Variable Selection for General Transformation Models

LI, Jianbo

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Statistics

The Chinese University of Hong Kong
June 2011

Thesis/Assessment Committee

Professor Cheung Siuhung (Chair)
Professor Gu, Minggao (Thesis Supervisor)
Professor Chan Pingshing (Committee Member)
Professor Zhu, Lixing (External Examiner)

Abstract of thesis entitled:

Variable Selection for General Transformation Models

Submitted by LI, Jianbo

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in June 2011

General transformation models are a class of semiparametric survival models. The models generalize simple transformation models with more flexibility in modeling data coming from statistical practice. The models include many popular survival models as their special cases, e.g., proportional hazard Cox regression models, proportional odds models, generalized probit models, frailty survival models and heteroscedastic hazard regression models etc. Although the maximum marginal likelihood estimate of parameters in general transformation models with interval censored data is very satisfactory, its large sample properties are open. In this thesis, we will consider the problem and use discretization technique to establish the large sample properties of maximum marginal likelihood estimates with interval censored data.

In general, to reduce possible model bias, many covariates will be collected into a model. Hence a high-dimensional regression model is built. But at the same time, some non-significant variables may be also included in. So one of tasks to build an efficient survival model is to select significant variables. In this thesis, we will focus on the variable selection for general transformation models with ranking data, right censored data and interval censored data. Ranking data are widely seen in epidemiological studies, population pharmacokinetics and economics. Right censored data are the most common data in clinical trials. Interval censored data are another type common data in medical studies, financial, epidemiological, demographical and sociological studies. For example, a patient visits a doctor with a prespecified schedule. In his last visit, the doctor did not find occurrence of an interested event but at the current visit, the doctor found the event has

occurred. Then the exact occurrence time of this event was censored in an interval bracketed by the two consecutive visiting dates. Based on rank-based penalized log-marginal likelihood approach, we will propose an uniform variable selection procedure for all three types of data mentioned above. In the penalized marginal likelihood function, we will consider non-concave and Adaptive-LASSO (ALASSO) penalties. For the non-concave penalties, we will adopt HARD thresholding, SCAD and LASSO penalties. ALASSO is an extended version of LASSO. The key of ALASSO is that it can assign weights to effects adaptively according to the importance of corresponding covariates. Therefore it has received more attention recently. By incorporating Monte Carlo Markov Chain stochastic approximation (MCMC-SA) algorithm, we also propose an uniform algorithm to find the rank-based penalized maximum marginal likelihood estimates. Based on the numeric approximation for marginal likelihood function, we propose two evaluation criteria — approximated GCV and BIC — to select proper tuning parameters. Using the procedure, we not only can select important variables but also be able to estimate corresponding effects simultaneously. An advantage of the proposed procedure is that it is baseline-free and censoring-distribution-free. With some regular conditions and proper penalties, we can establish the $\sqrt{n}$-consistency and oracle properties of penalized maximum marginal likelihood estimates. We illustrate our proposed procedure by some simulations studies and some real data examples. At last, we will extend the procedures to analyze stratified survival data.

Keywords: General transformation models; Marginal likelihood; Ranking data; Right censored data; Interval censored data; Variable selection; HARD; SCAD; LASSO; ALASSO; Consistency; Oracle.

摘要

廣義變換模型是一類半參數生存模型。該模型在數據建模靈方面極大地推廣了線性變換模型，很多常見生存分析模型可看作是它的特例。例如：比例風險 Cox 迴歸模型，比例勝算迴歸模型，廣義 probit 模型，脆弱性生存模型，異方差生存模型等等。基於區間刪失數據，雖然廣義變換模型中參數的極大邊際似然估計已經非常的好，但是它的大樣本性質目前仍然是一個空白。在本文中，我們將考慮這個問題並通過離散化技術建立基於區間刪失數據的極大邊際似然估計的大樣本性質。

為了減小模型可能存在的偏，通常我們會把盡可能多的協變量包含到所考慮的模型中來，從而建立了一個高維模型。然而這樣做也有可能把一些非顯著變量也考慮進來，從而可能會給模型帶來更大的偏。所以建立一個有效模型的任務之一就是變量選擇。在本文中，我們將考慮基於秩數據，右刪失數據和區間刪失數據的廣義變換模型的變量選擇問題。基於秩的懲罰邊際似然方法，我們對這三種類型數據提出一個統一的變量選擇程序。在邊際懲罰似然函數中，我們主要考慮非凹懲罰函數(LASSO，SCAD 和 HARD)和 Adaptive－LASSO(ALASSO)懲罰函數。基於蒙特卡羅馬爾可夫鏈隨機近似算法，我們為廣義變換模型的變量選擇提出了一個有效的隨機近似算法。通過對邊際似然函數的數值近似，我們還提出了兩個最優懲罰參數的選擇標準——近似 GCV 和 BIC。在一定的規則條件下，我們可以建立懲罰極大邊際似然估計的一致性和 oracle 性質。最后，我們把所提方法推廣到了分層生存數據研究中。

關鍵詞：廣義變換模型；邊際似然；秩數據；右刪失數據；區間刪失數據；變量選擇；SCAD；LASSO；HARD；ALASSO；一致性；Oracle。

Acknowledgement

I would like to thank my supervisor, Professor Gu, Ming Gao, for his inspiring and encouraging guidance and friendly way to guide me into deep understanding of this work. Without his consistent, patient and illuminating guidance, this thesis would not have reached its present form. His serious working attitude, broad knowledge and noble personality set a good example for me.

I would like to thank Professor Zhu Lixing, Professor Cheung Siuhung and Professor Chan Pingshing, who have served on my dissertation committee, for their time and good advice. I would like to thank the Department of Statistics for providing me the good study and research environment. I am also grateful all the current classmates and former graduate students for their help, support and good suggestions during my PhD study in CUHK.

Finally, special thanks are given my parents, my wife and my daughter for their love, support and encouragement through these years.

This work is dedicated to

my parents

Li Yueming and Zhu Jingying,

my wife

Liu Honghua,

and my daughter

Li Yile.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Simple transformation models [21] can flexibly and concisely reflect the relationship between survival time or duration time and its corresponding covariates. The models include many popular survival models as their special cases. Simple transformation models have been widely used in economics, biostatistics, pharmacokinetics, financial risk managements and many other areas. Gu, *et al.* (2005) [40] extended the simple transformation models with more generality and proposed a class of general transformation models and corresponding rank-based maximum marginal likelihood estimation procedure. They also proposed a three-stage MCMC stochastic approximation algorithm to find the rank-based maximum marginal likelihood estimate. In general, to reduce possible model bias, many covariates are collected into a model. At the same time, some non-significant variables may be also included in which, in turn, may enlarge model bias. Motivated by this, in this thesis, we will consider variable selection for general transformation models with ranking data, right censored data and interval censored data.

1.1 Simple Transformation Models

Let $T \in R^+$ be the survival time variable and $Z \in R^p$ be the corresponding covariate vector. The simple transformation models [21] are given by

$$h(T) = Z^T\beta + \varepsilon \tag{1.1}$$

where $h(\cdot)$ is an unknown link function and usually assumed to be a monotonically increasing function, β is a p-dimensional regression parameter vector and ε is a random

error with known cumulative density function $F(\cdot)$.

Models (1.1), actually, are semi-parametric because of the "nuisance" parameter $h(\cdot)$. They can flexibly capture the relationship between survival time and its corresponding covariates. They include many popular models as their special cases. For example, when $F(\cdot)$ is the standard extreme value type-I cumulative density function, models (1.1) reduce to proportional hazards Cox regression models [17, 18]; when $F(\cdot)$ is the standard logistic cumulative density function, models (1.1) reduce to proportional odds regression models [7] and $F(\cdot)$ is the standard normal cumulative density function, models (1.1) become generalized probit regression models [12, 66, 67].

Because of the model flexibility, many authors have proposed lots of effective estimation procedures for β in (1.1) with uncensored or censored data. Dabrowska and Doksum (1988) [21] proposed a partial likelihood method. Murphy *et al.* (1997) [52] gave a maximum semiparametric likelihood approach. Other methods include maximum marginal likelihood [48], rank approximation [20, 55], profile likelihood [15], generalized estimating equation [13, 16, 30, 31, 72], nonparametric maximum likelihood [76, 77] etc. Although the approaches above can estimate β efficiently, most of them need to estimate "nuisance" parameter $h(\cdot)$. This loses the baseline-free property enjoyed by proportional hazard regression models. Although maximum marginal likelihood and estimation equation methods do not depend on baseline function, they need to estimate censoring cumulative density function. Gu, *et al.* (2005) [40] extended the model (1.1) and proposed a class of general transformation models. In their paper, they proposed a rank-based maximum marginal likelihood estimation procedure to find the rank-based maximum marginal likelihood estimate. An advantage of their estimation procedure is that it is baseline-free and censoring-distribution-free. Comparing with the approaches mentioned above , it should be a preferable alternative for the statistical inference of survival data.

1.2 General Transformation Models

It can be easily shown that the simple transformation models (1.1) are equivalent to

$$S_Z(t) = g^{-1}(g(S_0(t)) - Z^T\beta) \tag{1.2}$$

where $S_Z(t)$ is the conditional survival function of T given covariate vector Z; $S_0(t)$ is the unknown baseline survival function with $Z = \mathbf{0}$; $g^{-1}(t) = 1 - F(t)$ and $F(t)$ is defined in (1.1). Relaxing the special structure of covariates and random error term in simple transformation models (1.1) or (1.2), Gu, *et al.* (2005) [40] proposed a class of general transformation models. The general transformation models, in terms of survival function, are given by

$$S_Z(t) = \Phi(S_0(t), Z, \boldsymbol{\beta}) \tag{1.3}$$

where $S_0(\cdot)$ and $S_Z(t)$ are the same as the ones in (1.2); $\boldsymbol{\beta}$ is a parameter vector including regression parameters with respect to Z and model transformation parameters in $\Phi(u, v, w)$ (if applicable); $\Phi(u, v, w)$ is assumed to be known and satisfies that $\Phi(0, v, w) = 0$ and $\Phi(1, v, w) = 1$ for any v and w. To conduct statistical inference conveniently, they also assume the following restrictions on $\Phi(u, v, w)$:

- It is an increasing continuous function with respective to u for any v and w;
- It is first-order and second-order differentiable for u and w respectively for any v.

General transformation models (1.3) are non-trivially more general than the models (1.1) or (1.2). Models (1.3) include not only the simple transformation models, but also frailty models, heteroscedastic hazard regression models [42] and other important survival models. More examples please refer to Gu, *et al.* (2005) [40] and therein. They are also a class of semi-parametric models because of "nuisance" parameter $S_0(t)$.

Gu, *et al.* (2005) [40] proposed a rank-based maximum marginal likelihood estimation procedure for $\boldsymbol{\beta}$ in models (1.3) with interval censored data. The estimation procedure is free of baseline survival function and censoring distribution, enjoyed by partial likelihood approach. In their paper, they gave a three-stage Markov Chain Monte Carlo stochastic approximation (MCMC-SA) algorithm to find rank-based maximum marginal likelihood estimates (MMLE). Through many simulation studies, they empirically showed that MMLE may be consistent and distributed asymptotically normally. Huang (2005) [43] established the asymptotic properties of MMLE by discretization technique and martingale methods. However, his proofs are not very rigorous and he only considered no censoring case. Wu (2008) [71] gave a rigorous proof for the asymptotic properties of MMLE with right censored data by similar discretization technique. In her thesis, she

also considered general transformation models with time-varying covariates and general transformation model with measurement error. Ni (2008) [53] considered the misspecified general transformation models and mixed-effect general transformation models with ranking data. Based on maximum marginal likelihood estimation method of Gu, *et al.* (2005) [40], he proposed a quasi maximum marginal likelihood estimation procedure for β in misspecified general transformation models. By discretization technique, he showed that quasi maximum marginal likelihood estimate (QMMLE) is consistent and distributed asymptotically normally. Based on the asymptotic properties, he also proposed Wald test, Lagrange multiplier test and information matrix test for QMMLE. At last, he also established large sample properties of MMLE in mixed-effect general transformation models with ranking data. For the interval censored data, the theoretical properties of MMLE have been not studied because of its complicated censoring mechanism. This is one of our interests in this thesis.

Zeng and Lin (2007a, b) [76, 77] proposed another class of general transformation models defined by, in terms of cumulative intensity function,

$$\Lambda(t|Z) = G(\int_0^t R^*(s)\exp(Z(s)^T\beta)d\Lambda(s)) \tag{1.4}$$

where $G(\cdot)$ is a continuously differentiable and strictly increasing function with $G(0) = 0$ and $G(\infty) = \infty$, $R^*(\cdot)$ is an indicator process, β is a regression parameter vector and $\Lambda(\cdot)$ is an unspecified increasing function. Zeng *et al.* (2009) [75] and Zeng and Lin (2010) [78] also extended the models (1.4). The models are not the direct extension of simple transformation models (1.1) or (1.2) because models (1.4) do not include generalized probit models, a nature and important transformation models, as their special cases. While models (1.4) also include proportional hazard regression models, proportional odds regression models and corresponding time-varying covariate models, it can not deal with interval censored data as like models (1.3).

1.3 Variable Selection Methods

Similar to the context of ordinary linear regression, to reduce possible model bias in the analysis of survival data using general transformation models (1.3), we usually collect

as many covariates as possible into the models. However, at the same time, some non-significant covariates may be also included in, which, in turn, may enlarge model bias. Therefore it is meaningful to select important covariates in all the pre-collected covariates by some variable selection procedure.

Many variable selection approaches have been proposed for ordinary linear regression models and some of them have been extended to the context of survival data analysis. Fan and Lv (2010) [27] overviewed the variable selection methods in details. The popular variable selection methods are usually conducted by minimizing penalized sum of squared error or maximizing penalized log-likelihood function. So, in the sense, choosing a variable selection method is equivalent to choose an adequate penalty function $p_\lambda(\cdot)$. To make the minimization of penalized sum of squared error or maximization of penalized log-likelihood function an effective variable selection procedure, the penalty function should be irregular at origin, that is, $p'_\lambda(0_+) > 0$ [25].

In classical model selection procedure, L_0-penalty function with the form of $p_\lambda(\theta) = \lambda I(\theta \neq 0)$ (called entropy penalty in Wavelet studies [3, 23]) has a nice interpretation of best subset selection and admits nice sample properties [6]. However, the computation is infeasible in high dimensional statistical endeavors [27].

