in Table 5.2 are less than satisfactory. This observation is related to the so-called post-selection inference problem (Buehler and Feddersen 1963, Leeb and Potscher 2005; 2006, Potscher and Leeb 2009), which exists for all inference procedures that involve model selection process. The conventional statistical inference does not take into account the fact that the selected model itself is stochastic, and thereby distort the true sampling distribution of the estimates. This topic is beyond the scope of this dissertation. Fortunately, when the variable selection performance is reasonably good, the parameter estimation and inference results are acceptable.

5.7 Proof of Theorems

Proof of Lemma 5.4.1. We first consider the penalized estimate under the true model, $\hat{\beta}_{\lambda_0}$, the support of which is $\{1, ..., k_n\}$. By definition, $\hat{\beta}_{\lambda_0}$ solves the equations

$$\frac{\partial \ell_n(\hat{\beta}_{\lambda_0})}{\partial \beta_j} - n b_{\lambda_0 j} = 0, \qquad j = 1, ..., k_n,$$

where $b_{\lambda_0 j} = P'_\lambda(|\hat{\beta}_{\lambda_0 j}|)\mathrm{sgn}(\hat{\beta}_{\lambda_0 j})$ and $\hat{\beta}_{\lambda_0 j}$ is the jth component of $\hat{\beta}_{\lambda_0}$. Under Conditions (A) to (H), $\hat{\beta}_{\lambda_0}$ possesses the oracle property. Therefore, $|\hat{\beta}_{\lambda_0 j}| \to |\beta_{0j}| \geq \min_{1\leq j\leq s_n} |\beta_{0j}|$, and $|\hat{\beta}_{\lambda_0 j}|/\lambda_0 \to \infty$. Consequently, $\mathrm{pr}(P'_\lambda(|\hat{\beta}_{\lambda_0 j}|) = 0) \to 1$ by the formula of the SCAD penalty, and therefore $\mathrm{pr}(b_{\lambda_0 j} = 0) \to 1$ for all $j = 1, ..., k_n$. As a result, with probability tending to one, $\hat{\beta}_{\lambda_0}$ solves the equations

$$\frac{\partial \ell_n(\hat{\beta}_{\lambda_0})}{\partial \beta_j} = 0, \qquad j = 1, ..., k_n,$$

which are the same equations that $\hat{\beta}_{\alpha_0}$ solves by definition. This implies that $\hat{\beta}_\lambda = \hat{\beta}_{\alpha_0}$ with probability tending to one. It follows that

$$\mathrm{pr}\{\mathrm{GIC}(\lambda_0) = \mathrm{GIC}^*(\alpha_0)\} \to 1. \tag{5.6}$$

On the other hand, for any $\lambda \in \Omega$ and any model α_λ, by the definition of $\hat{\beta}_{\alpha_\lambda}$ we have

$$\mathrm{GIC}(\lambda) \geq \mathrm{GIC}^*(\alpha_\lambda). \tag{5.7}$$

By (5.6) and (5.7), Lemma 5.4.1 is proved. □

The log-partial likelihood function under Cox proportional hazards model can be written in the summation format as

$$\ell_n(\beta) = \sum_{i=1}^{n} \left[\beta^T Z_i(t_i) - \log \textstyle\sum_{j=1}^{n} Y_j(t_i) \exp\{\beta^T Z_j(t_i)\}\right] \Delta_i.$$

Since the log-partial likelihood is a sum of dependent random variables, we introduce the following intermediate function to facilitate the theoretical derivation:

$$\bar{\ell}_n(\beta) = \sum_{i=1}^{n} \left[\beta^T Z_i(t_i) - \log\{n s_n^{(0)}(\beta, t_i)\}\right] \Delta_i,$$

where $s_n^{(0)}(\beta, t)$ is defined in Section 5.3. It is obvious that $\mathrm{E}\{\bar{\ell}_n(\beta)\} = \mathrm{E}\{\ell_n(\beta)\}$. Define $\mathrm{supp}(\beta)$ to be the support of β consisting of indices of nonzero components of β. Define the set $\mathscr{B}_{\alpha_\lambda}(N) := \{\beta \in \mathbb{R}^{d_n} : \|\beta - \beta^0_{\alpha_\lambda}\| \le N,\ \mathrm{supp}(\beta) = \alpha_\lambda\} \cup \{\beta^0_{\alpha_\lambda}\}$ for some $N > 0$. We then define

$$Z_{\alpha_\lambda, N}(\beta) := \frac{1}{n}\Big|\ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda}) - \big[\mathrm{E}\{\ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda})\}\big]\Big|,$$

for any $\beta \in \mathscr{B}_{\alpha_\lambda}(N)$. Fan and Tang (2013) studied the stochastic order of the supremum of $Z_{\alpha_\lambda, N}(\beta)$ over all $\beta \in \mathscr{B}_{\alpha_\lambda}(N)$ in a generalized linear model by using the Lipschitz property of the log likelihood. In Cox model, however, the log partial-likelihood does not possess Lipschitz property (Kong and Nan 2014). Therefore, we only consider pointwise stochastic order of $Z_{\alpha_\lambda, N}(\beta)$ for any given $\beta \in \mathscr{B}_{\alpha_\lambda}(N)$, which is adequate for our purpose because our focus is only on the penalized estimator.

Lemma 5.7.1. *Under Conditions (A) to (I), uniformly for all model α_λ,*

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} Z_{\alpha_\lambda, N}(\beta) = O_p\left[N\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right].$$

Proof. We first restate some of the theorems from Van de Geer (2008) that will be used in our proofs.

Theorem A.1 in Van de Geer (2008) (Bousquet concentration theorem):

Let $X_1, ..., X_n$ be independent random variables in space $\mathscr{X}$ and let Γ be a class of real-valued functions on $\mathscr{X}$ satisfying for some positive constants η_n and τ_n

$$\|\gamma\|_\infty \le \eta_n \quad \textit{and} \quad \frac{1}{n}\sum_{i=1}^{n} \mathrm{var}\{\gamma(X_i)\} \le \tau_n^2 \quad \forall \gamma \in \Gamma.$$

Define $Z := \sup_{\gamma\in\Gamma} \left| n^{-1} \sum_{i=1}^{n} \{\gamma(X_i) - \mathrm{E}\gamma(X_i)\} \right|$. *Then for any* $\varepsilon > 0$,

$$\mathrm{pr}\left[Z \geq \mathrm{E}Z + \varepsilon \left\{ 2(\tau_n^2 + 2\eta_n \mathrm{E}Z) \right\}^{1/2} + \frac{2\varepsilon^2 \eta_n}{3} \right] \leq \exp(-n\varepsilon^2).$$

Theorem A.2 in Van de Geer (2008) (Symmetrization theorem):

Let $X_1, ..., X_n$ *be independent random variables in space* $\mathscr{X}$ *and let* $\epsilon_1, ..., \epsilon_n$ *be a Rademacher sequence independent of* $X_1, ..., X_n$, *where* $\mathrm{pr}(\epsilon_i = 1) = \mathrm{pr}(\epsilon_i = -1) = 1/2$ *for all* i. *Let* Γ *be a class of real-valued functions on* $\mathscr{X}$. *Then*

$$\mathrm{E}\left[\sup_{\gamma\in\Gamma} \left| \sum_{i=1}^{n} \{\gamma(X_i) - \mathrm{E}\gamma(X_i)\} \right| \right] \leq 2\mathrm{E}\left\{ \sup_{\gamma\in\Gamma} \left| \sum_{i=1}^{n} \epsilon_i \gamma(X_i) \right| \right].$$

Lemma A.1 in Van de Geer (2008):

Let $X_1, ..., X_n$ *be independent random variables in space* $\mathscr{X}$ *and let* $\gamma_1, ..., \gamma_m$ *be real-valued functions on* $\mathscr{X}$ *satisfying for* $k = 1, ..., m$,

$$\mathrm{E}\gamma_k(X_i) = 0 \ \ \forall i \qquad \|\gamma_k\|_\infty \leq \eta_n \qquad \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\gamma_k^2(X_i) \leq \tau_n^2.$$

Then

$$\mathrm{E}\left\{ \max_{1\leq k\leq m} \left| \frac{1}{n} \sum_{i=1}^{2} \gamma_k(X_i) \right| \right\} \leq \left\{ \frac{2\tau_n^2 \log(2m)}{n} \right\} + \frac{\eta_n \log(2m)}{n}.$$

We then introduce the following two intermediate quantities:

$$Q_{\alpha_\lambda, N}(\beta) := \frac{1}{n} \left| \bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda}) - \left[\mathrm{E}\{\ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda})\} \right] \right|,$$
$$R_{\alpha_\lambda, N}(\beta) := \frac{1}{n} \left| \ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda}) - \{\bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})\} \right|.$$

We will study the tail probabilities of the above two quantities separately.