The nature generalization of L_0-penalty is L_q-penalty ($0 < q < 2$) with the form of $p_\lambda(|\theta|) = \lambda|\theta|^q$ [35], which bridges the best subset penalty (L_0-penalty) and L_2-penalty. Obviously, L_0-penalty enjoys variable selection feature while L_2-penalty can give stable estimates. Tibshirani (1996) [69] proposed a L_1-penalty function, a special case of L_q-penalty, called LASSO, with the form of

$$p_\lambda(|\theta|) = \lambda|\theta| \tag{1.5}$$

Tibshirani (1997) [68] extended the LASSO penalty to consider variable selection for Cox's regression models by maximizing penalized partial likelihood function. The variable selection method with LASSO penalty is actually a shrinkage one. So its corresponding penalized estimate can be not taken as the final estimate.

For a good penalty function, Fan and Li (2001) [28] gave three conditions for its corresponding penalized estimate, that is,

- **Sparsity**: The resulting estimator automatically sets small estimated coefficients to

zero to accomplish variable selection and reduce model complexity;

- **Unbiasedness**: The resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias;

- **Continuity**: The resulting estimator is continuous in the data to reduce instability in model prediction.

From the penalty (1.5), we can see that when the tuning parameter λ is very large, the effects with large values will be heavily penalized and shrunk to some extent. So the penalized estimate with LASSO penalty may suffer from relative large bias. Zou (2006) [82] proposed an extended version of LASSO, called Adaptive-LASSO (ALASSO), with the form of

$$p_\lambda(|\theta|) = \lambda\tau|\theta| \tag{1.6}$$

where the weight τ is chosen by data adaptively. When the true value of θ is large, it will assign smaller weight to θ, otherwise it will assign larger weight to θ. Usually $\tau = |\tilde{\theta}|^{-\gamma}$ for some $\gamma > 0$, where $\tilde{\theta}$ is a consistent estimate of θ. The parameter γ can be a prespecified positive real number or can be seen as a tuning parameter, which can be selected on a grid of points by BIC, GCV or other evaluation criteria. Zou (2006) [82] showed the consistency and oracle properties of penalized estimates with ALASSO penalty in the context of ordinary linear models and generalized linear models. Following the penalty (1.6), we can reduce the penalty for large effects and obtain consistent estimates. Note that when $\tilde{\theta}$ is very close to 0, the penalty will become very large. Consequently, the corresponding estimate will go to 0 very fast. Zhang and Lu (2007) [79] applied ALASSO with $\gamma = 1$ to select important variables in Cox's regression models by penalized partial likelihood approach. They also gave the proof of consistency and oracle properties of penalized estimates with ALASSO penalty. Lu and Zhang (2007) [51] studied variable selection method with ALASSO penalty for proportional odds model by penalized maximum marginal likelihood approach. However, they did not gave the proofs of large sample properties of corresponding estimate because of the complicated marginal likelihood function. We will solve the problem using the general transformation models (1.3) in Chapters 2, 3 and 4 with ranking data, right censored data and interval censored data respectively. In this thesis, following Zhang and Lu (2007) [79] and Lu and Zhang (2007) [51], we also see γ

as a prespecified positive real number and take $\gamma = 1$ in all the simulation studies and applications of variable selection methods with ALASSO penalty. However, we will give the oracle properties of penalized estimates with ALASSO penalty for any $\gamma > 0$.

Fan (1997) [24] proposed another penalty function – HARD thresholding penalty function with the form of

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda) \tag{1.7}$$

This penalty was improved by Antoniadis (1997) [2]. The penalty does not overpenalize the effects with large values. In the penalized least squares, when design matrix X is orthonormal, best subset selection and stepwise deletion are equivalent to the penalized least squares with HARD thresholding penalty. However, this states does not hold in other cases [25]. Fan (2002) [25] summarized that the estimates with both the L_q- and HARD thresholding penalties do not satisfy the three mathematical conditions mentioned above. Fan (1997) [24] proposed a penalty function with corresponding estimates satisfying the conditions – smoothly clipped absolute deviation (SCAD) penalty with the form of

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}, \tag{1.8}$$

for some $a > 2$ and $\theta > 0$. Based on Bayesian statistical point of view, Fan and Li (2001) [28] suggested $a = 3.7$. In this thesis, we also take the value for a. Fan and Li (2001) [28] showed that the penalized estimates with SCAD penalty enjoy consistency and oracle properties in the context of linear models. Fan and Li (2002) [25] extended the variable selection methods with SCAD penalty to Cox's regression models and frailty Cox's regression models. They also proved that the penalized estimates with SCAD and HARD penalties enjoy oracle properties and they are also consistent in the context of Cox regression models.

Figure 1.1 displays the plot of non-concave penalty functions — LASSO, HARD and SCAD. From the figure, we can see that both SCAD and HARD thresholding do not overpenalize θ with large values since it is a constant when $|\theta|$ is very large while LASSO would assign large penalty to such effects. In addition, SCAD penalty is smoother than HARD and hence SCAD can produce continuous sparse estimates. On the other hand, we can also find that when θ is very close to 0, all the three methods can shrink θ to 0 very fast.

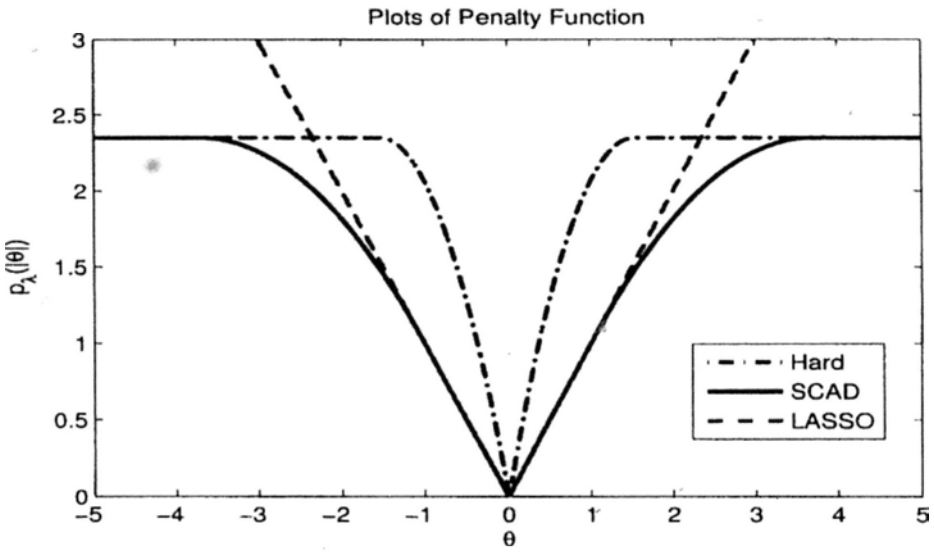

Figure 1.1: The plot of non-concave penalty functions.

Other variable selection methods also include Bayesian variable selection [29, 49] and Dantzig selector [4, 11] etc. Fan and Li (2006, 2010) [26, 27] gave an overview of the variable selection methods following penalized least squares and penalized log-likelihood methods.

Since the penalty functions are irregular at origin, some approximations of penalty functions should be needed for the stability of variable selection procedure. Fan and Li (2001) [28] proposed a local square approximation method for penalty functions and the method has been used widely in the context of variable selection. In this thesis, we also follow the method to approximate all the penalty functions.

Another key of variable selection is the choice of the proper tuning parameter λ. Many selection criteria for λ have been proposed by various authors and the criteria are widely used in model selection and variable selection. Among the criteria, k-fold cross validation (k-fold-CV), BIC and generalized cross validation (GCV) are most popular ([1, 10, 19, 28, 36, 69, 70, 82]). However, in the context of survival analysis, there are no closed forms for k-fold-CV, BIC and GCV. Craven and Wahba (1979) [19] proposed an approximated GCV for the selection of degree of spline functions. Fan and Li (2002) [25], Lu and Zhang (2007) [51], Zhang and Lu (2007) [79] and others applied the approximated GCV in the context of survival analysis to select tuning parameters. In our studies, we will also use the approximated GCV to select proper tuning parameter λ for ranking data

and right censored data while we will propose a BIC-type criterion to select the proper tuning parameter for interval censored data.

In this thesis, we mainly consider the variable selection with HARD, SCAD LASSO and ALASSO penalties for general transformation models (1.3). We will conduct the variable selection procedures by maximizing the following rank-based penalized log-marginal likelihood function

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{i=1}^{p} p_\lambda(|\beta_i|) \tag{1.9}$$

where $\ell(\boldsymbol{\beta})$ is the log-marginal likelihood function for model (1.3), n is the number of observations, $\lambda \geq 0$ is the tuning parameter or penalty parameter and p is the number of regression parameters with respect to covariates Z in (1.3). Note that the parameters in penalty term of (1.9) are the partial components of $\boldsymbol{\beta}$ in (1.3) when $\boldsymbol{\beta}$ contains model transformation parameters.

The rest of this thesis is organized as follows. In Chapter 2, we study the variable selection methods for general transformation models with ranking data. We will propose an effective algorithm to implement the procedure. With proper penalty function, we establish the consistency and oracle properties of penalized maximum marginal likelihood estimate (PMMLE). We also further extend the proposed variable selection procedures into the stratified general transformation models. Some simulation studies and Hong Kong horse racing data analysis will be used to illustrate our proposed variable selection procedures. In Chapter 3, we consider the variable selection for general transformation models with right censored data. The algorithm used for ranking data case will be used to implement the variable selection procedures. The $\sqrt{n}$-consistency and oracle properties for right censored data will be established. We use some simulation studies and *PBC* data analysis to illustrate the proposed procedures for right censored data. In Chapter 4, we will firstly study the asymptotic properties of rank-based maximum marginal likelihood estimate of parameters in general transformation models with interval censored data, based on which we further study the variable selection for the models with interval censored data. At last, we make some conclusions and present some further studies in Chapter 5.

□ **End of chapter.**

Chapter 2

Variable Selection for General Transformation Models with Ranking Data

Ranking data are widely seen in epidemiological studies, population pharmacokinetics and economics. In this chapter, we will consider variable selection for general transformation models with ranking data by non-concave (SCAD, HARD thresholding and LASSO) and ALASSO penalty. The variable selection procedures will be done by maximizing rank-based penalized log-marginal likelihood function. Based on MCMC stochastic approximation (MCMC-SA) algorithms in Gu, *et al.* (2005)[40] and Gu and Kong (1998) [39], we propose a three-step MCMC-SA algorithm to implement the variable selection. We also establish consistency and oracle properties for penalized maximum marginal likelihood estimates. We illustrate the proposed procedures by some simulation studies and Hong Kong horse racing data analysis.

2.1 Penalized Log-marginal Likelihood

Denote $T \in R^+$ as the failure time variable and $Z \in R^p$ as the corresponding covariate vector. It is assumed that T and Z are modeled by the general transformation models (1.3). In this Chapter, we assume that $\boldsymbol{\beta}$ only contains the regression parameters for easy presentation of our proposed variable selection procedures. It does not add any difficulties

when $\boldsymbol{\beta}$ includes model transformation parameters, which should be not penalized in variable selection procedures.

Let $\{Z_i\}_{i=1}^n$ be n i.i.d. copies of the population Z, $\mathcal{R}_n = (r_1, r_2, \cdots, r_n)^T$ be the observed ranking vector and $\mathcal{C}_n = \{(t_1, t_2, \cdots, t_n)^T : t_{\alpha_1} < t_{\alpha_2} < \cdots < t_{\alpha_n}\}$ be the set of underlying failure times consistent with the ranking $\mathcal{R}_n$, where α_i is an antirank, that is, $\alpha_i = j$ if and only if $r_j = i$. Denote T_i as the underlying failure time for ith individual, $\mathbf{T}_n = (T_1, T_2, \cdots, T_n)^T$ and $\mathbf{Z}_n = (Z_1, Z_2, \cdots, Z_n)^T$. Then the rank-based marginal likelihood function for models (1.3) with ranking data is given by

$$\begin{aligned} L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n) &= P_r(\mathcal{R}_n|\mathbf{Z}_n) = P_r(\mathbf{T}_n \in \mathcal{C}_n|\mathbf{Z}_n) \\ &= \int_{\mathcal{C}_n} (-1)^n \prod_{i=1}^n \phi(S_0(t_i), Z_i, \boldsymbol{\beta}) \prod_{i=1}^n dS_0(t_i) \qquad (2.1) \\ &= \int_{\mathcal{D}_n} \prod_{i=1}^n \phi(1-u_i, Z_i, \boldsymbol{\beta}) \prod_{i=1}^n du_i \end{aligned}$$

where $\phi(u, v, w) = \frac{\partial \Phi(u,v,w)}{\partial u}$ and $\mathcal{D}_n = \{(u_1, u_2, \cdots, u_n) : 0 \leq u_{\alpha_1} < u_{\alpha_2} < \cdots < u_{\alpha_n} \leq 1\}$. The last equality of (2.1) holds because of the simple transformation $u_i = 1 - S_0(t_i)$. Following (2.1), we can get the rank-based penalized log-marginal likelihood function as follows

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n \sum_{i=1}^p p_\lambda(|\beta_i|) \qquad (2.2)$$

where $\ell(\boldsymbol{\beta}) = \log(L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n))$ and $p_\lambda(\cdot)$ is a penalty function, which is irregular at origin. Then the rank-based penalized maximum marginal likelihood estimate (PMMLE) of $\boldsymbol{\beta}$ can be obtained by maximizing (2.2) with respect to $\boldsymbol{\beta}$. With proper penalty function $p_\lambda(\cdot)$, some components of $\hat{\boldsymbol{\beta}}_n$ will be zero, then they will disappear in selected models, which reaches to the purpose of variable selection.

2.2 Consistency and Oracle Properties

In this section, we will study the consistency and oracle properties of PMMLEs with non-concave and ALASSO penalties. Some regular conditions should be needed.

(Al) Suppose that $\boldsymbol{\beta} \in \Theta$, a compact subset of the Euclidean space R^p and the true value $\boldsymbol{\beta}_0$ is an interior of Θ. The covariate vector Z is exogenous and assumed to be

bounded, equivalently, there exists a constant $M_1 > 0$ such that $P(||Z|| \leq M_1) = 1$, where and hereafter $||\cdot||$ denotes the Euclidian norm. And for all $w \in \Theta, u \in (0,1)$ and v satisfying $||v|| \leq M_1$,

$$\phi_3(u,v,w) = \frac{\partial \phi(u,v,w)}{\partial w} \quad \phi_{33}(u,v,w) = \frac{\partial^2 \phi(u,v,w)}{\partial w \partial w^T}$$

exist and are continuous with respect to $w \in \Theta$.