We would like to use Theorem A.1 in Van de Geer (2008) to establish a probability bound for $Q_{\alpha_\lambda,N}(\beta)$. However, the theorem involves the expectation of the quantity under study. Thus, we first derive a bound for $\mathrm{E}\{Q_{\alpha_\lambda,N}(\beta)\}$. Let $\epsilon_1,...,\epsilon_n$ be a Rademacher sequence, independent of $\bar{\ell}_1(\beta)-\bar{\ell}_1(\beta^0_{\alpha_\lambda}),...,\bar{\ell}_n(\beta)-\bar{\ell}_n(\beta^0_{\alpha_\lambda})$. By Theorem A.2 in Van de Geer (2008) with $X_i=\bar{\ell}_n(\beta)-\bar{\ell}_n(\beta^0_{\alpha_\lambda})$, γ being the identity function, and $\Gamma=\{\gamma\}$,

$$
\begin{aligned}
&\mathrm{E}\{Q_{\alpha_\lambda,N}(\beta)\}=\frac{1}{n}\mathrm{E}\left|\bar{\ell}_n(\beta)-\bar{\ell}_n(\beta^0_{\alpha_\lambda})-\left[\mathrm{E}\{\ell_n(\beta)-\ell_n(\beta^0_{\alpha_\lambda})\}\right]\right|\le\frac{2}{n}\mathrm{E}\left|\sum_{i=1}^n\epsilon_i\{\bar{\ell}_i(\beta)-\bar{\ell}_i(\beta^0_{\alpha_\lambda})\}\right|\\
&=\frac{2}{n}\mathrm{E}\left|\sum_{i=1}^n\epsilon_i\left(\left[\beta^TZ_i(t_i)-\log\left\{ns_n^{(0)}(\beta,t_i)\right\}\right]\Delta_i-\left[(\beta^0_{\alpha_\lambda})^TZ_i(t_i)-\log\left\{ns_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)\right\}\right]\Delta_i\right)\right|\\
&\le\frac{2}{n}\mathrm{E}\left|\sum_{i=1}^n\epsilon_i\left\{\beta^TZ_i(t_i)-(\beta^0_{\alpha_\lambda})^TZ_i(t_i)\right\}\Delta_i\right|+\frac{2}{n}\mathrm{E}\left|\sum_{i=1}^n\epsilon_i\left\{\log s_n^{(0)}(\beta,t_i)-\log s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)\right\}\Delta_i\right|\\
&=I_1+I_2.
\end{aligned}
$$

We first consider I_1. By Cauchy-Schwarz inequality,

$$
\begin{aligned}
I_1&=\frac{2}{n}\mathrm{E}\left|\sum_{i=1}^n\epsilon_i\left\{\sum_{j=1}^{|\alpha_\lambda|}(\beta_j-\beta^0_{\alpha_\lambda j})Z_{ij}(t_i)\right\}\Delta_i\right|=2\mathrm{E}\left|\sum_{j=1}^{|\alpha_\lambda|}\left\{(\beta_j-\beta^0_{\alpha_\lambda j})\frac{1}{n}\sum_{i=1}^n\epsilon_iZ_{ij}(t_i)\Delta_i\right\}\right|\\
&\le2\|\beta-\beta^0_{\alpha_\lambda}\|\mathrm{E}\left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\sum_{i=1}^n\frac{1}{n}\epsilon_iZ_{ij}(t_i)\Delta_i\right\}^2\right]^{1/2}\\
&\le2\|\beta-\beta^0_{\alpha_\lambda}\|\mathrm{E}\left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\max_{1\le j\le|\alpha_\lambda|}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_iZ_{ij}(t_i)\Delta_i\right|\right\}^2\right]^{1/2}\\
&=2\|\beta-\beta^0_{\alpha_\lambda}\||\alpha_\lambda|^{1/2}\mathrm{E}\left\{\max_{1\le j\le|\alpha_\lambda|}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_iZ_{ij}(t_i)\Delta_i\right|\right\}.
\end{aligned}
$$

Since $\mathrm{E}\{\epsilon_iZ_{ij}(t_i)\Delta_i\}=\mathrm{E}(\epsilon_i)\mathrm{E}\{Z_{ij}(t_i)\Delta_i\}=0$, $\|\epsilon_iZ_{ij}(t_i)\Delta_i\|_\infty\le\|Z_{ij}(t_i)\|_\infty\le K_n$, and $n^{-1}\sum_{i=1}^n\mathrm{E}\{\epsilon_iZ_{ij}(t_i)\Delta_i\}^2\le n^{-1}\sum_{i=1}^n\mathrm{E}\{Z_{ij}(t_i)^2\}\le n^{-1}\sum_{i=1}^n\mathrm{E}\|Z_{ij}(t_i)\|_\infty^2\le K_n^2$, by Lemma A.1 in Van de Geer (2008) with $X_i=\epsilon_i\Delta_iZ_i(t)$, $\gamma_k(\cdot)$ equal the k-th component of its argument,

$\eta_n = K_n$, and $\tau_n^2 = K_n^2$,

$$\mathrm{E}\left\{\max_{1\le j\le|\alpha_\lambda|}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i Z_{ij}(t_i)\Delta_i\right|\right\} \le \left\{\frac{2K_n^2\log(2|\alpha_\lambda|)}{n}\right\}^{1/2} + \frac{K_n\log(2|\alpha_\lambda|)}{n}.$$

It follows that,

$$I_1 \le 2|\alpha_\lambda|^{1/2}NK_n\left[\left\{\frac{2\log(2|\alpha_\lambda|)}{n}\right\}^{1/2} + \frac{\log(2|\alpha_\lambda|)}{n}\right].$$

Next we consider I_2. By mean value theorem, for some $\beta^*_{\alpha_\lambda}$ that lies between $\beta^0_{\alpha_\lambda}$ and β we have that

$$I_2 = \frac{2}{n}\mathrm{E}\left|\sum_{i=1}^{n}\epsilon_i\Delta_i\sum_{j=1}^{|\alpha_\lambda|}(\beta_j-\beta^0_{\alpha_\lambda j})\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t_i)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t_i)}\right| = 2\mathrm{E}\left|\sum_{j=1}^{|\alpha_\lambda|}(\beta_j-\beta^0_{\alpha_\lambda j})\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t_i)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t_i)}\right|,$$

where $s^{(1)}_{nj}(\beta,t)$ denotes the j-th component of $s^{(1)}_n(\beta,t)$, which is defined in Section 5.3. By Cauchy-Schwarz inequality we have that

$$\begin{aligned}
I_2 &\le 2\|\beta-\beta^0_{\alpha_\lambda}\|\mathrm{E}\left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t_i)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t_i)}\right\}^2\right]^{1/2}\\
&\le 2\|\beta-\beta^0_{\alpha_\lambda}\|\mathrm{E}\left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\max_{1\le j\le|\alpha_\lambda|}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t_i)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t_i)}\right|\right\}^2\right]^{1/2}\\
&= 2\|\beta-\beta^0_{\alpha_\lambda}\||\alpha_\lambda|^{1/2}\mathrm{E}\left\{\max_{1\le j\le|\alpha_\lambda|}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t_i)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t_i)}\right|\right\}.
\end{aligned}$$

By the definition of $s^{(1)}_{nj}(\beta,t)$ we have that

$$s^{(1)}_{nj}(\beta,t) = \mathrm{E}\left[Y(t)Z_j(t)\exp\{\beta^T Z(t)\}\right] \le K_n\mathrm{E}\left[Y(t)\exp\{\beta^T Z(t)\}\right] = K_n s^{(0)}_n(\beta,t).$$

Therefore, we have the following fact for all $i = 1, ..., n$ and $j = 1, ..., d_n$:

$$\mathrm{E}\left\{\epsilon_i \Delta_i \frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right\} = 0, \qquad \left\|\epsilon_i \Delta_i \frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right\|_\infty \leq \left\|\frac{K_n s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right\|_\infty = K_n,$$

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left\{\epsilon_i \Delta_i \frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right\}^2 \leq \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left\{\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right\}^2 \leq \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\frac{K_n s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right]^2 = K_n^2.$$

By Lemma A.1 in Van de Geer (2008) with $X_i = \epsilon_i \Delta_i s_n^{(1)}(\beta^*_{\alpha_\lambda}, t_i)\{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)\}^{-1}$, $\gamma_k(\cdot)$ equal the k-th component of its argument, $\eta_n = K_n$ and $\tau_n^2 = K_n^2$,

$$\mathrm{E}\left\{\max_{1\leq j\leq |\alpha_\lambda|} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i \Delta_i \frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right|\right\} \leq \left\{\frac{2K_n^2 \log(2|\alpha_\lambda|)}{n}\right\}^{1/2} + \frac{K_n \log(2|\alpha_\lambda|)}{n}.$$

It follows that,

$$I_2 \leq 2|\alpha_\lambda|^{1/2} N K_n \left[\left\{\frac{2\log(2|\alpha_\lambda|)}{n}\right\}^{1/2} + \frac{\log(2|\alpha_\lambda|)}{n}\right].$$

Therefore, for any $\beta \in \mathscr{B}_{\alpha_\lambda}(N)$,

$$\mathrm{E}\{Q_{\alpha_\lambda, N}(\beta)\} \leq I_1 + I_2 \leq 4|\alpha_\lambda|^{1/2} N K_n \left[\left\{\frac{2\log(2|\alpha_\lambda|)}{n}\right\}^{1/2} + \frac{\log(2|\alpha_\lambda|)}{n}\right].$$

Now we check the two conditions for Theorem A.1 in Van de Geer (2008). By Cauchy-Schwarz inequality and mean value theorem, for all i we have

$$\begin{aligned}
&\left|\bar{\ell}_i(\beta) - \bar{\ell}_i(\beta^0_{\alpha_\lambda})\right| \leq \left|\beta^T Z_i(t_i) - (\beta^0_{\alpha_\lambda})^T Z_i(t_i)\right|\Delta_i + \left|\log s_n^{(0)}(\beta, t_i) - \log s_n^{(0)}(\beta^0_{\alpha_\lambda}, t_i)\}\right|\Delta_i \\
&\leq \|\beta - \beta^0_{\alpha_\lambda}\| \left[\sum_{j=1}^{|\alpha_\lambda|} \{Z_{ij}(t_i)\}^2\right]^{1/2} + \left|\frac{\sum_{j=1}^{|\alpha_\lambda|}(\beta_j - \beta^0_{\alpha_\lambda j}) s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)}\right| \\
&\leq |\alpha_\lambda|^{1/2}\|\beta - \beta^0_{\alpha_\lambda}\| K_n + \|\beta - \beta^0_{\alpha_\lambda}\| \frac{\left\{\sum_{j=1}^{|\alpha_\lambda|} K_n^2 s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)^2\right\}^{1/2}}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t_i)} \\
&= |\alpha_\lambda|^{1/2}\|\beta - \beta^0_{\alpha_\lambda}\| K_n + |\alpha_\lambda|^{1/2}\|\beta - \beta^0_{\alpha_\lambda}\| K_n
\end{aligned}$$

$\leq 2|\alpha_\lambda|^{1/2} N K_n.$

Therefore, $\|\bar{\ell}_i(\beta) - \bar{\ell}_i(\beta^0_{\alpha_\lambda})\|_\infty \leq 2|\alpha_\lambda|^{1/2} N K_n$ and $\mathrm{var}\{\bar{\ell}_i(\beta) - \bar{\ell}_i(\beta^0_{\alpha_\lambda})\} \leq \mathrm{E}\{\bar{\ell}_i(\beta) - \bar{\ell}_i(\beta^0_{\alpha_\lambda})\}^2 \leq 4|\alpha_\lambda| N^2 K_n^2$. Let $\bar{a} := n^{-1/2}\{2\log(2|\alpha_\lambda|)\}^{1/2} + n^{-1}\log(2|\alpha_\lambda|)$, $\eta_n = 2|\alpha_\lambda|^{1/2} N K_n$, and $\tau_n^2 = 4|\alpha_\lambda| N^2 K_n^2$. Then by Theorem A.1 in Van de Geer (2008) with $X_i = \bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})$, γ being the identity function, and $\Gamma = \{\gamma\}$, for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{pr}\left[Q_{\alpha_\lambda, N}(\beta) \geq 4|\alpha_\lambda|^{1/2} N K_n \bar{a} + \varepsilon\{2(4|\alpha_\lambda| N^2 K_n^2 + 16|\alpha_\lambda| N^2 K_n^2 \bar{a})\}^{1/2} + \frac{4\varepsilon^2 |\alpha_\lambda|^{1/2} N K_n}{3} \right] \\
&= \mathrm{pr}\left[Q_{\alpha_\lambda, N}(\beta) \geq 2|\alpha_\lambda|^{1/2} N K_n \left\{ 2\bar{a} + \varepsilon(2 + 8\bar{a})^{1/2} + \frac{2\varepsilon^2}{3} \right\} \right] \leq \exp(-n\varepsilon^2). \qquad (5.8)
\end{aligned}$$

Next we consider $R_{\alpha_\lambda, N}(\beta)$. By mean value theorem, for some $\beta^*_{\alpha_\lambda}$ that lies between $\beta^0_{\alpha_\lambda}$ and β we have that

$$\begin{aligned}
R_{\alpha_\lambda, N}(\beta) &= \frac{1}{n}\sum_{i=1}^n \left| \left(\log\left[\frac{1}{n}\sum_{j=1}^n \frac{Y_j(t_i)\exp\{\beta^T Z_j(t_i)\}}{s_n^{(0)}(\beta, t_i)} \right] \right. \right. \\
&\quad \left. \left. - \log\left[\frac{1}{n}\sum_{j=1}^n \frac{Y_j(t_i)\exp\{(\beta^0_{\alpha_\lambda})^T Z_j(t_i)\}}{s_n^{(0)}(\beta^0_{\alpha_\lambda}, t_i)} \right] \right) \Delta_i \right| \\
&\leq \sup_{0\leq t\leq\tau} \left| \log\left[\sum_{j=1}^n \frac{Y_j(t)\exp\{\beta^T Z_j(t)\}}{s_n^{(0)}(\beta, t)} \right] - \log\left[\sum_{j=1}^n \frac{Y_j(t)\exp\{(\beta^0_{\alpha_\lambda})^T Z_j(t)\}}{s_n^{(0)}(\beta^0_{\alpha_\lambda}, t)} \right] \right| \\
&= \sup_{0\leq t\leq\tau} \left| \log\left\{ \frac{S_n^{(0)}(\beta, t)}{s_n^{(0)}(\beta, t)} \right\} - \log\left\{ \frac{S_n^{(0)}(\beta^0_{\alpha_\lambda}, t)}{s_n^{(0)}(\beta^0_{\alpha_\lambda}, t)} \right\} \right| \\
&= \sup_{0\leq t\leq\tau} \left| (\beta - \beta^0_{\alpha_\lambda})^T \left\{ \frac{S_n^{(1)}(\beta^*_{\alpha_\lambda}, t)}{S_n^{(0)}(\beta^*_{\alpha_\lambda}, t)} - \frac{s_n^{(1)}(\beta^*_{\alpha_\lambda}, t)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t)} \right\} \right| \\
&\leq \sup_{0\leq t\leq\tau} \|\beta - \beta^0_{\alpha_\lambda}\| \left[\sum_{j=1}^{|\alpha_\lambda|} \left\{ \frac{S_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t)}{S_n^{(0)}(\beta^*_{\alpha_\lambda}, t)} - \frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t)} \right\}^2 \right]^{1/2} \\
&= \sup_{0\leq t\leq\tau} \|\beta - \beta^0_{\alpha_\lambda}\| \left\{ \sum_{j=1}^{|\alpha_\lambda|} \left(\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda}, t)} \left[S_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t) - s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t) \right. \right. \right. \\
&\quad \left. \left. \left. + \frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda}, t)}{s_n^{(0)}(\beta^*_{\alpha_\lambda}, t)} \{ s_n^{(0)}(\beta^*_{\alpha_\lambda}, t) - S_n^{(0)}(\beta^*_{\alpha_\lambda}, t) \} \right] \right)^2 \right\}^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{0\leq t\leq\tau} \|\beta-\beta^0_{\alpha_\lambda}\| \left\{\sum_{j=1}^{|\alpha_\lambda|}\left(\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\left[\max_{1\leq j\leq|\alpha_\lambda|}\left|S_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)-s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)\right|\right.\right.\right.\\
&\quad \left.\left.\left.+K_n\left|S_n^{(0)}(\beta^*_{\alpha_\lambda},t)-s_n^{(0)}(\beta^*_{\alpha_\lambda},t)\right|\right]\right)^2\right\}^{1/2}\\
&= \sup_{0\leq t\leq\tau} \|\beta-\beta^0_{\alpha_\lambda}\||\alpha_\lambda|^{1/2}\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\left\{\max_{1\leq j\leq|\alpha_\lambda|}\left|S_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)-s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)\right|\right.\\
&\quad \left.+K_n\left|S_n^{(0)}(\beta^*_{\alpha_\lambda},t)-s_n^{(0)}(\beta^*_{\alpha_\lambda},t)\right|\right\}\\
&\leq \|\beta-\beta^0_{\alpha_\lambda}\||\alpha_\lambda|^{1/2}\sup_{0\leq t\leq\tau}\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\sup_{0\leq t\leq\tau}\left\{\max_{1\leq j\leq|\alpha_\lambda|}\left|S_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)-s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)\right|\right.\\
&\quad \left.+K_n\left|S_n^{(0)}(\beta^*_{\alpha_\lambda},t)-s_n^{(0)}(\beta^*_{\alpha_\lambda},t)\right|\right\}. \qquad (5.9)
\end{aligned}
$$

Under Condition (I) we have that

$$
S_n^{(0)}(\beta^*_{\alpha_\lambda},t)\geq\frac{1}{n}\sum_{i=1}^n Y_i(t)\inf_{\beta,Z_i}\exp\{\beta^TZ_i(t)\}=\frac{1}{n}\sum_{i=1}^n Y_i(t)\exp\{-\sup_{\beta,Z_i}\beta^TZ_i(t)\}=U_n^{-1}\frac{1}{n}\sum_{i=1}^n Y_i(t).
$$

Since $Y(t)$ is a non-increasing function of t, we have that

$$
\inf_{0\leq t\leq\tau} S_n^{(0)}(\beta^*_{\alpha_\lambda},t)\geq U_n^{-1}\frac{1}{n}\sum_{i=1}^n Y_i(\tau),
$$

and therefore

$$
\sup_{0\leq t\leq\tau}\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\leq U_n\left\{\frac{1}{n}\sum_{i=1}^n Y_i(\tau)\right\}^{-1}.
$$

Define $\mu\coloneqq \mathrm{E}\{Y(\tau)\}$. By Lemma 3.2 in Kong and Nan (2014),

$$
\mathrm{pr}\left\{\frac{1}{n}\sum_{i=1}^n Y_i(\tau)\leq\frac{\mu}{2}\right\}=\mathrm{pr}\left[\left\{\frac{1}{n}\sum_{i=1}^n Y_i(\tau)\right\}^{-1}\geq\frac{2}{\mu}\right]\leq 2\exp\left(-\frac{n\mu^2}{2}\right).
$$

Therefore,

$$\text{pr}\left\{\sup_{0\le t\le\tau}\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\ge\frac{2U_n}{\mu}\right\}\le 2\exp\left(-\frac{n\mu^2}{2}\right).$$

By a modification of Lemma 3.3 and 3.4 in Kong and Nan (2014) we have that for any positive constant ε,

$$\text{pr}\left\{\sup_{0\le t\le\tau}\left|S_n^{(0)}(\beta^*_{\alpha_\lambda},t)-s_n^{(0)}(\beta^*_{\alpha_\lambda},t)\right|\ge U_n\varepsilon\right\}\le\frac{1}{5}W^2\exp(-n\varepsilon^2),$$

$$\text{pr}\left\{\sup_{0\le t\le\tau}\max_{1\le j\le|\alpha_\lambda|}\left|S_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)-s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)\right|\ge U_nK_n\varepsilon\right\}\le\frac{1}{5}|\alpha_\lambda|W^2\exp(-n\varepsilon^2),$$

where W is a constant determined by the bracketing number of the class of functions indexed by t, $\mathscr{F}=\{Y(t)\exp\{\beta^TZ(t)\}U_n^{-1}:t\in[0,\tau],\exp\{\beta^TZ(t)\}\le U_n\}$. Applying these results to (5.9) we have

$$\text{pr}\left\{R_{\alpha_\lambda,N}(\beta)\ge\frac{2N|\alpha_\lambda|^{1/2}U_n^2K_n\varepsilon}{\mu}\right\}\le 2\exp\left(-\frac{n\mu^2}{2}\right)+\frac{1}{5}(|\alpha_\lambda|+1)W^2\exp(-n\varepsilon^2).\qquad(5.10)$$

Since $Z_{\alpha_\lambda,N}(\beta)\le Q_{\alpha_\lambda,N}(\beta)+R_{\alpha_\lambda,N}(\beta)$, by (5.8) and (5.10) we have that

$$\begin{aligned}&\text{pr}\left[Z_{\alpha_\lambda,N}(\beta)\ge 2NK_n|\alpha_\lambda|^{1/2}\left\{2\bar{a}+\varepsilon(2+8\bar{a})^{1/2}+\frac{2\varepsilon^2}{3}+\frac{U_n^2\varepsilon}{\mu}\right\}\right]\\&\le 2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(|\alpha_\lambda|+1)W^2+1\right\}\exp(-n\varepsilon^2).\end{aligned}$$