(A2) For any v satisfying $||v|| \leq M_1$, there are functions $F_1(u,v)$ and $F_2(u,v)$, integrable with respect to u over $(0,1)$ such that

$$||\phi_3(u,v,w)|| < F_1(u,v), \quad ||\phi_{33}(u,v,w)|| < F_2(u,v) \text{ for all } w \in \Theta.$$

(A3) Denote $\psi(u,v,w) = \frac{\phi_3(u,v,w)}{\phi(u,v,w)}$ and $U = 1 - S_0(T)$. For any $\boldsymbol{\beta} \in \Theta$,

$$\begin{aligned}
\mu(\boldsymbol{\beta}) &= \mathrm{E}_{\boldsymbol{\beta}_0}\left\{\psi(1-U,Z,\boldsymbol{\beta})\right\}, \\
A(\boldsymbol{\beta}) &= E_{\boldsymbol{\beta}_0}\left\{\psi^2(1-U,Z,\boldsymbol{\beta})\right\}, \\
B(\boldsymbol{\beta}) &= E_{\boldsymbol{\beta}_0}\left\{-\frac{\partial}{\partial \boldsymbol{\beta}^T}\psi(1-U,Z,\boldsymbol{\beta})\right\}
\end{aligned}$$

exist and it is assumed that $A(\boldsymbol{\beta}_0)$ is positive definite. Here and below $\mathbf{a}^2$ for a column vector $\mathbf{a}$ means $\mathbf{a}\mathbf{a}^T$.

Obviously under conditions (A1)-(A3), it can be easily shown that $A(\boldsymbol{\beta}_0) = B(\boldsymbol{\beta}_0)$.

(A4) For any $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, where $\mathcal{O}_{\boldsymbol{\beta}_0}$ is a neighborhood of $\boldsymbol{\beta}_0$ in Θ, there exists a positive real number M_2 such that $||B(\boldsymbol{\beta})|| < M_2$.

(A5) The function $\psi(u,v,w)$ is continuous with respect to $u \in (0,1)$ and satisfies Lipschitz condition with respect to u in any closed subset of $(0,1)$. That is, for any set $[L,R] \subset (0,1)$, there exists a finite number $M_3(L,R)$, which is related with L and R, such that

$$||\psi(u_1,v,w) - \psi(u_2,v,w)|| \leq M_3|u_1 - u_2|$$

holds for all $u_1, u_2 \in [L,R]$, $||v|| \leq M_1$, and $w \in \mathcal{O}_{\boldsymbol{\beta}_0}$.

(A6) The function $\frac{\partial}{\partial w^T}\psi(u,v,w)$ is continuous with respect to $u \in (0,1)$ and satisfies Lipschitz condition, that is, there exists a positive real number M_4, for all $u_1, u_2 \in (0,1)$, $||v|| \leq M_1$, and $w \in \mathcal{O}_{\boldsymbol{\beta}_0}$ such that

$$\left|\left|\frac{\partial \psi(u_1,v,w)}{\partial w^T} - \frac{\partial \psi(u_2,v,w)}{\partial w^T}\right|\right| \leq M_4|u_1 - u_2|.$$

(A7) For any fixed discretization (for the discretization technique and corresponding related notations, please see the Section 2.3 of Ni (2008) [53]), there exists N such that when $n > N$,

$$E_{\beta 0}\left[-\frac{\partial}{\partial \beta^T} S_n(\beta)\right] \geq E_{\beta 0}\left[-\frac{\partial}{\partial \beta^T} S_{n,m}(\beta)\right], \quad \text{for all} \quad \beta \in \mathcal{O}_{\beta_0}.$$

where $S_n(\beta)$ and $-\frac{\partial}{\partial \beta^T} S_n(\beta)$ are the score function and Fisher information matrix with respect to the marginal likelihood function (2.1). $S_{n,m}(\beta)$ and $-\frac{\partial}{\partial \beta^T} S_{n,m}(\beta)$ are the discretized versions of $S_n(\beta)$ and $-\frac{\partial}{\partial \beta^T} S_n(\beta)$, which can be found in the Section 2.3 of Ni (2008).

Denote $U_i = 1 - S_0(T_i)$ and $\mathbf{u}_n = (u_1, u_2, \cdots, u_n)^T$, then $S_n(\beta)$ and $\frac{\partial}{\partial \beta} S_n(\beta)$ can be given by

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^{n} \int_{\mathcal{D}_n} \psi(1-u_i, Z_i, \beta) p(\mathbf{u}_n; \beta | \mathcal{R}_n, \mathbf{Z}_n) d\mathbf{u}_n \\ &= \sum_{i=1}^{n} \mathrm{E}_\beta \left[\psi(1-U_i, Z_i, \beta) | \mathcal{R}_n, \mathbf{Z}_n\right] \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} &\frac{\partial S_n(\beta)}{\partial \beta^T} \\ &= \sum_{i=1}^{n} \mathrm{E}_\beta \left[\frac{\partial}{\partial \beta^T} \psi(1-U_i, Z_i, \beta) \Big| \mathscr{F}_n\right] \\ &\quad + \mathrm{E}_\beta \left\{ \left[\sum_{i=1}^{n} \psi(1-U_i, Z_i, \beta)\right]^T \left[\sum_{i=1}^{n} \psi(1-U_i, Z_i, \beta)\right] \Big| \mathscr{F}_n \right\} - S_n^2(\beta) \\ &= \sum_{i=1}^{n} \mathrm{E}_\beta \left[\frac{\partial}{\partial \beta^T} \psi(1-U_i, Z_i, \beta) \Big| \mathscr{F}_n\right] + \mathrm{Var}_\beta \left[\sum_{i=1}^{n} \psi(1-U_i, Z_i, \beta) \Big| \mathscr{F}_n\right]. \end{aligned} \tag{2.4}$$

where

$$p(\mathbf{u}_n; \beta | \mathcal{R}_n, \mathbf{Z}_n) = \frac{I(\mathbf{u}_n \in \mathcal{D}_n)}{L_n(\beta | \mathcal{R}_n, \mathbf{Z}_n)} \prod_{i=1}^{n} \phi(1-u_i, Z_i, \beta) \tag{2.5}$$

is the conditional density function of $\mathbf{U} = (U_1, U_2, \cdots, U_n)^T$ given the ranking $\mathcal{R}_n$ and covariates $\mathbf{Z}_n$. Therefore the expectations and variance in (2.3) and (2.4) are respective to the density function (2.5). The subscript β here and below means that the expectations and variance in (2.3) and (2.4) are under the regression parameter β to distinguish from the true value, β_0, of β.

(A1) is a regular condition for models (1.3) whereas (A2) essentially allows the interchangeability of order for differentiation and integration or sum. (A3)-(A7) are sufficient conditions for asymptotic normality of $S_n(\boldsymbol{\beta}_0)$ and the uniform convergence of $\frac{\partial}{\partial \boldsymbol{\beta}^T} S_n(\boldsymbol{\beta})$ in $\mathcal{O}_{\boldsymbol{\beta}_0}$, which will be used to prove the oracle properties of PMMLEs. The condition (A5) is weaker than the Lipschitz condition and it is used to show that $S_n(\boldsymbol{\beta})$ can be approximated by a discretized score function $S_{n,m}(\boldsymbol{\beta})$. The inequality in condition (A7) should hold for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ otherwise, use of the discretized data will be better. Moreover, when $\frac{\partial}{\partial \boldsymbol{\beta}^T} S_n(\boldsymbol{\beta})$ enjoys some continuity in $\mathcal{O}_{\boldsymbol{\beta}_0}$, the inequality can be satisfied.

With proper penalty functions and above conditions, we can show that the PMMLEs are $\sqrt{n}$-consistent and perform as well as the oracle estimates. For the easy presentation of variable selection procedure, without loss of generality, we partition $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \cdots, \beta_{p0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ such that $\boldsymbol{\beta}_{10}$ contains all the nonzero components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{20} = \mathbf{0}$ contains all the zero effects. We also assume that the length of $\boldsymbol{\beta}_{10}$ is s. In addition, we declare that $\boldsymbol{\beta}_1$ consists of first s components of $\boldsymbol{\beta}$ while $\boldsymbol{\beta}_2$ includes the remaining ones of $\boldsymbol{\beta}$ such that we also have the corresponding partition for $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$.

2.2.1 Results of Variable Selection Methods with Non-concave Penalties

To avoid confusion in notations, here we declare that $\hat{\boldsymbol{\beta}}_n$ in this section is PMMLE of $\boldsymbol{\beta}$ with non-concave penalties while in the next section it becomes PMMLE of $\boldsymbol{\beta}$ with ALASSO penalty. Therefore, in this section, $p_\lambda(\cdot)$ can be any one of functions (1.8), (1.7) and (1.5) and in the next section $p_\lambda(\cdot)$ is the function (1.6).

Denote

$$a_n = \max_{1 \le j \le s} \{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\} \tag{2.6}$$

$$b_n = \max_{1 \le j \le s} \{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \tag{2.7}$$

Theorem 2.1 (Consistency) *Under the conditions (A1)-(A7), if $b_n \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ of $Q(\boldsymbol{\beta})$such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$.*

From Theorem 2.1, we can see that the PMMLE $\hat{\boldsymbol{\beta}}_n$ is $\sqrt{n}$-consistent when $a_n = O(n^{-1/2})$.

Denote

$$\mathbf{b}_{\lambda_n} = (p'_{\lambda_n}(|\beta_{10}|)\text{sign}(\beta_{10}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)\text{sign}(\beta_{s0}))^T$$

and

$$\Sigma_{\lambda_n} = \text{diag}(p''_{\lambda_n}(|\beta_{10}|), p''_{\lambda_n}(|\beta_{20}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)),$$

where λ_n means that the value of tuning parameter in penalty function varies as the sample size n. The following theorem presents the oracle properties of PMMLE with nonconcave penalties.

Theorem 2.2 (Oracle) *Under the conditions of Theorem 2.1 and*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0,$$

if $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *and* $a_n = O(n^{-1/2})$, *then with probability tending to 1, the* $\sqrt{n}$*-consistent local maximizer* $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \hat{\boldsymbol{\beta}}_{2n}^T)^T$ *in Theorem 2.1 must satisfy*

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$;

(ii) (Asymptotic Normality)

$$\sqrt{n}(B_1 + \Sigma_{\lambda_n})(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10} + (B_1 + \Sigma_{\lambda_n})^{-1}\mathbf{b}_{\lambda_n}) \xrightarrow{D} N(\mathbf{0}, B_1) \tag{2.8}$$

where B_1 *the upper leading* $s \times s$ *submatrix of* $B(\boldsymbol{\beta}_0)$.

Remark 2.1: Obviously for HARD thresholding and SCAD penalty functions, if $\lambda_n \to 0$, then for sufficiently large n, $a_n = 0$, $\boldsymbol{b}_{\lambda_n} = \mathbf{0}$ and $\Sigma_{\lambda_n} = \mathbf{0}$. Therefore, if $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \xrightarrow{D} \text{N}(\mathbf{0}, B_1^{-1}) \text{ and } \hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}.$$

This implies that with proper tuning parameter λ, PMMLEs with HARD and SCAD penalties enjoy oracle properties. It is as like that we have known $\boldsymbol{\beta}_2 = 0$ in advance when we estimate $\boldsymbol{\beta}_1$. On the other hand, for LASSO penalty function, $a_n = \lambda_n$, then the conditions $\sqrt{n}\lambda_n \to \infty$ and $a_n = O(n^{-1/2})$ in Theorem 2.2 contradict. So PMMLEs with LASSO penalty can not enjoy oracle properties.

2.2.2 Results of Variable Selection Method with Adaptive - LASSO Penalty

In this section, we present the consistency and Oracle properties of PMMLEs with ALASSO penalty (1.6).

Theorem 2.3 (Consistency) *Under the conditions (A1)-(A7), if* $\sqrt{n}\lambda_n = O(1)$, *then there exists a local maximizer* $\hat{\boldsymbol{\beta}}_n$ *of* $Q(\boldsymbol{\beta})$ *with ALASSO penalty satisfies* $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$.

This theorem shows that, with proper tuning parameter λ_n, the PMMLEs with ALASSO penalty enjoy $\sqrt{n}$-consistency.

Theorem 2.4 (Oracle) *If* $\sqrt{n}\lambda_n \to 0$ *and* $n^{(1+\gamma)/2}\lambda_n \to \infty$ *for some* $\gamma > 0$, *then under conditions of Theorem 2.3, the maximizer* $\hat{\boldsymbol{\beta}}_n$ *in Theorem 2.3 must satisfy:*

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$;

(ii) (Asymptotic Normality)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(\mathbf{0}, B_1^{-1}) \quad as \quad n \to \infty, \tag{2.9}$$

where B_1 *is defined in Theorem 2.2.*

From the Theorem 2.4, we can see that, with proper tuning parameter λ_n, the PMMLEs with ALASSO penalty also enjoy oracle properties as if we have known which effects are equal to 0 in advance when we estimate $\boldsymbol{\beta}_1$ in models (1.3).

2.2.3 Proof of Theorems

In this section, we will prove Theorems in Section 2.2.1 and 2.2.2 under regular conditions (A1)-(A7). Without loss of generality, we suppose that the continuous baseline cumulative density function $F_0(t) = 1 - S_0(t)$ satisfies $F_0(t) = 0$ for $t < 0$ and $F_0(t)$ is strictly increasing in $(0, \infty)$. Before proving the Theorems 2.1-2.4, we firstly introduce some Lemmas.

Lemma 2.1 Under the conditions (A1)-(A6), we have

$$\frac{1}{\sqrt{n}} S_n(\boldsymbol{\beta}_0) \xrightarrow{D} N(0, A(\boldsymbol{\beta}_0)) \quad \text{as} \quad n \longrightarrow \infty \tag{2.10}$$

where $A(\boldsymbol{\beta})$ is defined in condition (A3) and "$\xrightarrow{D}$" means convergence with distribution.

Lemma 2.2 Under the assumptions (A1)-(A7), it holds that

$$-\frac{1}{n}\frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \xrightarrow{P} B(\boldsymbol{\beta}) \quad \text{as} \quad n \to \infty \tag{2.11}$$

uniformly for all $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, where $B(\boldsymbol{\beta})$ is defined in condition (A3).

Lemma 2.3 Under conditions (A1) - (A7), there exists a sequence of $\tilde{\boldsymbol{\beta}}_n$ satisfying $S_n(\tilde{\boldsymbol{\beta}}_n) = 0$ such that as $n \longrightarrow 0$

(i) $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}_0$;

(ii) $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, A(\boldsymbol{\beta}_0)^{-1})$.

This Lemma shows that the maximum marginal likelihood estimate of $\boldsymbol{\beta}$ is consistent and distributed asymptotically normally.

Following the arguments for misspecified general transformation models with ranking data in Section 2.3 of Ni (2008) [53], one can prove Lemmas 2.1-2.3 without adding any difficulties. Therefore we omit the details here.