To establish the stochastic order of $Z_{\alpha_\lambda,N}(\beta)$, we use the following result: for any random sequence X_n, a_n, b_n and a diverging sequence γ_n, $\text{pr}(X_n\ge a_n+b_n\gamma_n)=o(1)$ implies that $X_n=O_p(a_n+b_n)$. Let $\varepsilon=\{|\alpha_\lambda|\log(d_n)\}^{1/2}n^{-1/2}\gamma_n$, where γ_n is any diverging sequence. Then,

$$\text{pr}\left(Z_{\alpha_\lambda,N}(\beta)\ge 2NK_n|\alpha_\lambda|^{1/2}\left[2\bar{a}+\gamma_n\left\{\frac{|\alpha_\lambda|\log(d_n)(2+8\bar{a})}{n}\right\}^{1/2}+\frac{2|\alpha_\lambda|\log(d_n)\gamma_n^2}{3n}\right.\right.$$

$$+\frac{U_n^2\{|\alpha_\lambda|\log(d_n)\}^{1/2}\gamma_n}{n^{1/2}\mu}\Bigg]\Bigg)$$
$$\leq 2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(|\alpha_\lambda|+1)W^2+1\right\}\exp\{-|\alpha_\lambda|\log(d_n)\gamma_n^2\}. \tag{5.11}$$

Since $\bar{a}=n^{-1/2}\{2\log(2|\alpha_\lambda|)\}^{1/2}+n^{-1}\log(2|\alpha_\lambda|)$ and $\log(2|\alpha_\lambda|)<|\alpha_\lambda|$ for all $|\alpha_\lambda|$, we have that $\bar{a}<(2|\alpha_\lambda|)^{1/2}n^{-1/2}+2|\alpha_\lambda|n^{-1}<2(2|\alpha_\lambda|)^{1/2}n^{-1/2}$. Hence,

$$\begin{aligned}
&\mathrm{pr}\Bigg(Z_{\alpha_\lambda,N}(\beta)\geq 2NK_n|\alpha_\lambda|^{1/2}\Bigg[2\bar{a}+\gamma_n\left\{\frac{|\alpha_\lambda|\log(d_n)(2+8\bar{a})}{n}\right\}^{1/2}+\frac{2|\alpha_\lambda|\log(d_n)\gamma_n^2}{3n}\\
&\quad+\frac{U_n^2\{|\alpha_\lambda|\log(d_n)\}^{1/2}\gamma_n}{n^{1/2}\mu}\Bigg]\Bigg)\\
&\geq \mathrm{pr}\Bigg[Z_{\alpha_\lambda,N}(\beta)\geq 2NK_n|\alpha_\lambda|^{1/2}\Bigg(4\left(\frac{2|\alpha_\lambda|}{n}\right)^{1/2}+\gamma_n\left[\frac{|\alpha_\lambda|\log(d_n)\{2+16(2|\alpha_\lambda|)^{1/2}n^{-1/2}\}}{n}\right]^{1/2}\\
&\quad+\frac{2|\alpha_\lambda|^{1/2}d_n^{1/2}\log(d_n)\gamma_n^2}{3n}+\frac{U_n^2\{|\alpha_\lambda|\log(d_n)\}^{1/2}\gamma_n}{n^{1/2}\mu}\Bigg)\Bigg]\\
&=\mathrm{pr}\Bigg[Z_{\alpha_\lambda,N}(\beta)\geq\frac{2NK_n|\alpha_\lambda|}{n^{1/2}}\Bigg(4*2^{1/2}+\gamma_n\left[\log(d_n)\{2+16(2|\alpha_\lambda|)^{1/2}n^{-1/2}\}\right]^{1/2}\\
&\quad+\frac{2d_n^{1/2}\log(d_n)\gamma_n^2}{3n^{1/2}}+\frac{U_n^2\{\log(d_n)\}^{1/2}\gamma_n}{\mu}\Bigg)\Bigg].
\end{aligned}\tag{5.12}$$

By (5.11) and (5.12) we have that

$$\begin{aligned}
&\mathrm{pr}\Bigg[Z_{\alpha_\lambda,N}(\beta)\geq\frac{2NK_n|\alpha_\lambda|}{n^{1/2}}\Bigg(4*2^{1/2}+\gamma_n\left[\log(d_n)\{2+16(2|\alpha_\lambda|)^{1/2}n^{-1/2}\}\right]^{1/2}\\
&\quad+\frac{2d_n^{1/2}\log(d_n)\gamma_n^2}{3n^{1/2}}+\frac{U_n^2\{\log(d_n)\}^{1/2}\gamma_n}{\mu}\Bigg)\Bigg]\\
&\leq 2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(|\alpha_\lambda|+1)W^2+1\right\}\exp\{-|\alpha_\lambda|\log(d_n)\gamma_n^2\}.
\end{aligned}$$

Now we derive the probability bound for the supremum of $Z_{\alpha_\lambda,N}(\beta)$ over all possible model $|\alpha_\lambda|$. We use the fact that $\binom{d_n}{k}\leq(d_ne/k)^k$ for any $0\leq k\leq d_n$, where e is the Euler's

number.

$$
\begin{aligned}
&\mathrm{pr}\left[\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|}Z_{\alpha_\lambda,N}(\beta)\geq\frac{2NK_n}{n^{1/2}}\left(4*2^{1/2}+\gamma_n\left[\log(d_n)\{2+16(2|\alpha_\lambda|)^{1/2}n^{-1/2}\}\right]^{1/2}\right.\right.\\
&\quad\left.\left.+\frac{2d_n^{1/2}\log(d_n)\gamma_n^2}{3n^{1/2}}+\frac{U_n^2\{\log(d_n)\}^{1/2}\gamma_n}{\mu}\right)\right]\\
&\leq\sum_{|\alpha_\lambda|=1}^{d_n}\mathrm{pr}\left[Z_{\alpha_\lambda,N}(\beta)\geq\frac{2NK_n|\alpha_\lambda|}{n^{1/2}}\left(4*2^{1/2}+\gamma_n\left[\log(d_n)\{2+16(2|\alpha_\lambda|)^{1/2}n^{-1/2}\}\right]^{1/2}\right.\right.\\
&\quad\left.\left.+\frac{2d_n^{1/2}\log(d_n)\gamma_n^2}{3n^{1/2}}+\frac{U_n^2\{\log(d_n)\}^{1/2}\gamma_n}{\mu}\right)\right]\\
&\leq\sum_{k=1}^{d_n}\binom{d_n}{k}\left[2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(k+1)W^2+1\right\}\exp\{-k\log(d_n)\gamma_n^2\}\right]\\
&\leq\sum_{k=1}^{d_n}\left(\frac{d_ne}{k}\right)^k\left[2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(k+1)W^2+1\right\}\exp\{-k\log(d_n)\gamma_n^2\}\right]\\
&=\sum_{k=1}^{d_n}\left(\frac{e}{k}\right)^k\left[2d_n^k\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(k+1)W^2+1\right\}d_n^{(1-\gamma_n^2)k}\right].
\end{aligned}
\tag{5.13}
$$

By Condition (H), $\{(d_n+1)\log(d_n)/n\}=o(1)$. Thus $d_n^{d_n+1}=o\{\exp(n)\}$ and the first term in the square brackets in (5.13) is $o(d_n^{-1})$. Since γ_n diverges to infinity, the second term in the square brackets in (5.13) is also $o(d_n^{-1})$. Moreover, $(e/k)^k<1$ for all $k\geq 3$. Therefore, it is easy to see that (5.13) goes to 0 as $n\to\infty$. Since γ_n diverges at an arbitrary rate, it follows that,

$$
\begin{aligned}
\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|}Z_{\alpha_\lambda,N}(\beta)&=O_p\left[\frac{2NK_n}{n^{1/2}}\left(4*2^{1/2}+\left[\log(d_n)\{2+16(2|\alpha_\lambda|)^{1/2}n^{-1/2}\}\right]^{1/2}\right.\right.\\
&\quad\left.\left.+\frac{2d_n^{1/2}\log(d_n)}{3n^{1/2}}+\frac{U_n^2\{\log(d_n)\}^{1/2}}{\mu}\right)\right]=O_p\left[N\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right]
\end{aligned}
$$

since $\{d_n\log(d_n)/n\}=o(1)$ under Condition (H).

□

Lemma 5.7.2. *Under Conditions (A) to (I), uniformly for all model α_λ,*

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| = O_p\left[\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right].$$

Proof. Denote $\|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| = N_{n,\alpha_\lambda}$. Since $\hat{\beta}_{\alpha_\lambda}$ maximizes $\ell_n(\beta_{\alpha_\lambda})$, we have that $\ell_n(\beta^0_{\alpha_\lambda}) \leq \ell_n(\hat{\beta}_{\alpha_\lambda})$. Since $\beta^0_{\alpha_\lambda}$ minimizes the Kullback-Leibler distance, we have that $\mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \geq \mathrm{E}\{\ell_n(\hat{\beta}_{\alpha_\lambda})\}$ and $\partial \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}/\partial\beta = 0$, where the expectation is taken under the true model. It then follows that,

$$\begin{aligned} 0 &\leq \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_\lambda})\} \leq \ell_n(\hat{\beta}_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \left[\ell_n(\beta^0_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}\right] \\ &\leq n Z_{\alpha_\lambda, N_{n,\alpha_\lambda}}(\hat{\beta}_{\alpha_\lambda}). \end{aligned} \tag{5.14}$$

By Taylor expansion, for some $\beta^*_{\alpha_\lambda}$ that lies between $\hat{\beta}_{\alpha_\lambda}$ and $\beta^0_{\alpha_\lambda}$ we have that

$$\begin{aligned} &\mathrm{E}\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda})\} \\ &= (\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T \frac{\partial \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}}{\partial\beta} + \frac{1}{2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T \frac{\partial^2 \mathrm{E}\{\ell_n(\beta^*_{\alpha_\lambda})\}}{\partial\beta^2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) \\ &= -\frac{n}{2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T I_n(\beta^*_{\alpha_\lambda})(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) \\ &\leq -\frac{n}{2}\|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\|^2 \mathrm{eigen}_{\min}\{I_n(\beta^*_{\alpha_\lambda})\} \\ &\leq -\frac{n}{2} N^2_{n,\alpha_\lambda} C_3. \end{aligned} \tag{5.15}$$

The last two inequalities in (5.15) hold by spectral decomposition on $I_n(\beta^*_{\alpha_\lambda})$ and Condition (E). By (5.14) and (5.15) it must hold for $\hat{\beta}_{\alpha_\lambda}$ that $N_{n,\alpha_\lambda} \leq \{2Z_{\alpha_\lambda, N_{n,\alpha_\lambda}}(\hat{\beta}_{\alpha_\lambda})\}^{1/2} C_3^{-1/2}$. Then by Lemma 5.7.1 it can be shown that $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1/2} \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| = O_p[\{\log(d_n)/n\}^{1/2}]$.