Proof of Theorem 2.1 Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any given $1 - \varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \sup_{||\boldsymbol{u}||=C} Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) < Q(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon. \tag{2.12}$$

Based on that $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\theta) > 0$ for $\theta > 0$, we have

$$\begin{aligned} &\frac{1}{n}\left[Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - Q(\boldsymbol{\beta}_0)\right] \\ \leq &\frac{1}{n}\left[\ell(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}_0)\right] - \sum_{j=1}^{s}[p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)], \end{aligned} \tag{2.13}$$

By Lemmas 2.1 and 2.2, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \quad ||\boldsymbol{u}|| = C\}$, we have

$$\begin{aligned} &\frac{1}{n}\left[\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0)\right] \\ =&\frac{1}{n}\left[\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T B(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\{1 + O_P(1)\}\right] \\ =&-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T[B(\boldsymbol{\beta}_0) + O_P(1)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_P(n^{-1/2}) \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \\ =&-\frac{1}{2}\alpha_n^2 \boldsymbol{u}^T[B(\boldsymbol{\beta}_0) + O_P(1)]\boldsymbol{u} + O_P(n^{-1/2}\alpha_n||\boldsymbol{u}||) \end{aligned} \tag{2.14}$$

Note that $B(\boldsymbol{\beta}_0)$ is a positive definite matrix. The order for first term in the last equality of (2.14) is $C^2\alpha_n^2$ and for second one is $\alpha_n^2 C$. Therefore, for a sufficiently large C, the second term is dominated by the first term in the last equation of (2.14). On the other hand, by Taylor's expansion, the second term of (2.13) is bounded by

$$\sqrt{s}\alpha_n a_n||\boldsymbol{u}|| + \alpha_n^2 b_n||\boldsymbol{u}||^2 = C\alpha_n^2(\sqrt{s} + b_n C).$$

If $b_n \to 0$, the second term of (2.13) is dominated by the first term of (2.14). Thus, for a sufficiently large C, (2.12) holds, which means that there exists a local maximizer for

$Q(\boldsymbol{\beta})$ in the ball $\{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ \ ||\boldsymbol{u}|| \leq C\}$ with probability at least $1 - \varepsilon > 0$. Therefore, there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$. □

Proof of Theorem 2.2 (i) It is sufficient to prove that

$$Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) = \max_{||\boldsymbol{\beta}_2|| \leq Cn^{-1/2}} \{Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T)\} \tag{2.15}$$

for any constant C and any given $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_P(n^{-1/2})$.

From Lemma 2.1 and $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$,

$$\begin{aligned} ||S_n(\boldsymbol{\beta})|| &= ||S_n(\boldsymbol{\beta}_0) + \frac{\partial S_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| + O_P(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||^2) \\ &= O_P(\sqrt{n}) \end{aligned} \tag{2.16}$$

So for $j = (s+1), (s+2), \cdots, p$

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= S_{nj}(\boldsymbol{\beta}) - np'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) \\ &= -np'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) + O_P(\sqrt{n}) \\ &= n\lambda_n\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) + O_P(\frac{1}{\sqrt{n}\lambda_n})\}. \end{aligned} \tag{2.17}$$

where $S_{nj}(\boldsymbol{\beta})$ is jth element of $S_n(\boldsymbol{\beta})$. Since $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$, and $\frac{1}{\sqrt{n}\lambda_n} \to 0$, the derivative $\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}$ and $-\beta_j$ have the same sign. Therefore (2.15) holds.

(ii) From $a_n = O(n^{-1/2})$ and Theorem 2.1, there exists a local $\sqrt{n}$-consistent maximizer, $\hat{\boldsymbol{\beta}}_{1n}$, of $Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T)$ satisfying

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T} = 0 \ \text{ for } \ j = 1, 2, \cdots, s. \tag{2.18}$$

Set $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T$ and $S_{1n}(\boldsymbol{\beta})$ as the upper leading $s \times s$ submatrix of $S_n(\boldsymbol{\beta})$, then

$$\begin{aligned} 0 = \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n} &= \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \\ &= S_{1n}(\boldsymbol{\beta}_0) - n\boldsymbol{b}_{\lambda_n} + \frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) - n\Sigma_{\lambda_n}(\boldsymbol{\beta}_1^*)(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \end{aligned} \tag{2.19}$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{T*}, \boldsymbol{\beta}_2^{T*})^T$ lies on the line segment between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$; $\Sigma_{\lambda_n}(\boldsymbol{\beta}_1) = \text{diag}(p''_{\lambda_n}(|\beta_1|), p''_{\lambda_n}(|\beta_2|), \cdots, p''_{\lambda_n}(|\beta_s|))$. From Theorem 2.1 and Lemmas 2.1-2.2, (2.8) holds. This completes the proof. □

Proof of Theorem 2.3 Similar to the proof of Theorem 2.1, let $\alpha_n = n^{-1/2}$. It is sufficient to show that for any given $1-\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{||\boldsymbol{u}||=C} Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) < Q(\boldsymbol{\beta}_0)\right\} \geq 1-\varepsilon. \tag{2.20}$$

Note that

$$\begin{aligned}
&\frac{1}{n}\left[Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - Q(\boldsymbol{\beta}_0)\right] \\
\leq &\frac{1}{n}\left[\ell(\boldsymbol{\beta} + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}_0)\right] - \lambda_n \sum_{j=1}^{s}[(|\beta_{j0} + \alpha_n u_j| - |\beta_{j0}|)/|\tilde{\beta}_j|^{\gamma}] \\
\leq &\frac{1}{n}\left[\ell(\boldsymbol{\beta} + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}_0)\right] + \lambda_n \alpha_n \sum_{j=1}^{s}[|u_j|/|\tilde{\beta}_j|^{\gamma}]
\end{aligned} \tag{2.21}$$

By Lemmas 2.1-2.2, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ \ ||\boldsymbol{u}|| = C\}$, we have

$$\begin{aligned}
&\frac{1}{n}\left[\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0)\right] \\
= &\frac{1}{n}\left[S_n(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \frac{\partial S_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\{1 + O_P(1)\}\right] \\
= &-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T[B(\boldsymbol{\beta}_0) + O_P(1)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_P(n^{-1/2}) \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \\
= &-\frac{1}{2}\alpha_n^2 \boldsymbol{u}^T[B(\boldsymbol{\beta}_0) + O_P(1)]\boldsymbol{u} + O_P(n^{-1/2}\alpha_n ||\boldsymbol{u}||)
\end{aligned} \tag{2.22}$$

Note that $B(\boldsymbol{\beta}_0)$ is a positive definite matrix. The order for first term in the last equality of (2.22) is $C^2\alpha_n^2$ and for second one is $\alpha_n^2 C$. Therefore, for a sufficiently large C, the second term is dominated by the first term in the last equality. On the other hand, by Taylor's expansion and Lemma 2.3,

$$\frac{1}{|\tilde{\beta}_j|^{\gamma}} = \frac{1}{|\beta_{j0}|^{\gamma}} - \frac{\gamma \text{sign}(\beta_{j0})}{|\beta_{j0}|^{\gamma+1}}(\tilde{\beta}_j - \beta_{j0}) + o_P(\tilde{\beta}_j - \beta_{j0}) = \frac{1}{|\beta_{j0}|^{\gamma}} + \frac{O_P(1)}{\sqrt{n}}$$

and hence the second term of (2.21) is bounded by $C\alpha_n^2$, since

$$\begin{aligned}
|\alpha_n \lambda_n| \sum_{j=1}^{s}[|u_j|/|\tilde{\beta}_j|^{\gamma}] &= \frac{1}{\sqrt{n}}\lambda_n \sum_{j=1}^{s}\left[\frac{|u_j|}{|\beta_{j0}|^{\gamma}} + \frac{|u_j|}{\sqrt{n}}O_P(1)\right] \\
\leq &Cn^{-1/2}\lambda_n O_P(1) = Cn^{-1}(\sqrt{n}\lambda_n)O_P(1) \leq C\alpha_n^2 O_P(1)
\end{aligned}$$

and

$$\sqrt{n}\lambda = O(1).$$

Therefore, the second term of (2.21) is also dominated by the first term of (2.22). Thus, for a sufficiently large C, (2.20) holds, which means that there exists a local maximizer

in the ball $\{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ ||\boldsymbol{u}|| \leq C\}$ with probability at least $1 - \varepsilon > 0$. Therefore, there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$. □

Proof of Theorem 2.4 (i)Similar to the proof of Theorem 2.2, it is sufficient to prove that

$$Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) = \max_{||\boldsymbol{\beta}_2|| \leq Cn^{-1/2}} \{Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T)\} \tag{2.23}$$

for any given $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_P(n^{-1/2})$ and any constant C.

Since $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$,

$$\begin{aligned} ||S_n(\boldsymbol{\beta})|| &= ||S_n(\boldsymbol{\beta}_0) + \frac{\partial S_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| + O_P(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||^2) \\ &= O_P(\sqrt{n}) \end{aligned}$$

From Lemma 2.3, for $j = s+1, s+2, \cdots, p$, we have

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= S_{nj}(\boldsymbol{\beta}) - n\lambda_n \frac{\text{sign}(\beta_j)}{|\tilde{\beta}_j|^\gamma} \\ &= -(n\lambda_n) n^{\gamma/2} \frac{\text{sign}(\beta_j)}{|\sqrt{n}\tilde{\beta}_j|^\gamma} + O_P(\sqrt{n}) \\ &= \sqrt{n}\left[O_P(1) - (n^{(\gamma+1)/2}\lambda_n)\frac{\text{sign}(\beta_j)}{|O_P(1)|}\right] \end{aligned}$$

where $S_{nj}(\boldsymbol{\beta})$ is jth element of $S_n(\boldsymbol{\beta})$. Since $n^{(\gamma+1)/2}\lambda_n \to \infty$ as $n \to \infty$, the derivative $\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}$ and $-\beta_j$ have the same sign. Therefore (2.23) holds.

(ii) From Theorem 2.3, there exists a local $\sqrt{n}$-consistent maximizer, $\hat{\boldsymbol{\beta}}_{1n}$, of $Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T)$ satisfying

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T} = 0 \quad \text{for} \quad j = 1, 2, \cdots, s. \tag{2.24}$$

Set $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T$ and denote $S_{1n}(\boldsymbol{\beta})$ as a vector consisting of the first s components of $S_n(\boldsymbol{\beta})$, then

$$\begin{aligned} 0 = \frac{\partial Q(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_1} &= \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \\ &= S_{1n}(\boldsymbol{\beta}_0) - n\lambda_n \left(\frac{\text{sign}(\beta_{10})}{|\tilde{\beta}_1|^\gamma}, \frac{\text{sign}(\beta_{20})}{|\tilde{\beta}_2|^\gamma}, \cdots, \frac{\text{sign}(\beta_{s0})}{|\tilde{\beta}_s|^\gamma}\right)^T \\ &+ \frac{\partial S_n(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) - n\lambda_n \Sigma_{\lambda_n}(\boldsymbol{\beta}_1^*)(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \end{aligned} \tag{2.25}$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{T*}, \boldsymbol{\beta}_2^{T*})$ lies on the line segment between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. From $\sqrt{n}\lambda_n \to 0$, $\tilde{\boldsymbol{\beta}}_1 \xrightarrow{P} \boldsymbol{\beta}_{10}$, Theorem 2.1 and Lemmas 2.2-2.3, (2.9) holds. This completes the proof. □

2.3 Implementation

In this section, we propose an uniform algorithm for our proposed variable selection procedures by incorporating the MCMC-SA algorithm in Gu, *et al.* (2005) [40] and Gu and Kong (1998) [39].

Through quadratic Taylor's expansion for (2.2), maximization of (2.2) can be reduced to a local quadratic maximization problem, which leads to a modified Newton-Raphson algorithm. Note that there are high-dimensional integrations included in the $S_n(\boldsymbol{\beta})$ and $\frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$, first-order and second-order derivatives of $\ell(\boldsymbol{\beta})$. Moreover, the marginal likelihood function $L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$ is also a high-dimensional integration and has no closed expression. Thus we should use some approximation methods to estimate $S_n(\boldsymbol{\beta})$ and $\frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$. Because of Monte Carlo error, the Newton-Raphson algorithm may be very hard to converge. To overcome the difficulties, we will combine the three-stage MCMC-SA algorithms in Gu, *et al.* (2005) [40] and Gu and Kong (1998) [39] to maximize the penalized log-marginal likelihood function (2.2) with respect to $\boldsymbol{\beta}$. The basic idea is as follows: The first two stages of three-stage MCMC-SA algorithm of Gu, *et al* (2005) [40] are used to generate the initial estimate of $\boldsymbol{\beta}$; Then we update the estimate of $\boldsymbol{\beta}$ by developing MCMC-SA algorithm in Gu and Kong (1998) [39]. In the updating procedure, we not only can select significant variables but also can estimate corresponding effects; Finally, the third stage of three-stage MCMC-SA algorithm is used to calculate the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{1n}$. Three-stage MCMC stochastic approximation algorithm and corresponding notations are listed in Appendix A.

Another difficulty is that the penalty function in (2.2) is irregular at origin and may not be secondly differentiable at some points. Following Fan and Li (2001, 2002) [25, 28], we will approximate the penalty functions by local quadratic approximation method. Given an initial value $\boldsymbol{\beta}_0$ that is close to the maximizer of penalized log-marginal likelihood function (2.2), when β_{j0} is not very close to 0, the penalty $[p_\lambda(|\beta_j|)]'$ is locally approximated by

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sign}(\beta_j) \approx p'_\lambda(|\beta_{j0}|)\frac{\beta_j}{|\beta_{j0}|}, \tag{2.26}$$

otherwise, set $\hat{\beta}_j = 0$. As a consequence, for $\beta_j \approx \beta_{j0}$,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}[p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|](\beta_j^2 - \beta_{j0}^2)$$

and

$$[p_\lambda(|\beta_j|)]'' \approx p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|.$$

Next we give the standard deviation formula of the rank-based penalized maximum marginal likelihood estimate. Denote $\nabla\ell(\boldsymbol{\beta}) = -\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$ and $\nabla^2\ell(\boldsymbol{\beta}) = -\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$. Then, at the $(k+1)$th step in the Newton-Raphson algorithm, the penalized marginal likelihood estimate of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^{(k+1)}$, can be updated through

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - \left[\nabla^2\ell(\hat{\boldsymbol{\beta}}^{(k)}) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}^{(k)})\right]^{-1}\left[\nabla\ell(\hat{\boldsymbol{\beta}}^{(k)}) + n\boldsymbol{b}_\lambda(\hat{\boldsymbol{\beta}}^{(k)})\right] \tag{2.27}$$

where

$$\Sigma_\lambda(\boldsymbol{\beta}) = \mathrm{diag}\{p'_\lambda(|\beta_1|)/|\beta_1|, p'_\lambda(|\beta_2|)/|\beta_2|, \cdots, p'_\lambda(|\beta_p|)/|\beta_p|\}$$

and

$$\mathbf{b}_\lambda(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta})\boldsymbol{\beta}.$$

Thus at convergence, $Var(\hat{\boldsymbol{\beta}}_n)$ can be estimated by following sandwich matrix,

$$\left[\nabla^2\ell(\hat{\boldsymbol{\beta}}_n) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_n)\right]^{-1}\widehat{\mathrm{Cov}}(\nabla\ell(\hat{\boldsymbol{\beta}}_n) + n\boldsymbol{b}_\lambda(\hat{\boldsymbol{\beta}}_n))\left[\nabla^2\ell(\hat{\boldsymbol{\beta}}_n) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_n)\right]^{-1} \tag{2.28}$$

This formula is consistent with Theorems 2.2 and 2.4, which performs very well for moderate sample size.

With a proper tuning parameter λ, the rank-based penalized maximum marginal likelihood estimates can be found by the following three-step MCMC-SA algorithm.

Step 1 With Gibbs sampling procedure for ranking data, run the first two stages of the three-stage MCMC-SA algorithm and generate the off-line average estimates of $\boldsymbol{\beta}$ and Γ, denoted by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\Gamma}$, where Γ is the Fisher information matrix with respect to (2.1). Denote $\tilde{\mathbf{U}}$ as the last sample of $\mathbf{U}$ from Gibbs sampling procedure in this step.

Step 2 Set $\mathbf{U}_{0m} = \tilde{\mathbf{U}}$. Following the first step in the first stage of three-stage MCMC-SA algorithm, for fixed k, generate m samples from (2.5) - $\mathbf{U}_k = (\mathbf{U}_{k,1}, \mathbf{U}_{k,2}, \cdots, \mathbf{U}_{k,m})$ with $\mathbf{U}_{k,i} = (U_{k,i,1}, U_{k,i,2}, \cdots, U_{k,i,n})^T$. Take $\tilde{\boldsymbol{\beta}}$ and $\tilde{\Gamma}$ as the initial values of $\boldsymbol{\beta}$ and Γ in this step. With proper tuning parameter λ and the current value $\hat{\boldsymbol{\beta}}^{(k)}$, update $\hat{\boldsymbol{\beta}}^{(k+1)}$ by running

$$\Gamma_{k+1} = \Gamma_k + \gamma_{3k}(\bar{I}(\hat{\boldsymbol{\beta}}^{(k)}, \mathbf{U}_k) + \Sigma_\lambda(\hat{\boldsymbol{\beta}}^{(k)}) - \Gamma_k)$$

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \gamma_{3k}\Gamma_{k+1}^{-1}\left[\bar{H}(\hat{\boldsymbol{\beta}}^{(k)}; \mathbf{U}_k) - \mathbf{b}_\lambda(\hat{\boldsymbol{\beta}}^{(k)})\right]$$

where

$$\bar{I}(\boldsymbol{\beta}, \mathbf{U}_k) = -\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\boldsymbol{\beta}^T}H(\boldsymbol{\beta}; \mathbf{U}_{k,i}),$$

$$\bar{H}(\boldsymbol{\beta}; \mathbf{U}_k) = \frac{1}{m}\sum_{i=1}^{m}H(\boldsymbol{\beta}, \mathbf{U}_{k,i}),$$

$$H(\boldsymbol{\beta}; \mathbf{u}) = \sum_{i=1}^{n}\psi(1-u_i, Z_i, \boldsymbol{\beta}),$$

and $\gamma_{3k} = \frac{10}{k^c+10}$ with $c \in (0.5, 1)$. The updating step will terminate when $||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||$ is less than a pre-specified small positive real number, for example, 10^{-4}. With proper penalty function, this step not only can select important variables but also can estimate corresponding effects.

Step 3 The variance of the nonzero penalized estimates can be obtained by calculating $\nabla^2\ell(\hat{\boldsymbol{\beta}}_n)$ with the third stage of three-stage MCMC-SA algorithm.

Gibbs Sampling for Ranking Data:

Without loss of generality, we assume the components of $\mathbf{U}_n = (U_1, U_2, \cdots, U_n)^T$ has been sorted in advance in ascending order with their corresponding covariates $\mathbf{Z}_n$ ordered. Then the ranking vector is $\mathcal{R}_n = (1, 2, \cdots, n)^T$. Here $U_i = 1 - S_0(T_i)$. Then U_i has the survival function $\Phi(1-u, Z_i, \boldsymbol{\beta})$. Define

$$\mathcal{E} = \{(U_1, U_2, \cdots, U_n): \quad U_1 < U_2 < \cdots < U_n\}.$$

Then the density distribution of $\mathbf{U}_n$ condition on $\mathbf{U}_n \in \mathcal{E}$ is (2.5). Moreover, for fixed U_j's ($j \neq i$), U_i has the distribution function $1 - \Phi(1-u, Z_i, \boldsymbol{\beta})$ and it is restricted in (U_{i-1}, U_{i+1}) with $U_0 = 0$ and $U_{n+1} = 1$.

Given $\mathbf{Z}_n$, we can generate a sample of $\mathbf{U}_n$ from the stationary probability (2.5) through the following Gibbs sampling procedure. Let $\mathbf{U}_n^k = (U_{1,k}, U_{2,k}, \cdots, U_{n,k})^T$ be the current sample of $\mathbf{U}_n$. Then the next sample $\mathbf{U}_n^{k+1}$ can be generated as follows,

0. Set $j = 1$;

1. Set $u_j^- = 1 - \Phi(1 - U_{j-1,k}, Z_j, \boldsymbol{\beta})$ and $u_j^+ = 1 - \Phi(1 - U_{j+1,k}, Z_j, \boldsymbol{\beta})$ with $u_0^- = 0$ and $u_{n+1}^+ = 1$;

2. Generate u^* from Unif$[u_j^-, u_j^+]$ and set $U_{j,k+1} = 1 - \Phi^{-1}(1 - u^*, Z_j, \beta)$, where $\Phi^{-1}(u, v, w)$ is the inverse function of $\Phi(u, v, w)$ with respective to u.

3. If $j < n$, then $j = j + 1$ and go to step 1, otherwise stop.

Remark 2.3: Following Craven and Wahba (1979) [19] and Fan and Li, (2002) [25], the proper tuning parameter λ can be selected by an approximated generalized cross-validation (GCV) statistic. Following second order Taylor's expansion,

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}_0) - [\nabla\ell(\boldsymbol{\beta}_0)]^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\nabla^2\ell(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Let $\nabla^2\ell(\boldsymbol{\beta}_0) = X^TX$ as the Cholesky decomposition, then $\ell(\boldsymbol{\beta})$ can be approximated by

$$\ell(\boldsymbol{\beta}_0) - \frac{1}{2}\left[X(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + X^{-T}\nabla\ell(\boldsymbol{\beta}_0)\right]^T\left[X(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + X^{-T}\nabla\ell(\boldsymbol{\beta}_0)\right] + o(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||^2).$$

Then when $\boldsymbol{\beta}$ is very close to $\boldsymbol{\beta}_0$, the maximization of the penalized marginal likelihood function (2.2) can be reduced to the local quadratic minimization of

$$\frac{1}{2}[Y - X\boldsymbol{\beta}]^T[Y - X\boldsymbol{\beta}] + n\sum_{i=1}^{p} p_\lambda(|\beta_i|) \tag{2.29}$$

where $Y = X^{-T}[\nabla^2\ell(\boldsymbol{\beta})\boldsymbol{\beta} + \nabla\ell(\boldsymbol{\beta})]$. Then at convergence, $\boldsymbol{\beta}$ can be estimated by the ridge estimate $\hat{\boldsymbol{\beta}}_n = [\nabla^2\ell(\hat{\boldsymbol{\beta}}_n) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_n)]^{-1}X^TY$. So the degree of freedom for selected model can be approximated by

$$e(\lambda) = \text{tr}\left\{\left[\nabla^2\ell(\hat{\boldsymbol{\beta}}_n) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_n)\right]^{-1}\nabla^2\ell(\hat{\boldsymbol{\beta}}_n)\right\}. \tag{2.30}$$

The approximated GCV criterion can be defined by

$$\text{GCV}(\lambda) = -\frac{\ell(\hat{\boldsymbol{\beta}}_n)}{n[1 - e(\lambda)/n]^2} \tag{2.31}$$

where $\ell(\boldsymbol{\beta})$ can be approximated by the following important sampling procedure.

Intuitively, the GCV inflates the negative log-marginal likelihood by the factor involving the degree of freedom $e(\lambda)$. Large value $e(\lambda)$ will cause more inflation of the negative log-marginal likelihood.

Approximation of $\ell(\boldsymbol{\beta})$ with ranking data:

To approximate $\ell(\boldsymbol{\beta})$, we firstly approximate $L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$. Note that when $\mathbf{Z}_n = \mathbf{0}$,

$$p_0(\mathbf{u}_n; \mathcal{R}_n) = n!I(\mathbf{u}_n \in \mathcal{D}_n) \tag{2.32}$$

is the baseline conditional density function of $\mathbf{U}_n$ given $\mathcal{R}_n$. Multiply and divide the integrand in (2.1) by the density (2.32), then the marginal likelihood (2.1) can be rewritten as a conditional expectation with respect to the baseline density (2.32), namely,

$$L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n) = \frac{1}{n!}\mathrm{E}_{p_0}\left[\prod_{i=1}^{n}\phi(1-U_i, Z_i, \boldsymbol{\beta})\middle|\mathcal{R}_n, \mathbf{Z}_n\right]. \tag{2.33}$$

Therefore $L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$ can be approximated by important sampling. Specifically, we first generate M_0 simulated samples $\tilde{U}_i = (U_{i,1}, U_{i,2}, \cdots U_{i,n})^T$ of $\mathbf{U}$ from (2.32) by virtue of Gibbs sampling procedure with $\mathbf{Z}_n = \mathbf{0}$ and then $L_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$ can be estimated by

$$\hat{L}_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n) = \frac{1}{n!}\frac{1}{M_0}\sum_{i=1}^{M_0}\left[\prod_{j=1}^{n}\phi(1-U_{i,j}, Z_j, \boldsymbol{\beta})\right].$$

The score function $S_n(\boldsymbol{\beta})$ and the Fisher information matrix $I_n(\boldsymbol{\beta}) = -\frac{\partial}{\partial\boldsymbol{\beta}^T}S_n(\boldsymbol{\beta})$ can be correspondingly approximated by

$$\hat{S}_n(\boldsymbol{\beta}) = \frac{1}{n!}\frac{1}{M_0}\sum_{i=1}^{M_0}\left\{H(\boldsymbol{\beta}; \tilde{U}_i)\left[\prod_{j=1}^{n}\phi(1-U_{i,j}, Z_j, \boldsymbol{\beta})\right]\right\}\Big/\hat{L}_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$$

and

$$\begin{aligned}&\hat{I}_n(\boldsymbol{\beta})\\ &= -\frac{1}{n!}\frac{1}{M_0}\sum_{i=1}^{M_0}\left\{\left[\frac{\partial}{\partial\boldsymbol{\beta}^T}H(\boldsymbol{\beta}; \tilde{U}_i) + H^2(\boldsymbol{\beta}; \tilde{U}_i)\right]\left[\prod_{j=1}^{n}\phi(1-U_{i,j}, Z_j, \boldsymbol{\beta})\right]\right\}\Big/\hat{L}_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)\\ &\quad + \hat{S}_n^2(\boldsymbol{\beta})\end{aligned} \tag{2.34}$$

Then we can use $\hat{L}_n(\boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$ and $\hat{I}_n(\boldsymbol{\beta})$ to approximate (2.31). Given covariate sample $\mathbf{Z}_n$, to assess the selected models efficiently, the same simulated samples of $\mathbf{U}_n$ will be used to approximate (2.31) regardless of the values of $\boldsymbol{\beta}$ and λ.

2.4 Extension to Stratified General Transformation Models

The general transformation models (1.3) and corresponding variable selection procedures can be easily extended into stratified case.

Denote $S_{0i}(t)$ as the baseline survival function in the ith stratum and then the stratified general transformation models can be defined by

$$S_{i|Z_i}(t) = \Phi(S_{0i}(t), Z_i, \boldsymbol{\beta}) \quad \text{for} \quad i = 1, 2, \cdots, N. \tag{2.35}$$

where the assumptions about $\Phi(u, v, w)$ and β are the same as the ones for (1.3). Moreover, due to the difference of external environments, the baseline survival function $S_{0i}(\beta)$ may vary across strata.

Suppose that $\{Z_{ij}\}_{j=1}^{n_i}$ are n_i i.i.d. copies of Z_i in ith stratum, $i = 1, 2, \cdots, N$. Denoted $T_{i,j}$ as the underlying failure time of jth individual in ith stratum, $\mathcal{R}_i$ as the observed ranking in the ith stratum. Then the marginal likelihood function $L_i(\beta|\mathbf{Z}_i, \mathcal{R}_i)$ in ith stratum has the same form as (2.1) for general transformation models (1.3). Consequently, the marginal likelihood function for (2.35) is given by

$$L(\beta|\mathbf{Z}, \mathcal{R}) = \prod_{i=1}^{N} L_i(\beta|\mathbf{Z}_i, \mathcal{R}_i),$$

where $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_N)$, $\mathbf{Z}_i = \{Z_{i1}, Z_{i2}, \cdots, Z_{in_i}\}$ and $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_N)$. Following the variable selection methods for non-stratified general transformation models (1.3), we can easily develop the variable selection procedures and corresponding oracle properties for the stratified general transformation models. The variable selection procedures for stratified general transformation models (2.35) can be also implemented by slightly modified algorithm discussed in Section 2.3. Here we omit it. The stratified general transformation models and corresponding algorithm will be used to analyze the horse racing data set in Section 2.5.2.

2.5 Numeric Studies

In this section, we illustrate our proposed variable selection procedures by three simulation examples and one real data application. Firstly we conduct some simulations for three special models of general transformation models with ranking data — proportional hazard regression model, proportional odds regression model and generalized probit transformation model. Then we apply proposed variable selection procedures to analyze *Hong Kong Horse Racing Data* through stratified generalized probit model.

According to Tibshirani (1997) [68], we use median of mean squared error (MMSE) $(\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0)$ over 100 runs to evaluate the efficiency of the proposed variable selection methods, where Σ is the population covariance matrix of regressors and it can be estimated by covariate sample.

2.5.1 Simulation Examples

In this example, 100 data sets consisting of $n = 100$ and 200 observations are simulated from the models

$$S_Z(t) = \Phi(S_0(t), Z, \beta) \tag{2.36}$$

where $\Phi(1-u, v, w) = h^{-1}(h(u) + v^T w)$ and $h^{-1}(u)$ takes standard extreme value survival function, standard logistic survival function and standard normal survival function, which correspond to proportional hazards models (PH), proportional odds models (PO) and generalized probit models (GP) respectively; $S_0(t) = e^{-t}$, $Z \in R^9$ and $\beta = (0.8, 0, -0.8, 0, 0, 0.8, 0, 0, -0.8)^T$. All the elements of Z follow the standard normal distribution independently.