□

Lemma 5.7.3. *Under Conditions (A) to (I), uniformly for all model α_λ,*

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) \right| = O_p\left\{\log(d_n)\right\}.$$

Proof. Define the event $\mathscr{A}_n := \{\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1/2} \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| \leq \gamma_n \{\log(d_n)/n\}^{1/2}\}$, where γ_n is any diverging sequence. Denote $\mathscr{A}_n^c$ as the complement of $\mathscr{A}_n$. Then for any positive number ε,

$$\mathrm{pr}\left\{ \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) \right| \geq \varepsilon \right\} \leq \mathrm{pr}\left\{ \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) \right| \geq \varepsilon \mid \mathscr{A}_n \right\} + \mathrm{pr}(\mathscr{A}_n^c).$$

By the definition of $\hat{\beta}_{\alpha_\lambda}$ and $\beta^0_{\alpha_\lambda}$, we know that $\ell_n(\beta^0_{\alpha_\lambda}) \leq \ell_n(\hat{\beta}_{\alpha_\lambda})$ and $\mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \geq \mathrm{E}\{\ell_n(\hat{\beta}_{\alpha_\lambda})\}$ for any model α_λ. Thus,

$$\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) \leq \ell_n(\hat{\beta}_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \left[\ell_n(\beta^0_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}\right] \leq n Z_{\alpha_\lambda, N_{n,\alpha_\lambda}}(\hat{\beta}_{\alpha_\lambda}), \tag{5.16}$$

where $N_{n,\alpha_\lambda} = \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\|$. Define $N_n^* = \gamma_n \{\log(d_n)/n\}^{1/2}\}$. By Lemma 5.7.1, we have that $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1} Z_{\alpha_\lambda, N_n^*}(\hat{\beta}_{\alpha_\lambda}) = O_p\{\gamma_n \log(d_n)/n\}$. Therefore, we also have that $\mathrm{E}\left\{\sup_{\alpha_\lambda} n|\alpha_\lambda|^{-1} Z_{\alpha_\lambda, N_n^*}(\hat{\beta}_{\alpha_\lambda})\right\} = O\{\gamma_n \log(d_n)\}$. Since $\sup_{\alpha_\lambda} n|\alpha_\lambda|^{-1} Z_{\alpha_\lambda, N_n^*}(\hat{\beta}_{\alpha_\lambda})$ is a positive integrable random variable, by Markov inequality,

$$\mathrm{pr}\left\{ \sup_{\alpha_\lambda} \frac{n}{|\alpha_\lambda|} Z_{\alpha_\lambda, N_n^*}(\hat{\beta}_{\alpha_\lambda}) \geq \varepsilon \right\} \leq \mathrm{E}\left\{ \sup_{\alpha_\lambda} \frac{n}{|\alpha_\lambda|} Z_{\alpha_\lambda, N_n^*}(\hat{\beta}_{\alpha_\lambda}) \right\} \varepsilon^{-1} = \frac{O\{\gamma_n \log(d_n)\}}{\varepsilon}.$$

Let $\varepsilon = \gamma_n^2 \log(d_n)$, then

$$\mathrm{pr}\left\{ \sup_{\alpha_\lambda} \frac{n}{|\alpha_\lambda|} Z_{\alpha_\lambda, N_n^*}(\hat{\beta}_{\alpha_\lambda}) \geq \gamma_n^2 \log(d_n) \right\} \leq O(\gamma_n^{-1}) = o(1). \tag{5.17}$$

By (5.16) and (5.17), it can be shown that

$$\mathrm{pr}\left\{\sup_{\alpha_\lambda}|\alpha_\lambda|^{-1}\left|\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\beta^0_{\alpha_\lambda})\right|\geq\gamma_n^2\log(d_n)\mid\mathscr{A}_n\right\}=o(1).$$

By Lemma 5.7.2, $\mathrm{pr}(\mathscr{A}_n^c)=o(1)$. Therefore,

$$\mathrm{pr}\left\{\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|}\left|\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\beta^0_{\alpha_\lambda})\right|\geq\gamma_n^2\log(d_n)\right\}\leq o(1)+o(1)=o(1).$$

It follows that $\sup_{\alpha_\lambda}|\alpha_\lambda|^{-1}\left|\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\beta^0_{\alpha_\lambda})\right|=O_p\left\{\log(d_n)\right\}$.

□

Lemma 5.7.4. *Under Conditions (A) to (I), uniformly for all model α_λ,*

$$\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|^{1/2}}\left|\ell_n(\beta^0_{\alpha_\lambda})-\mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}\right|=O_p\left[\{n\log(d_n)\}^{1/2}\right].$$

Proof. Since $\ell_n(\beta^0_{\alpha_\lambda})$ is a sum of dependent random variables, we decompose the quantity in the statement of the lemma as follows,

$$\begin{aligned}
&\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|^{1/2}}\left|\ell_n(\beta^0_{\alpha_\lambda})-\mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}\right|\\
&\leq\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|^{1/2}}\left\{\left|\ell_n(\beta^0_{\alpha_\lambda})-\bar{\ell}_n(\beta^0_{\alpha_\lambda})\right|+\left|\bar{\ell}_n(\beta^0_{\alpha_\lambda})-\mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\}\right|\right\}\\
&=\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|^{1/2}}(I_1+I_2).
\end{aligned}$$

We first consider I_1.

$$\begin{aligned}
I_1&=\left|\sum_{i=1}^n\log\left\{\frac{s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)}{S_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)}\right\}\Delta_i\right|\leq n\left|\sup_{0\leq t\leq\tau}\log\left\{\frac{s_n^{(0)}(\beta^0_{\alpha_\lambda},t)}{S_n^{(0)}(\beta^0_{\alpha_\lambda},t)}\right\}\right|\\
&\leq n\sup_{0\leq t\leq\tau}\left|\log\{S_n^{(0)}(\beta^0_{\alpha_\lambda},t)\}-\log\{s_n^{(0)}(\beta^0_{\alpha_\lambda},t)\}\right|. \qquad (5.18)
\end{aligned}$$

By mean value theorem, $\log\{S_n^{(0)}(\beta^0_{\alpha_\lambda},t)\}-\log\{s_n^{(0)}(\beta^0_{\alpha_\lambda},t)\}=(S_n^\star)^{-1}\{S_n^{(0)}(\beta^0_{\alpha_\lambda},t)-s_n^{(0)}(\beta^0_{\alpha_\lambda},t)\}$,

where S_n^* lies between $S_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$ and $s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$. By Lemma 3.3 in Kong and Nan (2014) we have that for any positive number ε,

$$\mathrm{pr}\left\{\sup_{0\le t\le\tau}\left|S_n^{(0)}(\beta_{\alpha_\lambda}^0, t) - s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\right| \ge U_n\varepsilon\right\} \le \frac{1}{5}W^2\exp(-n\varepsilon^2), \tag{5.19}$$

where W is a constant determined by the bracketing number of the class of functions indexed by t, $\mathscr{F} = \{Y(t)\exp\{\beta^T Z(t)\}/U_n : t \in [0,\tau], \exp\{\beta^T Z(t)\} \le U_n\}$. It follows from (5.19) that $S_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$ converges to $s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$ in probability uniformly on $t \in [0,\tau]$, and therefore so does S_n^*. By Condition (D), $s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$ is uniformly bounded away from 0. Let C_5 be a constant satisfying $0 < C_5 < \inf_{0\le t\le\tau} s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$. Define the event $\mathscr{A}_n := \{S_n^* > C_5\}$. Denote $\mathscr{A}_n^c$ as the complement of $\mathscr{A}$. Consider

$$\begin{aligned}
&\mathrm{pr}\left[\sup_{0\le t\le\tau}\left|\log\{S_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\} - \log\{s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\}\right| \ge \frac{U_n\varepsilon}{C_5}\right] \\
&= \mathrm{pr}\left[\sup_{0\le t\le\tau}\left|\frac{1}{S_n^*}\{S_n^{(0)}(\beta_{\alpha_\lambda}^0, t) - s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\}\right| \ge \frac{U_n\varepsilon}{C_5}\right] \\
&\le \mathrm{pr}\left[\sup_{0\le t\le\tau}\left|\frac{1}{S_n^*}\{S_n^{(0)}(\beta_{\alpha_\lambda}^0, t) - s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\}\right| \ge \frac{U_n\varepsilon}{C_5} \mid \mathscr{A}_n\right] + \mathrm{pr}(\mathscr{A}_n^c) \\
&= I_{11} + I_{12}.
\end{aligned}$$

By (5.19) we have

$$\begin{aligned}
I_{11} &\le \mathrm{pr}\left[\sup_{0\le t\le\tau}\left|\frac{1}{C_5}\{S_n^{(0)}(\beta_{\alpha_\lambda}^0, t) - s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\}\right| \ge \frac{U_n\varepsilon}{C_5}\right] \\
&= \mathrm{pr}\left\{\sup_{0\le t\le\tau}\left|S_n^{(0)}(\beta_{\alpha_\lambda}^0, t) - s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)\right| \ge U_n\varepsilon\right\} \\
&\le \frac{1}{5}W^2\exp(-n\varepsilon^2).
\end{aligned}$$