We run the three-step MCMC-SA algorithm with the following parameter setting: For the first two stages of the three-stage MCMC-SA algorithm, we take $m = 100$, $K_0 = 100$, $c_1 = 0.3$ and $c_2 = 0.6$ while for the second step in our three-step MCMC-SA algorithm, we take $c_3 = 0.9$. The tuning parameters in the variable selection procedures are selected by approximated GCV (2.31). For the approximation of GCV (2.31), we will generate $M_0 = 20000$ samples from the conditional baseline density function (2.32).

MMSE is used to assess the efficiency of the proposed variable selection methods. MMSEs based on 100 simulations are listed in Table 2.1. The average numbers of selected zero coefficients are also reported in Table 2.1, in which the columns labeled as "correct" present the average number of zero effects detected correctly by our proposed procedures, while the columns labeled "incorrect" give the average number of coefficients erroneously set to be 0. Tables 2.2, 2.3 and 2.4 summarize the estimation of nonzero effects in PH, PO and GP models. The estimation results include the bias (Bias), sample standard deviations (SStd) and mean of estimated standard deviation (MStd). For PMMLE, MStd's are calculated from (2.28) while MStd's for oracle estimates are calculated based on the inverse of Fisher information matrix at oracle estimates. When one variable with true nonzero effect is excluded from the selected model, its estimate and corresponding estimated standard deviation are set to be 0.

Based on Table 2.1 and MMSE, it can be seen that the variable selection methods with SCAD, HARD thresholding and ALASSO penalties perform similarly and they also perform as well as the oracle estimates in the three models while the method with LASSO

Table 2.1: Variable Selection results for PH,PO,GP models with ranking data under $n = 100, 200$

		$n = 100$			$n = 200$		
			Aver. no. of 0 Coef.			Aver. no. of 0 Coef.	
Models	Penalty	MMSE	correct	incorrect	MMSE	correct	incorrect
PH	HARD	0.125	4.475	0.000*	0.087	4.841	0.000
	SCAD	0.077	4.670	0.003	0.028	5.000	0.000
	LASSO	0.675	4.680	0.004	0.638	4.922	0.000
	ALASSO	0.083	4.375	0.005	0.039	4.793	0.000
	Oracle	0.103	5.000	0.000	0.092	5.000	0.000
PO	HARD	0.254	4.931	0.070	0.065	4.863	0.010
	SCAD	0.192	4.888	0.060	0.065	4.932	0.011
	LASSO	0.939	4.670	0.080	0.820	4.822	0.034
	ALASSO	0.215	4.007	0.004	0.069	4.574	0.000
	Oracle	0.117	5.000	0.000	0.057	5.000	0.000
GP	HARD	0.067	4.678	0.000	0.027	4.728	0.000
	SCAD	0.062	4.630	0.000	0.024	4.860	0.000
	LASSO	0.461	4.525	0.000	0.441	4.940	0.000
	ALASSO	0.079	4.373	0.028	0.029	4.829	0.018
	Oracle	0.044	5.000	0.000	0.024	5.000	0.000

Note: 0.000*s indicate that the corresponding values are less than 0.0005.

penalty suffers from relatively large MMSE. Moreover, from the average number of zero coefficients, all the methods can select about the same correct number of significant variables for both $n = 100$ and $n = 200$. On the other hand, in all the settings, the values of MMSE decrease as the increasing of sample size. So the performance of variable selection methods will improve according to MMSE when sample size increases.

Table 2.2: Summary of estimation results for nonzero effects in PH model with ranking data under $n = 100, 200$

		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.1380	0.1164	-0.1433	0.1366	-0.1489	0.1423	-0.1433	0.1404
	SStd	0.0946	0.1105	0.1051	0.1037	0.0731	0.0730	0.0671	0.0855
	MStd	0.1175	0.1178	0.1158	0.1179	0.0779	0.0780	0.0788	0.0783
SCAD	Bias	-0.0528	0.0328	-0.0554	0.0338	-0.0293	0.0167	-0.0331	0.0373
	SStd	0.1325	0.1251	0.1262	0.1172	0.0759	0.0869	0.0856	0.0928
	MStd	0.1082	0.1075	0.1066	0.1090	0.0728	0.0733	0.0725	0.0716
LASSO	Bias	-0.4402	0.4429	-0.4527	0.4478	-0.4404	0.4342	-0.4306	0.4434
	SStd	0.0898	0.0803	0.1008	0.0992	0.0619	0.0520	0.0604	0.0574
	MStd	0.0878	0.0865	0.0847	0.0868	0.0611	0.0613	0.0602	0.0611
ALASSO	Bias	-0.1450	0.1526	-0.1471	0.1500	-0.1590	0.1411	-0.1523	0.1451
	SStd	0.1084	0.1191	0.1652	0.1187	0.0709	0.0743	0.0747	0.0775
	MStd	0.1169	0.1151	0.1152	0.1165	0.0778	0.0785	0.0778	0.0778
Oracle	Bias	-0.1293	0.1246	-0.1294	0.1484	-0.1388	0.1450	-0.1373	0.1427
	SStd	0.1079	0.1042	0.1030	0.0951	0.0661	0.0777	0.0677	0.0759
	MStd	0.1246	0.1294	0.1246	0.1220	0.0814	0.0821	0.0816	0.0815

Tables 2.2, 2.3 and 2.4 show that the biases of PMMLE with SCAD, HARD thresholding and ALASSO penalties are as small as that of oracle estimates and the Biases can be reasonably ignorable in the three model settings. However, the biases of PMMLE with LASSO penalty are relatively larger.

To test the accuracy of sandwich formula (2.28) in Section 2.3, we compare the sample standard deviations with the means of estimated standard deviation from (2.28). Note that SStd is the sample standard deviation. Without considering Monte Carlo error, it

Table 2.3: Summary of estimation results for nonzero effects in PO model with ranking data under $n = 100, 200$

		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.0379	0.0395	-0.0633	0.0884	-0.0152	0.0127	-0.0210	0.0209
	SStd	0.1626	0.2477	0.2245	0.2768	0.1463	0.1351	0.1661	0.1590
	MStd	0.1903	0.1804	0.1824	0.1851	0.1273	0.1276	0.1272	0.1276
SCAD	Bias	-0.0499	0.0516	-0.0664	0.0314	-0.0334	0.0013	-0.0208	0.0083
	SStd	0.2787	0.2701	0.2676	0.2390	0.1507	0.1536	0.1610	0.1329
	MStd	0.1674	0.1752	0.1710	0.1779	0.1200	0.1207	0.1200	0.1237
LASSO	Bias	-0.4980	0.4827	-0.5157	0.4878	-0.4527	0.4640	-0.4557	0.4759
	SStd	0.1702	0.1898	0.1788	0.1745	0.1395	0.1417	0.1324	0.1275
	MStd	0.1111	0.1104	0.1062	0.1123	0.0846	0.0835	0.0849	0.0835
ALASSO	Bias	-0.0359	0.0478	-0.0554	0.0562	-0.0107	0.0297	-0.0274	0.0076
	SStd	0.1879	0.1793	0.2310	0.1860	0.1395	0.1376	0.1422	0.1475
	MStd	0.1894	0.1904	0.1859	0.1885	0.1273	0.1268	0.1265	0.1263
Oracle	Bias	0.014	-0.0086	0.0347	-0.0084	0.0154	-0.0236	0.0116	-0.0003
	SStd	0.1881	0.1872	0.2167	0.1908	0.1449	0.1517	0.1315	0.1589
	MStd	0.2007	0.2020	0.2006	0.2030	0.1390	0.1398	0.1385	0.1384

can be seen as the true standard deviation. From Tables 2.2, 2.3 and 2.4, we can find that all the means of estimated standard deviation are reasonably close to their corresponding sample standard deviations in all the settings. Moreover, the values of MStd, SStd and their difference decrease as the increasing of sample size. This shows that when the sample size increases, the performance of sandwich formula (2.28) will improve significantly.

In a word, for the general transformation models with ranking data, the variable selections with SCAD, HARD thresholding and ALASSO penalties outperform the method

Table 2.4: Summary of estimation results for nonzero effects in GP model with ranking data under $n = 100, 200$

		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.0250	-0.0426	0.0258	-0.0315	0.0116	-0.0046	0.0108	-0.0093
	SStd	0.1162	0.1356	0.1382	0.1548	0.0940	0.0834	0.0820	0.0850
	MStd	0.1213	0.1237	0.1245	0.1225	0.0777	0.0777	0.0777	0.0778
SCAD	Bias	0.0426	-0.0424	0.0267	-0.0408	0.0115	-0.0093	0.0069	-0.0050
	SStd	0.1871	0.1822	0.1894	0.1559	0.0848	0.0937	0.0903	0.0834
	MStd	0.1109	0.1103	0.1099	0.1102	0.0727	0.0730	0.0731	0.0731
LASSO	Bias	-0.3423	0.3378	-0.3471	0.3512	-0.3377	0.3264	-0.3446	0.3483
	SStd	0.0986	0.0929	0.0968	0.1198	0.0895	0.0801	0.0879	0.0772
	MStd	0.0927	0.0925	0.0914	0.0918	0.0636	0.0636	0.0636	0.0637
ALASSO	Bias	-0.0415	0.0164	-0.0286	0.0262	-0.0170	0.0184	-0.0357	0.0257
	SStd	0.2023	0.1703	0.2032	0.2010	0.1196	0.1220	0.1137	0.1174
	MStd	0.1313	0.1310	0.1313	0.1316	0.0878	0.0861	0.0828	0.0863
Oracle	Bias	0.0279	-0.0141	0.0165	-0.0087	0.0075	-0.0252	0.0231	-0.0192
	SStd	0.1296	0.1132	0.1277	0.1251	0.0873	0.0857	0.0974	0.0907
	MStd	0.1342	0.1331	0.1328	0.1302	0.0836	0.0841	0.0848	0.0837

with LASSO penalty according to MMSE and estimation. Moreover they also perform as well as the oracle estimates. We also concluded that the standard error formula can perform very well for moderate sample size.

2.5.2 Horse Racing Data Analysis

In this section, stratified general transformation models and proposed variable selection procedures are applied to analyze the horse racing data in Hong Kong. The data set dates

from September 9, 2007 to July 3, 2008. A total of 730 races were recorded and there were $5 \sim 14$ horses involved in each race. For each race, only the ranking as racing results and some characteristics of participating horses, jockeys and trainers were recorded. The data set includes 20 variables listed in Tables 2.5.

The aim of this application is to select important variables for interpreting the race-track betting markets and to investigate the extent to which publicly available information is efficiently incorporated into the payoff of a race, which has been studied by many authors [8, 9, 37, 38, 41, 50, 65]. Because of the factors such as weather, pace, time (afternoon or evening) etc, the baseline survival function in general transformation model may be different for each race. Therefore, the stratified general transformation models should be considered. In this application, we use the generalized probit transformation models with strata to analyze the data. Variable selection methods with Hard thresholding, SCAD, LASSO and ALASSO penalties are used to select important variables. We run the proposed three-step MCMC-SA algorithm with the same programme parameter setting as simulation studies to conduct the application. The tuning parameters are selected by approximated GCV (2.31).

Before implementing the variable selection procedures, we first standardize the covariates. The results are summarized in Tables 2.6, 2.7 and 2.8. In the tables, "Est." and "Std" stand for the parameter estimates and corresponding estimated standard deviation calculated from (2.28). The last columns in these Tables list Z-values (the ratio of estimate and standard deviation) for the parameter estimates. By use of approximated GCV, the tuning parameters λ for variable selection methods with Hard Thresholding, SCAD, LASSO and ALASSO penalties are 0.021, 0.031, 0.022 and 0.0003 respectively. To compare the efficiencies of proposed variable selection methods, the maximum marginal likelihood estimation approach is also used to analyze the data set and the corresponding standard deviation are taken as the inverse of information matrix. Table 2.6 summarizes results for MMLE and SCAD; Table 2.7 displays results for LASSO and ALASSO penalties. The summarized results for HARD penalty are given in Table 2.8.

From Tables 2.6, 2.7 and 2.8, Z_1 is the most important factor and it is selected by all the four procedures. Moreover, the Z-value of Z_1 gets larger in the selected models. We also find that while the Z-value of Z_{17} is very small according to MMLE, it is selected

Table 2.5: The interpretation of variables considered in the horse racing application

Variables	Interpretations
Z_1	Relative strength of the horse by the public;
Z_2	Average horse speed rating;
Z_3	Number of Barrier trials since last race;
Z_4	Number of days between last race and current race;
Z_5	Age of the horse;
Z_6	Horse speed rating in last race;
Z_7	Logarithm of horse body weight, truncated at 1200 lbs;
Z_8	Weight carried change from last race;
Z_9	Weight carried by the horse in current race;
Z_{10}	Logarithm of weight allowance of the jockey;
Z_{11}	Barrial draw, centered within race;
Z_{12}	Jockey place percentage;
Z_{13}	Logarithm of place betting percentage in last race;
$Z_{13.5}$	Distance of the race;
Z_{14}	Interaction between Z_7 and $Z_{13.5}$;
$Z_{14.5}$	Number of today's race on the same surface;
Z_{15}	Interaction between Z_{11} and $Z_{14.5}$;
$Z_{15.5}$	Indicator whether the race was held in Happy Vally
Z_{16}	Interaction between Z_{11} and $Z_{15.5}$;
Z_{17}	Interaction between Z_{11} and Logrithm of $Z_{13.5}$;
$Z_{17.5}$	Indicator whether the race was held in all weather track;
Z_{18}	Interaction between Z_{11} and $Z_{17.5}$
$Z_{18.5}$	Numeric value of the hardness of the racing surface;
Z_{19}	Interaction between Z_7 and $Z_{18.5}$;
Z_{20}	Interaction among Z_{11}, sqrt($Z_{14.5}$) and $Z_{15.5}$.