Further, we have that $I_{12} = o(1)$ since S_n^* converges to $s_n^{(0)}(\beta_{\alpha_\lambda}^0, t)$ in probability uniformly

on $t \in [0, \tau]$. Therefore, by replacing ε with $n^{-1/2}\varepsilon$, from (5.18) we have that

$$\mathrm{pr}\left(I_1 \geq \frac{n^{1/2} U_n \varepsilon}{C_5}\right) \leq \frac{1}{5} W^2 \exp(-\varepsilon^2). \tag{5.20}$$

Next we consider I_2. For any i, $|\bar{\ell}_i(\beta^0_{\alpha_\lambda})| \leq |(\beta^0_{\alpha_\lambda})^T Z_i(t_i) - \log\{s^{(0)}_n(\beta^0_{\alpha_\lambda}, t_i)\}| \leq |(\beta^0_{\alpha_\lambda})^T Z_i(t_i)| + |\log\{s^{(0)}_n(\beta^0_{\alpha_\lambda}, t_i)\}| \leq |\log(U_n)| + |\log(\mathrm{E}[\exp\{(\beta^0_{\alpha_\lambda})^T Z_i(t_i)\}])| \leq 2|\log(U_n)|$. It implies that $-2\log(U_n) \leq \bar{\ell}_i(\beta^0_{\alpha_\lambda}) \leq 2\log(U_n)$ for all i. Therefore, by Hoeffding's inequality, for any positive number ε,

$$\mathrm{pr}(I_2 \geq \varepsilon) \leq 2\exp\left[-\frac{\varepsilon^2}{2\sum_{i=1}^n 4\{\log(U_n)\}^2}\right] = 2\exp\left[-\frac{\varepsilon^2}{8n\{\log(U_n)\}^2}\right].$$

By replacing ε with $n^{1/2}\varepsilon$ we have

$$\mathrm{pr}(I_2 \geq n^{1/2}\varepsilon) \leq 2\exp\left[-\frac{\varepsilon^2}{2\sum_{i=1}^n 4\{\log(U_n)\}^2}\right] = 2\exp\left[-\frac{\varepsilon^2}{8\{\log(U_n)\}^2}\right]. \tag{5.21}$$

From (5.20) and (5.21) we get

$$\mathrm{pr}\left(I_1 + I_2 \geq \frac{n^{1/2} U_n \varepsilon}{C_5} + n^{1/2}\varepsilon\right) \leq \frac{1}{5} W^2 \exp(-\varepsilon^2) + 2\exp\left[-\frac{\varepsilon^2}{8\{\log(U_n)\}^2}\right].$$

Let $\varepsilon = \{\gamma_n |\alpha_\lambda| \log(d_n)\}^{1/2}$, where γ_n is any diverging sequence. Then,

$$\begin{aligned}
&\mathrm{pr}\left[I_1 + I_2 \geq \{n\gamma_n |\alpha_\lambda| \log(d_n)\}^{1/2}\left(\frac{U_n}{C_5} + 1\right)\right] \\
&\leq \frac{1}{5} W^2 \exp\{-\gamma_n |\alpha_\lambda| \log(d_n)\} + 2\exp\left[-\frac{\gamma_n |\alpha_\lambda| \log(d_n)}{8\{\log(U_n)\}^2}\right].
\end{aligned}$$

By using the fact that $\binom{d_n}{k} \leq (d_n e/k)^k$ for any $0 \leq k \leq d_n$, we have that

$$\mathrm{pr}\left[\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}}(I_1 + I_2) \geq \{n\gamma_n \log(d_n)\}^{1/2}\left(\frac{U_n}{C_5} + 1\right)\right]$$

$$
\begin{aligned}
&\leq \sum_{|\alpha_\lambda|=1}^{d_n} \mathrm{pr}\left[I_1 + I_2 \geq \{n\gamma_n |\alpha_\lambda| \log(d_n)\}^{1/2} \left(\frac{U_n}{C_5} + 1 \right) \right] \\
&\leq \sum_{k=1}^{d_n} \left(\frac{d_n e}{k} \right)^k \left(\frac{1}{5} W^2 \exp\{-\gamma_n k \log(d_n)\} + 2 \exp\left[-\frac{\gamma_n k \log(d_n)}{8\{\log(U_n)\}^2} \right] \right) \\
&= \sum_{k=1}^{d_n} \left(\frac{e}{k} \right)^k \left[\frac{1}{5} W^2 d_n^{k - k\gamma_n} + 2 d_n^{k - \frac{k\gamma_n}{8\{\log(U_n)\}^2}} \right]. \qquad (5.22)
\end{aligned}
$$

Since γ_n diverges to infinity, the two terms in the square brackets are both $o(d_n^{-1})$. Moreover, $(e/k)^k < 1$ for all $k \geq 3$. Therefore, (5.22) goes to 0 as $n \to \infty$. Hence, $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1/2}(I_1 + I_2) = O_p[\{n \log(d_n)\}^{1/2}]$. It then follows that $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1/2} \left| \ell_n(\beta^0_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| = O_p\left[\{n \log(d_n)\}^{1/2}\right]$.

□

Proof of Theorem 5.4.2. For all model α_λ we have that

$$
\begin{aligned}
&\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| \\
&= \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) + \ell_n(\beta^0_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| \\
&\leq \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) \right| + \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \left| \ell_n(\beta^0_{\alpha_\lambda}) - \mathrm{E}\{\ell_n(\beta^0_{\alpha_\lambda})\} \right|. \qquad (5.23)
\end{aligned}
$$

By Lemma 5.7.3 and 5.7.4, (5.23) $= O_p\{\log(d_n)\} + O_p[\{n \log(d_n)\}^{1/2}] = O_p[\{n \log(d_n)\}^{1/2}]$ under Condition (H). □

Proof of Theorem 5.4.3. We first restate the corollary of Lemma 1 in Laurent and Massart (2000) that will be used in our proof.

Corollary of Lemma 1 in Laurent and Massart (2000):

Let U be a χ^2 statistic with D degrees of freedom. For any positive ε,

$$
\mathrm{pr}\{U - D \geq 2(D\varepsilon)^{1/2} + 2\varepsilon\} \leq \exp(-\varepsilon).
$$

By Taylor expansion, for some $\beta^*_{\alpha_\lambda}$ that lies between $\hat{\beta}_{\alpha_\lambda}$ and $\hat{\beta}_{\alpha_0}$,

$$\begin{aligned}\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) &= (\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T \ell_n'(\beta^0_{\alpha_\lambda}) + \frac{1}{2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T \ell_n''(\beta^0_{\alpha_\lambda})(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) \\ &\quad + \frac{1}{6}\sum_{i=1}^{n}\sum_{j,k,l=1}^{d_n} \tilde{\ell}_i'''(\beta^*_{\alpha_\lambda})_{jkl}(\hat{\beta}_{\alpha_\lambda j} - \beta^0_{\alpha_\lambda j})(\hat{\beta}_{\alpha_\lambda k} - \beta^0_{\alpha_\lambda k})(\hat{\beta}_{\alpha_\lambda l} - \beta^0_{\alpha_\lambda l}) \\ &= I_1 + I_2 + I_3.\end{aligned}$$

Since $\alpha_\lambda \not\supseteq \alpha_0$, $\beta^0_{\alpha_\lambda} = \beta_0$, the true parameter. As the regular Cox proportional hazards model is a special case of that with a case-cohort design with the subcohort sampling probability being one, from Theorem 3.3.1 in Chapter 3 we have that $\|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| = O_p(|\alpha_\lambda|^{1/2} n^{-1/2})$ for any $\alpha_\lambda \not\supseteq \alpha_0$. By using Lemma 3.7.5 and 3.7.6 in Chapter 3 we can derive the stochastic orders of I_1, I_2, and I_3 for any $\alpha_\lambda \not\supseteq \alpha_0$ as follows. $I_1 \le \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| \|\ell_n'(\beta^0_{\alpha_\lambda})\| = O_p(|\alpha_\lambda|)$. We decompose I_2 as

$$\begin{aligned}I_2 &= \frac{1}{2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T \{\ell_n''(\beta^0_{\alpha_\lambda}) + nI_n(\beta^0_{\alpha_\lambda})\}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) - \frac{1}{2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T nI_n(\beta^0_{\alpha_\lambda})(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) \\ &= I_{21} - I_{22}.\end{aligned}$$

Since $I_{21} \le \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\|^2 O_p(n^{1/2}|\alpha_\lambda|) = O_p(|\alpha_\lambda|^2 n^{-1/2}) = o_p(|\alpha_\lambda|)$ under Condition (H) and $I_{22} \ge n\|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\|^2 \text{eigen}_{\min}\{I_n(\beta^0_{\alpha_\lambda})\}/2 \ge n\|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\|^2 C_3/2 = O_p(|\alpha_\lambda|)$, it follows that $I_{21} = o_p(I_{22})$. Also, in the proof of Theorem 3.3.1 we established that $\tilde{\ell}_i'''(\beta^*_{\alpha_\lambda})_{jkl}$ is $O_p(1)$. Thus, $I_3 \le O_p\{(|\alpha_\lambda|/n)^{3/2}|\alpha_\lambda|^{3/2} n\} = O_p(|\alpha_\lambda|^3 n^{-1/2}) = o_p(|\alpha_\lambda|)$. Thus, $I_3 = o_p(I_{22})$. Let $R_1 = I_{21} + I_3 = o_p(I_{22}) = o_p(|\alpha_\lambda|)$, then

$$\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) = (\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T \ell_n'(\beta^0_{\alpha_\lambda}) - \frac{1}{2}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda})^T nI_n(\beta^0_{\alpha_\lambda})(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) + R_1. \quad (5.24)$$