Table 2.6: Summary of estimation results for horse racing data (I)

	MMLE			SCAD		
Parameters	Est.	Std	Z-value	Est.	Std	Z-value
β_1	0.9814	0.0366	26.7846	1.0310	0.0316	32.6226
β_2	-0.2486	0.0409	-6.0784	-0.2024	0.0349	-5.7977
β_3	0.0488	0.0241	2.0256	0	—	—
β_4	-0.0025	0.0004	-6.9410	-0.0020	0.0003	-6.2021
β_5	-0.0102	0.0096	-1.0642	0	—	—
β_6	-0.0972	0.0238	-4.0902	-0.0992	0.0236	-4.1996
β_7	0.7763	0.2322	3.3425	0.7251	0.2314	3.1328
β_8	-0.0057	0.0023	-2.4366	-0.0056	0.0020	-2.7495
β_9	-0.0118	0.0026	-4.5364	-0.0078	0.0022	-3.4235
β_{10}	-0.0599	0.0203	-2.9569	0	—	—
β_{11}	0.0544	0.1228	0.1230	0	—	—
β_{12}	0.3119	0.1893	1.6473	0.4282	0.1751	2.4453
β_{13}	0.0111	0.0165	0.6683	0	—	—
β_{14}	-0.0004	0.0001	-2.5267	-0.0003	0.0003	-2.4317
β_{15}	-0.0013	0.0059	-0.2195	0	—	—
β_{16}	-0.0384	0.0250	-1.5356	0	—	—
β_{17}	-0.0176	0.0391	-0.4512	-0.0018	0.0010	-1.9160
β_{18}	-0.0153	0.0115	-1.3275	0	—	—
β_{19}	-0.0982	0.0409	-2.4011	-0.0840	0.0409	-2.0521
β_{20}	0.0152	0.011	1.3361	0	—	—

by all the four procedures and its Z-value increases in all the selected models. That is, following the procedures, the interaction between Barrial draw and logarithm of Distance

Table 2.7: Summary of estimation results for horse racing data (II)

	LASSO			ALASSO		
Parameters	Est.	Std	Z-value	Est.	Std	Z-value
β_1	0.9575	0.0286	33.4373	1.0139	0.0314	32.2416
β_2	-0.1020	0.0224	-4.5440	-0.2299	0.0353	-6.5184
β_3	0	—	—	0.0502	0.0233	2.1524
β_4	-0.0022	0.0007	-3.1198	-0.0025	0.0010	-2.4510
β_5	0	—	—	0	—	—
β_6	0	—	—	-0.0897	0.0235	-3.8221
β_7	0	—	—	0.6412	0.2223	2.8847
β_8	-0.0037	0.0020	-1.8355	-0.0056	0.0022	-2.5250
β_9	-0.0055	0.0022	-2.4713	-0.0114	0.0026	-4.3118
β_{10}	-0.0305	0.0122	-2.5083	0.0581	0.0198	-2.9363
β_{11}	0	—	—	0	—	—
β_{12}	0	—	—	0.2755	0.1772	1.5546
β_{13}	0.0275	0.0110	2.5099	0	—	—
β_{14}	0	—	—	-0.0003	0.0010	-0.2980
β_{15}	0	—	—	0	—	—
β_{16}	-0.0001	0.0010	-0.1150	0	—	—
β_{17}	-0.0022	0.0010	-2.1830	-0.0020	0.0010	-2.0430
β_{18}	-0.0002	0.0010	-0.1530	-0.0081	0.0087	-0.9301
β_{19}	-0.0022	0.0040	-0.5400	-0.0862	0.0397	-2.1733
β_{20}	0	—	—	0	—	—

of the race may be an important factor to interpret the horse race betting market even if its Z-value is also small based on MMLE. In addition, following MMLE, Z_5, Z_{11}, Z_{13},

Table 2.8: Summary of estimation results for horse racing data (III)

	HARD				HARD		
Parameters	Est.	Std	Z-value	Parameters	Est.	Std	Z-value
β_1	0.9926	0.0321	30.8979	β_{11}	0.1007	0.1224	0.8230
β_2	-0.2477	0.0354	-6.9901	β_{12}	0.3567	0.1890	1.8866
β_3	0.0515	0.0240	2.1416	β_{13}	0	—	—
β_4	-0.0025	0.0003	-7.9626	β_{14}	-0.0004	0.0003	-1.1163
β_5	0	—	—	β_{15}	0	—	—
β_6	-0.1015	0.0237	-4.2821	β_{16}	-0.0309	0.0214	-1.4463
β_7	0.8036	0.2315	3.4710	β_{17}	-0.0338	0.0015	-0.8724
β_8	-0.0058	0.0022	-2.5947	β_{18}	0	—	—
β_9	-0.0119	0.0026	-4.5031	β_{19}	-0.1118	0.0410	-2.7243
β_{10}	-0.0636	0.0202	-3.1462	β_{20}	0	—	—

Z_{15}, Z_{18} and Z_{20} are not significant and they are also excluded from the selected model by at least two methods. Therefore, the six covariates should be considered as unimportant factors and they can be excluded from working model. For most of the other covariates, there are no substantial difference about Z-values between MMLE and PMMLEs. That is, they are also important for interpreting the racing betting market.

Overall, the results from all the procedures strongly show that many public available informations have not been incorporated into public betting and the horse racing betting markets are, in general, not efficient in the semistrong form. This is in line with the earlier results reported in the literature listed at the beginning of this section.

□ **End of chapter.**

Chapter 3

Variable Selection for General Transformation Models with Right Censored Data

Right censored data are the most common data in survival analysis. For the right censored data, the failure time is only observed when it occurs before some censoring time. In this chapter, we consider variable selection for general transformation models (1.3) with right censored data by non-concave (SCAD, HARD thresholding and LASSO) and Adaptive-LASSO penalties. We conduct variable selection procedures by maximizing rank-based penalized log-marginal likelihood function. We will use the three-step MCMC-SA algorithm in Chapter 2 with Gibbs sampling procedure for right censored data to find the rank-based penalized maximum marginal likelihood estimates (PMMLE). With some conditions and proper penalties, we can also show the consistency and oracle properties of PMMLEs. We illustrate the proposed procedures by some simulation examples and Primary Biliary Cirrhosis Data analysis.

3.1 Penalized Log-marginal Likelihood

Let $T \in R^+$ be the survival time variable, $Z \in R^p$ be the corresponding covariate vector and $C \in R^+$ be the censoring time variable, independent of T given Z. In this Chapter, we mainly concentrate on the right non-informative censoring with finite discrete support

on $\{c_1, c_2, \cdots, c_s\}$ and $c_1 < c_2 < \cdots < c_s$. Let $Y = \min(T, C)$ be the event time until the occurrence of some interested event and $\delta = I(T \leq C)$ be the censoring indicator. Assuming that (T, Z) are modeled by the general transformation models (1.3). Similarly to the ranking data case in Chapter 2, we assume that $\boldsymbol{\beta}$ only contains regression parameters not any transformation parameters. Otherwise if there are transformation parameters in $\boldsymbol{\beta}$, such parameters should be not penalized in rank-based penalized log-marginal likelihood. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$ and we partition $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ such that $\boldsymbol{\beta}_{10}$ contains all the nonzero effects and $\boldsymbol{\beta}_{20} = \mathbf{0}$. In this chapter, we also consider the four type penalty functions (1.8), (1.7), (1.5) and (1.6) defined in Section 1.3.

Let $\{Y_i, Z_i, \delta_i\}_{i=1}^n$ be n i.i.d. copies of (Y, Z, δ). Denote $k_n = \sum_{i=1}^n \delta_i$ as the total number of uncensored times, $\mathcal{R}_n^*$ as the partial ranking among the k_n uncensored failure times an the specified observations between each pair of uncensored observations, and $\mathcal{R}_n$ as the complete ranking of underlying failure times $\mathbf{T}_n = (T_1, T_2, \cdots, T_n)^T$. Given $\mathcal{R}_n^*$, let $\mathcal{S}_n$ be ranking set containing all possible complete ranking $\mathcal{R}_n$ and

$$\mathcal{C}_n = \{(t_1, t_2, \cdots, t_n): \quad t_{i_1} < t_{i_2} < \cdots < t_{i_{k_n}}, t_j \geq t_{i_r}, \text{for} \quad j \in \mathcal{L}_{i_r} \quad \text{and} \quad 0 \leq r \leq k_n\}$$

be the failure time set consistent with $\mathcal{R}_n^*$, where i_r is the label of the rth ordered uncensored survival time and $\mathcal{L}_{i_r}$ is the set of labels corresponding to those observations censored in interval $[T_{i_r}, T_{i_{r+1}})$ with $T_{i_0} = 0$ and $T_{i_{k_n+1}} = \infty$. Then given the partial ranking $\mathcal{R}_n^*$ and covariates $\mathbf{Z}_n = (Z_1, Z_2, \cdots, Z_n)$, the rank-based marginal likelihood function for the models (1.3) with right censored data is given by

$$\begin{aligned} L_n(\boldsymbol{\beta}|\mathcal{R}_n^*, \mathbf{Z}_n) &= \Pr(\mathcal{R}_n \in \mathcal{S}_n|\mathbf{Z}_n) = \Pr(\mathbf{T}_n \in \mathcal{C}_n|\mathbf{Z}_n) \\ &= \int_{\mathbf{t}_n \in \mathcal{C}_n} (-1)^n \prod_{i=1}^n \phi(S_0(t_i), Z_i, \boldsymbol{\beta}) \prod_{i=1}^n dS_0(t_i) \\ &= \int_{\mathbf{u}_n \in \mathcal{D}_n} \prod_{i=1}^n \phi(1 - u_i, Z_i, \boldsymbol{\beta}) d\mathbf{u}_n \end{aligned} \tag{3.1}$$

where $\phi(u, v, w) = \frac{\partial \Phi(u,v,w)}{\partial u}$, $\mathbf{u}_n = (u_1, u_2, \cdots, u_n)$ and

$$\mathcal{D}_n = \{(u_1, u_2, \cdots, u_n): \quad u_{i_1} < u_{i_2} < \cdots < u_{i_{k_n}}, u_j \geq u_{i_r}, \text{for} \quad j \in \mathcal{L}_{i_r} \quad \text{and} \quad 0 \leq r \leq k_n\},$$

is the set of uniform (0,1) vectors consistent with $\mathcal{C}_n$. The last equality in (3.1) holds because of the simple transformation $u_i = 1 - S_0(t_i)$. With the marginal likelihood

function (3.1), the rank-based penalized log-marginal likelihood function is given by

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{i=1}^{p} p_\lambda(|\beta_i|) \tag{3.2}$$

where $\ell(\boldsymbol{\beta}) = \log(L_n(\boldsymbol{\beta}|\mathcal{R}_n^*, \mathbf{Z}_n))$, λ is a penalty tuning parameter and $p_\lambda(\cdot)$ is a penalty function, which should be irregular at origin for the purpose of variable selection. We can obtain the penalized maximum marginal likelihood estimate (PMMLE) of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_n$, by maximizing (3.2) with respect to $\boldsymbol{\beta}$. With proper penalty function $p_\lambda(\cdot)$, some components of $\hat{\boldsymbol{\beta}}_n$ will be zero and they do not appear in selected models, which achieves the purpose of variable selection for the model (1.3) with right censored data.

3.2 Consistency and Oracle Properties

In this section, we establish the oracle properties of PMMLEs with our proposed variable selection procedures for right censored data. We first describe some regular conditions.

(A0) $P(C > c_s) > 0$ and $\frac{k_n}{n} \xrightarrow{p} c$ $(c > 0)$, as $n \longrightarrow \infty$.

(A1) Suppose that $\boldsymbol{\beta} \in \Theta$, a compact subset of the Euclidean space R^p and the true value $\boldsymbol{\beta}_0$ is an interior of Θ. The covariate vector Z is exogenous, and assumed to be bounded, equivalently, there exists a constant $M_1 > 0$ such that $P(||Z|| \leq M_1) = 1$. And for all $u \in (0,1)$, $||v|| \leq M_1$ and $w \in \Theta$,

$$\phi_3(u,v,w) = \frac{\partial \phi(u,v,w)}{\partial w} \quad \phi_{33}(u,v,w) = \frac{\partial^2 \phi(u,v,w)}{\partial w \partial w^T}$$

exist and are continuous with respect to $w \in \Theta$. The condition holds with $\phi(u,v,w)$ replaced by $\Phi(u,v,w)$

(A2) For any v satisfying $||v|| \leq M_1$, there are functions $F_1(u,v)$ and $F_2(u,v)$, integrable with respect to u over $(0,1)$ such that

$$||\phi_3(u,v,w)|| < F_1(u,v), \quad ||\phi_{33}(u,v,w)|| < F_2(u,v), \quad \text{for all} \quad w \in \Theta.$$

This condition also holds when $\phi(u,v,w)$ is replaced by $\Phi(u,v,w)$.

(A3) Denote $\psi(u,v,w) = \frac{\phi_3(u,v,w)}{\phi(u,v,w)}$, $\Phi_3(u,v,w) = \partial\Phi(u,v,w)/\partial w$, $\Psi(u,v,w) = \frac{\Phi_3(u,v,w)}{\Phi(u,v,w)}$ and $U = F_0(T)$ with $F_0(t) = 1 - S_0(t)$. For any $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, an neighborhood of $\boldsymbol{\beta}_0$ in Θ, $\mathrm{E}_{\boldsymbol{\beta}_0}[\psi(1-U,Z,\boldsymbol{\beta})]$, $\mathrm{E}_{\boldsymbol{\beta}_0}[\psi^2(1-U,Z,\boldsymbol{\beta})]$ and $\mathrm{E}_{\boldsymbol{\beta}_0}[-\frac{\partial\psi(1-U,Z,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}]$ exist. These expectations

also hold with $\psi(u,v,w)$ replaced by $\Psi(u,v,w)$ and the covariance-variance matrix

$$V_1(\boldsymbol{\beta}) = Var_{\boldsymbol{\beta}_0}[\psi(1-F_0(Y),Z,\boldsymbol{\beta})\delta + \Psi(1-F_0(Y),Z,\boldsymbol{\beta})(1-\delta)] \tag{3.3}$$

is positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

Denote

$$V_2(\boldsymbol{\beta}) = -\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{\partial\psi(1-F_0(Y),Z,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_T}\delta + \frac{\partial\Psi(1-F_0(Y),Z,\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}^T}(1-\delta)\right\} \tag{3.4}$$

From conditions (A1)-(A3), it can be easily shown that $V_1(\boldsymbol{\beta}_0) = V_2(\boldsymbol{\beta}_0)$.

(A4) $\mathrm{E}_{\boldsymbol{\beta}_0}[\Psi^2(1-F_0(C),Z,\boldsymbol{\beta}_0)]$ exists and there exists a constant $M_2 > 0$ such that

$$\max\left\{\mathrm{E}_{\boldsymbol{\beta}_0}[\psi^2(1-U,Z,\boldsymbol{\beta}_0)], \mathrm{E}_{\boldsymbol{\beta}_0}\left[\Psi^2(1-U,Z,\boldsymbol{\beta}_0)\right], \mathrm{E}_{\boldsymbol{\beta}_0}[\Psi^2(1-F_0(C),Z,\boldsymbol{\beta}_0)]\right\} < M_2$$

(A5) The function $\psi(u,v,w)$ is continuous for $u \in (0,1)$ and satisfies Lipschitz condition with respect to u in any closed subset of (0,1). That is, for any $[L,R] \subset (0,1)$, there exists a constant $M_3(L,R)$, dependent on $[L,R]$, such that for all $u_1, u_2 \in [L,R]$, $|v| \le M_1$ and $w \in \mathcal{O}_{\boldsymbol{\beta}_0}$,

$$|\psi(u_1,v,w) - \psi(u_2,v,w)| \le M_3|u_1-u_2|.$$

This condition holds when $\psi(u,v,w)$ is replaced by $\Psi(u,v,w)$.