On the other hand, since $\hat{\beta}_{\alpha_\lambda}$ maximizes $\ell_n(\beta_{\alpha_\lambda})$, by Taylor expansion, for some $\beta^*_{\alpha_\lambda}$

that lies between $\hat{\beta}_{\alpha_\lambda}$ and $\hat{\beta}_{\alpha_0}$

$$
\begin{aligned}
0 = \ell_n'(\hat{\beta}_{\alpha_\lambda}) &= \ell_n'(\beta^0_{\alpha_\lambda}) + \{\ell_n''(\beta^0_{\alpha_\lambda}) + nI_n(\beta^0_{\alpha_\lambda})\}(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) - nI_n(\beta^0_{\alpha_\lambda})(\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}) \\
&\quad + \frac{1}{2}\left(\sum_{i=1}^{n} \sum_{j,k=1}^{d_n} \tilde{\ell}_i'''(\beta^*_{\alpha_\lambda})_{jk1}(\hat{\beta}_{\alpha_\lambda j} - \beta^0_{\alpha_\lambda j})(\hat{\beta}_{\alpha_\lambda k} - \beta^0_{\alpha_\lambda k}), \ldots, \right. \\
&\quad \left. \sum_{i=1}^{n} \sum_{j,k=1}^{d_n} \tilde{\ell}_i'''(\beta^*_{\alpha_\lambda})_{jkd_n}(\hat{\beta}_{\alpha_\lambda j} - \beta^0_{\alpha_\lambda j})(\hat{\beta}_{\alpha_\lambda k} - \beta^0_{\alpha_\lambda k}) \right)^T \\
&= J_1 + J_2 - J_3 + J_4. \qquad (5.25)
\end{aligned}
$$

Denote the vector J_2 as $(\nu_1, \ldots, \nu_{|\alpha_\lambda|})^T$ and J_3 as $(\upsilon_1, \ldots, \upsilon_{|\alpha_\lambda|})^T$. Since we have shown that $I_{21} = o_p(I_{22})$, it follows that $\sum_{j=1}^{|\alpha_\lambda|}(\hat{\beta}_{\alpha_\lambda j} - \beta^0_{\alpha_\lambda j})\nu_j = o_p\{\sum_{j=1}^{|\alpha_\lambda|}(\hat{\beta}_{\alpha_\lambda j} - \beta^0_{\alpha_\lambda j})\upsilon_j\}$. Since $\ell_n''(\beta^0_{\alpha_\lambda}) + nI_n(\beta^0_{\alpha_\lambda})$ and $nI_n(\beta^0_{\alpha_\lambda})$ are both symmetric matrices, under Condition (E) we have that $\nu_j = o_p(\upsilon_j)$ for all j, and therefore $J_2 = o_p(J_3)$ component-wise. Since $I_3 = o_p(I_{22})$, similar argument gives that $J_4 = o_p(J_3)$ component-wise. Let $R_2 = J_2 + J_4 = o_p(J_3)$, then $J_1 - J_3 + R_2 = 0$ by (5.25). Using proof by contradiction, it is necessary that $R_2 = o_p(J_1) = o_p\{\ell_n'(\beta^0_{\alpha_\lambda})\}$ component-wise. By solving (5.25) we have that $\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda} = n^{-1}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\{\ell_n'(\beta^0_{\alpha_\lambda}) + R_2\}$. Plug this result into (5.24) we get

$$
\begin{aligned}
\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) &= \{\ell_n'(\beta^0_{\alpha_\lambda}) + R_2\}^T n^{-1}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\ell_n'(\beta^0_{\alpha_\lambda}) \\
&\quad - \frac{1}{2}\{\ell_n'(\beta^0_{\alpha_\lambda}) + R_2\}^T n^{-1}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1} nI_n(\beta^0_{\alpha_\lambda}) n^{-1}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\{\ell_n'(\beta^0_{\alpha_\lambda}) + R_2\} + R_1 \\
&= \ell_n'(\beta^0_{\alpha_\lambda})^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\ell_n'(\beta^0_{\alpha_\lambda}) + R_2^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\ell_n'(\beta^0_{\alpha_\lambda}) \\
&\quad - \frac{1}{2}\ell_n'(\beta^0_{\alpha_\lambda})^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\ell_n'(\beta^0_{\alpha_\lambda}) - R_2^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\ell_n'(\beta^0_{\alpha_\lambda}) \\
&\quad - \frac{1}{2}R_2^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1} R_2 + R_1 \\
&= \frac{1}{2}\ell_n'(\beta^0_{\alpha_\lambda})^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}\ell_n'(\beta^0_{\alpha_\lambda}) - \frac{1}{2}R_2^T \frac{1}{n}\{I_n(\beta^0_{\alpha_\lambda})\}^{-1} R_2 + R_1 \\
&= K_1 - K_2 + R_1.
\end{aligned}
$$

Since $R_2 = o_p\{\ell_n'(\beta^0_{\alpha_\lambda})\}$ component-wise, $K_2 = o_p(K_1)$. Furthermore, by spectral decomposition and Condition (E) we have that $K_1 \geq \|\ell_n'(\beta^0_{\alpha_\lambda})\|^2 n^{-1}\text{eigen}_{\min}[\{I_n(\beta^0_{\alpha_\lambda})\}^{-1}]/2 = \|\ell_n'(\beta^0_{\alpha_\lambda})\|^2 n^{-1}[\text{eigen}_{\max}\{I_n(\beta^0_{\alpha_\lambda})\}]^{-1}/2 \geq O_p(|\alpha_\lambda|n)n^{-1}C_4^{-1} = O_p(|\alpha_\lambda|)$. Thus, $R_1 = o_p(K_1)$ since $R_1 = o_p(|\alpha_\lambda|)$. Since for any $\alpha_\lambda \not\supseteq \alpha_0$, $I_n(\beta^0_{\alpha_\lambda})$ is the covariance matrix of $n^{-1/2}\ell_n'(\beta^0_{\alpha_\lambda})$, it follows that $2K_1$ converges to a Chi-square distribution with degree of freedom $|\alpha_\lambda| - |\alpha_0|$. Therefore, $\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})$ converges to a Chi-square distribution with degree of freedom $|\alpha_\lambda| - |\alpha_0|$ for any $\alpha_\lambda \not\supseteq \alpha_0$. Then, by the corollary of Lemma 1 in Laurent and Massart (2000) as restated in the beginning of the proof, for any positive number ε,

$$\text{pr}\left[\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \geq |\alpha_\lambda| - |\alpha_0| + 2\left\{(|\alpha_\lambda| - |\alpha_0|)\varepsilon\right\}^{1/2} + 2\varepsilon\right] \leq \exp(-\varepsilon).$$

Let $\varepsilon = \gamma_n \log(d_n)(|\alpha_\lambda| - |\alpha_0|)$, where γ_n is any diverging sequence. Then

$$\begin{aligned}
&\text{pr}\left[\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \geq |\alpha_\lambda| - |\alpha_0| + 2\sqrt{(|\alpha_\lambda| - |\alpha_0|)^2\gamma_n \log(d_n)} + 2\gamma_n \log(d_n)(|\alpha_\lambda| - |\alpha_0|)\right] \\
&= \text{pr}\left(\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \geq (|\alpha_\lambda| - |\alpha_0|)\left[1 + 2\left\{\gamma_n \log(d_n)\right\}^{1/2} + 2\gamma_n \log(d_n)\right]\right) \\
&\leq \exp\left\{-\gamma_n \log(d_n)(|\alpha_\lambda| - |\alpha_0|)\right\}.
\end{aligned}$$

Therefore, by using the fact that $\binom{d_n}{k} \leq (d_n e/k)^k$ for any $0 \leq k \leq d_n$, we have that

$$\begin{aligned}
&\text{pr}\left[\sup_{\alpha_\lambda \not\supseteq \alpha_0} \frac{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})}{|\alpha_\lambda| - |\alpha_0|} \geq 1 + 2\left\{\gamma_n \log(d_n)\right\}^{1/2} + 2\gamma_n \log(d_n)\right] \\
&\leq \sum_{|\alpha_\lambda| = |\alpha_0|+1}^{d_n} \text{pr}\left(\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \geq (|\alpha_\lambda| - |\alpha_0|)\left[1 + 2\left\{\gamma_n \log(d_n)\right\}^{1/2} + 2\gamma_n \log(d_n)\right]\right) \\
&\leq \sum_{k=|\alpha_0|+1}^{d_n} \left(\frac{d_n e}{k}\right)^k \exp\left\{-\gamma_n \log(d_n)(k - |\alpha_0|)\right\} \\
&= \sum_{k=|\alpha_0|+1}^{d_n} \left(\frac{e}{k}\right)^k d_n^{\{k-(k-|\alpha_0|)\gamma_n\}}. \qquad (5.26)
\end{aligned}$$

Since γ_n diverges to infinity and $k = O(k - |\alpha_0|)$ as $k \to \infty$ under Condition (H), $d_n^{\{k-(k-|\alpha_0|)\gamma_n\}} =$

$o(d_n^{-1})$. Moreover, $(e/k)^k < 1$ for all $k \geq 3$. Therefore, (5.26) goes to 0 as $n \to \infty$. It follows that

$$\sup_{\alpha_\lambda \not\supseteq \alpha_0} \frac{1}{|\alpha_\lambda| - |\alpha_0|} \left\{ \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \right\} = O_p\left[1 + 2\left\{ \log(d_n) \right\}^{1/2} + 2\log(d_n) \right] = O_p\{\log(d_n)\}.$$

□

CHAPTER 6: SUMMARY AND FUTURE RESEARCH

In this dissertation we have studied the regularized variable selection procedure in both Cox proportional hazards model and additive hazards model with a case-cohort design and a diverging number of parameters. We focused on the smoothly clipped absolute deviation (SCAD) penalty, but the results can be extended to other penalty functions as well. We investigated both the asymptotic properties and finite sample performance of the variable selection procedures. Due to the non-predictability of the weight function $\rho_i(t)$ in the estimating equations, we employed modern empirical process techniques instead of traditional martingale theorems in most of the theoretical development. To accommodate the common features of case-cohort studies, we considered high censoring rates and large sample sizes in the simulation studies.

In Chapter 3, we proved that the SCAD-penalized variable selection procedure can identify the true model with probability tending to one as sample size goes to infinity under Cox proportional hazards model with a case-cohort design and a diverging dimension. The consistency and asymptotic normality of the penalized estimator were also established. Based on the simulation results, the BIC-based tuning parameter selection method outperforms the AIC-based one. The variable selection procedure was applied to the Busselton Health Study. In Chapter 4, we extended the SCAD-penalized variable selection procedure to additive hazards model with a stratified case-cohort design and a diverging dimension. We again proved the model selection consistency of the procedure as well as the consistency and asymptotic normality of the penalized estimator. We proposed a penalized cross-validation method for tuning parameter selection for additive hazards model and evaluated its finite

sample performance via simulation. It is found that the proposed penalized tuning parameter selection method outperforms the conventional five-fold cross-validation method. The variable selection procedure was applied to the Atherosclerosis Risk in Communities (ARIC) study. In Chapter 5, we shifted our focus to the tuning parameter selection for regularized variable selection method under Cox proportional hazards model with a diverging number of parameters in a random sample. We proposed a generalized information criterion (GIC) for tuning parameter selection and proved that, under certain conditions on the signal strength of the true model and the diverging sequence a_n, GIC can identify the true model with probability tending to one as sample size goes to infinity. We then conducted simulations to compare the variable selection performance of GIC with four different choices of a_n: 2, $\log(n)$, $\log\{\log(n)\}\log(d_n)$, and $\log\{\log(d_n)\}\log(d_n)$. It is found that the GIC with $a_n = \log\{\log(d_n)\}\log(d_n)$ gives better overall performance and therefore we recommended it for practical use. The proposed tuning parameter selection method was applied to the Framingham Heart Study.

There are several future directions where we can extend the research presented in this dissertation.

First, we have only investigated in this dissertation the scenarios where the dimension of the model is smaller than the sample size ($p \ll n$). With the increasing availability of the so-called "Big Data", it is desirable to extend the proposed variable selection procedures and the tuning parameter selection methods to the high-dimensional realm where $p \gg n$. The theoretical framework used in this dissertation will no longer be valid for this scenario. More advanced dimension-reduction techniques need to be developed for the case-cohort design with failure time outcome. One potential starting point could be to introduce the iterative sure independence screening (ISIS) method proposed by Fan and Lv (2008) into the Cox proportional hazards model and additive hazards model with a case-cohort design.

Second, we can extend the current variable selection methods to a case-cohort design

with multivariate failure time outcome. Multivariate failure time data arise frequently from biomedical research. For instance, elderly people may develop both coronary heart disease (CHD) and stroke; patients with kidney failure who are on dialysis may experience multiple events of infection. The potential correlation among failure times of different events poses additional challenge in the theoretical development. Meanwhile, a more efficient weight function is available for the case-cohort design with multivariate failure time outcome (Kim et al. 2013), the properties of which have not been studied in the context of regularized variable selection and tuning parameter selection.

Last, as a natural continuation, the proposed GIC statistic for tuning parameter selection in Chapter 5 needs to be extended from regular Cox proportional hazards model to one with a case-cohort design. The main challenge is to incorporate the weight function $\rho_i(t)$, which is not independent across subjects, into the empirical process techniques used to derive the probability bounds of various random quantities.

BIBLIOGRAPHY

Aalen, O. (1980), "A Model for Nonparametric Regression Analysis of Counting Processes," in *Lecture Notes in Statistics 2*, New York: Springer-Verlag, pp. 1–25.

Akaike, H. (1973), "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, 60, 255–265.

Andersen, P. K. and Gill, R. D. (1982), "Cox's regression model for counting processes: a large sample study," *The Annals of Statistics*, 10, 1100–1120.

Ballantyne, C. M., Hoogeveen, R. C., Bang, H., Coresh, J., Folsom, A. R., Heiss, G., and Sharrett, A. R. (2004), "Lipoprotein-Associated Phospholipase A2, high-sensitivity C-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study," *Circulation*, 109, 837 – 842.

Barlow, W. E. (1994), "Robust Variance Estimation for the Case-Cohort Design," *Biometrics*, 50, pp. 1064–1072.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000), "Exposure stratified case-cohort designs." *Lifetime data analysis*, 6, 39–58.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, pp. 2350–2383.

Buehler, R. J. and Feddersen, A. P. (1963), "Note on a Conditional Property of Student's t," *The Annals of Mathematical Statistics*, 1098–1100.

Cai, J., Fan, J., Li, R., and Zhou, H. (2005), "Variable selection for multivariate failure time data," *Biometrika*, 92, 303–316.

Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

Craven, P. and 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," *Australian Journal of Medicine*, 2, 714–718.

Dawber, T. R. (1980), *The Framingham Study: the epidemiology of atherosclerotic disease*, vol. 84, Cambridge, MA: Harvard University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, pp. 407–451.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman and Hall.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

— (2002), "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *The Annals of Statistics*, 30, 74–99.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fan, Y. and Tang, C. Y. (2013), "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society. Series B (Methodological)*, 75, 531–552.

Fu, W. J. (1998), "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, pp. 397–416.

Gaiffas, S. and Guilloux, A. (2012), "High-dimensional additive hazards models and the Lasso," *Electronic Journal of Statistics*, 6, 522–546.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning: data mining, inference and prediction*, Springer, 2nd ed.

Huber, P. J. (1973), "Robust regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics*, 1, 799–821.

Huffer, F. W. and McKeague, I. W. (1991), "Weighted Least Squares Estimation for Aalen's Additive Risk Model," *Journal of the American Statistical Association*, 86, pp. 114–129.

Kalbfleisch, J. D. and Lawless, J. F. (1988), "Likelihood Analysis of Multi-State Models For Disease Incidence and Mortality," *Statistics in Medicine*, 7, 149–160.

Kang, S. and Cai, J. (2009), "Marginal hazards model for case-cohort studies with multiple disease outcomes," *Biometrika*, 96, 887–901.

Kang, S., Cai, J., and Chambless, L. (2013), "Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the Atherosclerosis Risk in Communities (ARIC) study." *Biostatistics*, 14, 28–41.

Kim, S., Cai, J., and 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., and Bartholomew, H. C. (2003), "Serum Ferritin and Cardiovascular Disease: A 17-Year Follow-up Study in Busselton, Western Australia," *American Journal of Epidemiology*, 158, 144–149.

Kong, S. and Nan, B. (2014), "Non-Asymptotic Oracle Inequalities for the High-Dimensional Cox Regression via Lasso." *Statistica Sinica*, 24, 25–42.

Kulich, M. and Lin, D. (2000), "Additive hazards regression for case-cohort studies," *Biometrika*, 87, 73–87.

— (2004), "Improving the Efficiency of Relative-Risk Estimation in Case-Cohort Studies," *Journal of the American Statistical Association*, 99, 832–844.

Laurent, B. and Massart, P. (2000), "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, 1302–1338.

Leeb, H. and Potscher, B. M. (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

— (2006), "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" Cowles Foundation Discussion Papers 1444, Cowles Foundation for Research in Economics, Yale University.

Leng, C. and Ma, S. (2007), "Path consistent model selection in additive risk model via Lasso," *Statistics in Medicine*, 26, 3753–3770.

Lin, D. (2000), "On Fitting Cox's Proportional Hazards Models to Survey Data," *Biometrika*, 87, pp. 37–47.

Lin, D. and Ying, Z. (1993), "Cox regression with incomplete covariates measurements," *Journal of the American Statistical Association*, 88, 1341–1349.

— (1994), "Semiparametric Analysis of the Additive Risk Model," *Biometrika*, 81, pp. 61–71.

Lin, W. and Lv, J. (2013), "High-Dimensional Sparse Additive Hazards Regression," *Journal of the American Statistical Association*, 108, 247–264.

Lv, J. and Fan, Y. (2009), "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares," *The Annals of Statistics*, 37, pp. 3498–3528.

Ma, S. and Huang, J. (2005), "Lasso method for additive risk models with high dimensional covariates," in *Technical Report 347*, Department of Statistics and Actuarial Science, University of Iowa, Iowa.

Mallows, C. L. (1973), "Some Comments on CP," *Technometrics*, 15, pp. 661–675.

Martinussen, T. and Scheike, T. H. (2009), "Covariate Selection for the Semiparametric Additive Risk Model." *Scandinavian Journal of Statistics*, 36, 602 – 619.

Park, M. Y. and Hastie, T. (2007), "L1-Regularization Path Algorithm for Generalized Linear Models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69, pp. 659–677.

Peng, H. and Fan, J. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, 32, 928–961.

Portnoy, S. (1988), "Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity," *The Annals of Statistics*, 16, 356–366.

Potscher, B. M. and Leeb, H. (2009), "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding," *Journal of Multivariate Analysis*, 100, 2065 – 2082.

Prentice, R. L. (1986), "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, 73, 1–11.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, pp. 846–866.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, pp. 461–464.

Self, S. G. and Prentice, R. L. (1988), "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies," *The Annals of Statistics*, 16, 64–81.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society (Series B)*, 58, 267–288.

— (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385–395.

Van de Geer, S. A. (2008), "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, 614–645.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer.

Wacholder, S., Gail, M. H., Pee, D., and Brookmeyer, R. (1989), "Alternative Variance and Efficiency Calculations for the Case-Cohort Design," *Biometrika*, 76, pp. 117–123.

Wang, H., Li, B., and Leng, C. (2009), "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society Series B*, 71, 671–683.

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, 94, 553–568.

Wang, T. and Zhu, L. (2011), "Consistent tuning parameter selection in high dimensional sparse linear regression," *Journal of Multivariate Analysis*, 102, 1141 – 1151.

Wu, Y. (2012), "Elastic Net for Cox's Proportional Hazards Model with a Solution Path Algorithm," *Statistica Sinica*, 22, 271–294.

Zhang, C.-H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.

Zhang, H. H. and Lu, W. (2007), "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, 94, 691–703.

Zhang, Y., Li, R., and Tsai, C.-L. (2010), "Regularization Parameter Selections via Generalized Information Criterion," *Journal of the American Statistical Association*, 105, 312–323, pMID: 20676354.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, pp. 301–320.

Zou, H. and Zhang, H. H. (2009), "On the Adaptive Elastic-Net with a Diverging Number of Parameters," *The Annals of Statistics*, 37, pp. 1733–1751.