(A6) The function $\frac{\partial\psi(u,v,w)}{\partial w^T}$ satisfies Lipschitz condition with respect to u and this condition holds when $\psi(u,v,w)$ is replaced by $\Psi(u,v,w)$.

(A7) For any discretization (please refer to Section 2.3 in Wu (2008) [71]), there exists N such that when $n > N$,

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left[-\frac{\partial}{\partial\boldsymbol{\beta}^T}S_n(\boldsymbol{\beta})\right] \ge \mathrm{E}_{\boldsymbol{\beta}_0}\left[-\frac{\partial}{\partial\boldsymbol{\beta}^T}S_{n,m}(\boldsymbol{\beta})\right]$$

where $S_n(\boldsymbol{\beta})$ and $-\frac{\partial S_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}$ are the score function and Fisher information matrix with respect to marginal likelihood function (3.1); Given m discretization points for failure time variable, $S_{n,m}(\boldsymbol{\beta})$ and $-\frac{\partial S_{n,m}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}$ are the discretized versions of $S_n(\boldsymbol{\beta})$ and $-\frac{\partial S_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}$. The formula and meanings of discretized score function and Fisher information matrix can be found in Section 2.3 in Wu (2008) [71]. $S_n(\boldsymbol{\beta})$ and $\frac{\partial S_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}$ are given by

$$\begin{aligned} S_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\int_{\mathbf{u}_n\in\mathcal{D}_n}\psi(1-u_i,Z_i,\boldsymbol{\beta})p(\mathbf{u}_n|\mathcal{R}_n^*,\mathbf{Z}_n)d\mathbf{u}_n \\ &= \sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{\beta}}[\psi(1-U_i,Z_i,\boldsymbol{\beta})|\mathcal{R}_n^*,\mathbf{Z}_n] \end{aligned} \tag{3.5}$$

$$\frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = \mathrm{E}_{\boldsymbol{\beta}} \left\{ \left[\sum_{i=1}^{n} \xi_i(\boldsymbol{\beta}) \right] \middle| \mathscr{F}_n \right\} + \mathrm{Var}_{\boldsymbol{\beta}} \left\{ \left[\sum_{i=1}^{n} \xi_i(\boldsymbol{\beta}) \right] \middle| \mathscr{F}_n \right\} \tag{3.6}$$

where $\xi_i(\boldsymbol{\beta}) = \delta_i \frac{\partial \psi_i(1-U_i, Z_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} + (1-\delta_i) \sum_{j=1}^{n} \frac{\partial \Psi(1-U_j, Z_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \delta_j I(i \in \mathcal{L}_j)$ and

$$p(\mathbf{u}_n | \mathcal{R}_n^*, \mathbf{Z}_n) = \frac{I(\mathbf{u}_n \in \mathcal{D}_n)}{L_n(\boldsymbol{\beta} | \mathcal{R}_n^*, \mathbf{Z}_n)} \prod_{i=1}^{n} \phi(1-u_i, Z_i, \boldsymbol{\beta}) \tag{3.7}$$

is the joint conditional density function of $U_i = 1 - S_0(T_i)$ for $i = 1, 2, \cdots, n$, given $\mathcal{R}_n^*$ and $\mathbf{Z}_n$. Thus all the expectations and variance in (3.5) and (3.6) are respective to the density (3.7).

Under condition (A2), we can show that

$$\begin{aligned} S_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\beta}} \left[\psi(1-U_i, Z_i, \boldsymbol{\beta}) | \mathscr{F}_n \right] \\ &= \sum_{r=1}^{k_n} \mathrm{E}_{\boldsymbol{\beta}} \left[\psi(1-U_{i_r}, Z_{i_r}, \boldsymbol{\beta}) + \sum_{j \in \mathcal{L}_{i_r}} \Psi(1-U_{i_r}, Z_j, \boldsymbol{\beta}) \middle| \mathscr{F}_n \right] \\ &= \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\beta}} \left[\psi(1-U_i, Z_i, \boldsymbol{\beta}) \delta_i + (1-\delta_i) \sum_{j=1}^{n} \Psi(1-U_j, Z_i, \boldsymbol{\beta}) \delta_j I(i \in \mathcal{L}_j) \middle| \mathscr{F}_n \right]. \end{aligned} \tag{3.8}$$

Conditions (A1) and (A3) are the regular conditions for the models (1.3) while (A2) allows the interchangeability of order for differentiation and integration or sum. (A0) and (A4)-(A6) can be used to prove the asymptotic normality of score function $S_n(\boldsymbol{\beta}_0)$. (A6) and (A7) will be used to prove the oracle properties of PMMLEs. The inequality

$$\mathrm{E}_{\boldsymbol{\beta}_0} \left[-\frac{\partial}{\partial \boldsymbol{\beta}^T} S_n(\boldsymbol{\beta}) \right] \geq \mathrm{E}_{\boldsymbol{\beta}_0} \left[-\frac{\partial}{\partial \boldsymbol{\beta}^T} S_{n,m}(\boldsymbol{\beta}) \right]$$

should hold for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, otherwise, use of discretized data is better than use of original data.

3.2.1 Results of Variable Selection Methods with Nonconcave Penalties

We first present the results of variable selection methods with non-concave penalties. Denote

$$a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \quad \beta_{j0} \neq 0\} \quad \text{and} \quad b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \quad \beta_{j0} \neq 0\},$$

where $p_\lambda(\cdot)$ can be any one of (1.8), (1.7) and (1.5). Then we have,

Theorem 3.1 (Consistency) *Under conditions (A0)-(A7), if $b_n \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ of $Q(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$.*

From this theorem, we can see that with proper tuning parameter λ_n, as long as $a_n = O(n^{-1/2})$, there exists a $\sqrt{n}$-consistent penalized maximum marginal likelihood estimate of $\boldsymbol{\beta}_0$.

To present the oracle properties, denote

$$\Sigma_{\lambda_n} = \text{diag}\{p''_{\lambda_n}(|\beta_{10}|), p''_{\lambda_n}(|\beta_{20}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)\}$$

and

$$\mathbf{b}_{\lambda_n} = (p'_{\lambda_n}(|\beta_{10}|)\text{sign}(\beta_{10}), p'_{\lambda_n}(|\beta_{20}|)\text{sign}(\beta_{20}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)\text{sign}(\beta_{s0}))^T$$

Then we have **Theorem 3.2** (Oracle) *Assume that the penalty function $p_{\lambda_n}(\beta)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\beta\to 0+} p'_{\lambda_n}(\beta)/\lambda_n > 0.$$

If $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem 3.1, with probability tending to 1, the $\sqrt{n}$-consistent local maximizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \hat{\boldsymbol{\beta}}_{2n}^T)^T$ in Theorem 3.1 must satisfy:

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$;

(ii) (Asymptotic normality)

$$\sqrt{n}(V + \Sigma_{\lambda_n})\{\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10} + (V + \Sigma_{\lambda_n})^{-1}\mathbf{b}_{\lambda_n}\} \to N(\mathbf{0}, V) \tag{3.9}$$

where V is the upper leading $s \times s$ submatrix of $V_2(\boldsymbol{\beta}_0)$ defined by (3.4).

Remark 3.2 From HARD thresholding and SCAD penalties, we can see that if $\lambda_n \to 0$, then for sufficiently large n, $a_n = 0$, $\Sigma_{\lambda_n} = \mathbf{0}$ and $\mathbf{b}_{\lambda_n} = \mathbf{0}$. Therefore, when $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \to N(\mathbf{0}, V^{-1}) \text{ and } \hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}.$$

That is, with proper tuning parameter, PMMLEs with HARD and SCAD penalties enjoy the oracle properties. They perform very well as if we have known $\boldsymbol{\beta}_2 = \mathbf{0}$ in advance when we estimate $\boldsymbol{\beta}_1$. However, for LASSO penalty, $a_n = \lambda_n$ then the conditions $a_n = O(n^{-1/2})$ and $\sqrt{n}\lambda_n \to \infty$ in Theorem 3.2 contradict. So PMMLEs with LASSO penalty can not enjoy oracle properties.

3.2.2 Results of Variable Selection Method with Adaptive-LASSO Penalty

In this section, we present the consistency and Oracle properties of PMMLEs with Adaptive-LASSO penalty (1.6).

Theorem 3.3 (Consistency) *Under the conditions (A0)-(A7), if* $\sqrt{n}\lambda_n = O(1)$, *then there exists a maximizer* $\hat{\boldsymbol{\beta}}_n$ *of* $Q(\boldsymbol{\beta})$ *with ALASSO penalty such that* $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$.

This theorem shows that, with proper tuning parameter λ_n, the PMMLEs $\hat{\boldsymbol{\beta}}_n$ with ALASSO penalty enjoy $\sqrt{n}$-consistency.

Theorem 3.4 (Oracle) *If* $\sqrt{n}\lambda_n \to 0$ *and* $n^{(1+\gamma)/2}\lambda_n \to \infty$ *for some* $\gamma > 0$, *then under conditions of Theorem 3.3, the local maximizer* $\hat{\boldsymbol{\beta}}_n$ *must satisfy:*

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$;

(ii) (Asymptotic Normality)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(\mathbf{0}, V^{-1}) \quad as \quad n \to \infty, \tag{3.10}$$

where V *is defined in Theorem 3.2.*

Based on the Theorem 3.4, we can see that, with proper tuning parameter λ_n, the PMMLEs with Adaptive-LASSO penalty also enjoy oracle properties as if we have known which effects are equal to 0 in advance when we estimate $\boldsymbol{\beta}_1$.

3.2.3 Proof of Theorems

In this section, we prove the Theorems 3.1-3.4 in Sections 3.2.1 and 3.2.2 by similar method to that in Section 2.2.3. Before proving the Theorems, we also introduce some Lemmas as follows. All the proofs of Lemmas can be found in Section 2.3 of Wu (2008) [71].

Lemma 3.1 Under the conditions (A0)-(A6),

$$\frac{1}{\sqrt{n}} S_n(\boldsymbol{\beta}_0) \xrightarrow{D} N(0, V_1(\boldsymbol{\beta}_0)) \quad \text{as} \quad n \longrightarrow \infty \tag{3.11}$$

where $V_1(\boldsymbol{\beta})$ is define by (3.3).

Lemma 3.2 Under the assumptions (A0)-(A7), it holds that

$$-\frac{1}{n}\frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \xrightarrow{P} V_2(\boldsymbol{\beta}) \quad \text{as} \quad n \to \infty \tag{3.12}$$

uniformly for all $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, where $V_2(\boldsymbol{\beta})$ is defined in (3.4).

Lemma 3.3 Under the conditions (A0) - (A7), there exists a sequence of $\tilde{\boldsymbol{\beta}}_n$ satisfying $S_n(\tilde{\boldsymbol{\beta}}_n) = 0$ such that as $n \longrightarrow 0$

(i) $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}_0$;

(ii) $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, V_2(\boldsymbol{\beta}_0)^{-1})$.

This Lemma shows that the maximum marginal likelihood estimate of $\boldsymbol{\beta}$ in general transformation models with right censored data is $\sqrt{n}$-consistent and is distributed asymptotically normally.

Proof of Theorem 3.1 Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any given $1 - \varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \sup_{||\boldsymbol{u}||=C} Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) < Q(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon. \tag{3.13}$$

Based on that $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\theta) > 0$, we have

$$\begin{aligned} &\frac{1}{n}\left[Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - Q(\boldsymbol{\beta}_0)\right] \\ \leq &\frac{1}{n}\left[\ell(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}_0)\right] - \sum_{j=1}^{s}[p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)], \end{aligned} \tag{3.14}$$

By Lemmas 3.1 and 3.2, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ \ ||\boldsymbol{u}|| = C\}$, we have

$$\begin{aligned} &\frac{1}{n}\left[\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0)\right] \\ =&\frac{1}{n}\left[\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T V_2(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\{1 + O_P(1)\}\right] \\ =&-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T[V_2(\boldsymbol{\beta}_0) + O_P(1)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_P(n^{-1/2}) \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \\ =&-\frac{1}{2}\alpha_n^2 \boldsymbol{u}^T[V_2(\boldsymbol{\beta}_0) + O_P(1)]\boldsymbol{u} + O_P(n^{-1/2}\alpha_n||\boldsymbol{u}||) \end{aligned} \tag{3.15}$$

Note that $V_2(\boldsymbol{\beta}_0)$ is a positive definite matrix. The order for the first term in the last equality of (3.15) is $C^2\alpha_n^2$ and for second one is $\alpha_n^2 C$. Therefore, for a sufficiently large C, the second term is dominated by the first term in the last equation of (3.15). On the other hand, by Taylor's expansion, the second term of (3.14) is bounded by

$$\sqrt{s}\alpha_n a_n||\boldsymbol{u}|| + \alpha_n^2 b_n||\boldsymbol{u}||^2 = C\alpha_n^2(\sqrt{s} + b_n C).$$

If $b_n \to 0$, the second term of (3.14) is dominated by the first term of (3.15). Thus, for a sufficiently large C, (3.13) holds, which means that there exists a local maximum in the ball $\{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ ||\boldsymbol{u}|| \leq C\}$ with probability at least $1 - \varepsilon > 0$. Therefore, there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$. □

Proof of Theorem 3.2 (i) It is sufficient to prove that

$$Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) = \max_{||\boldsymbol{\beta}_2|| \leq Cn^{-1/2}} Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) \tag{3.16}$$

for any given $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_P(n^{-1/2})$ and any constant C.

From Lemmas 3.1-3.2 and $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$, we have that

$$\begin{aligned} ||S_n(\boldsymbol{\beta})|| &= ||S_n(\boldsymbol{\beta}_0) + \frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| + O_P(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||^2) \\ &= O_P(n^{1/2}) \end{aligned} \tag{3.17}$$

So for $j = s+1, s+2, \cdots, p$,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= S_{nj}(\boldsymbol{\beta}) - n p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) \\ &= -n p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) + O_P(\sqrt{n}) \\ &= n\lambda_n\{-\lambda_n^{-1} p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) + O_P(\frac{1}{\sqrt{n}\lambda_n})\}. \end{aligned} \tag{3.18}$$

where $S_{nj}(\boldsymbol{\beta})$ is jth element of $S_n(\boldsymbol{\beta})$. Since $\liminf_{n\to\infty} \liminf_{\theta \to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$, and $\frac{1}{\sqrt{n}\lambda_n} \to 0$, the derivative and $-\beta_j$ have the same sign. Therefore (3.16) holds.

(ii) From $a_n = O(n^{-1/2})$ and Theorem 3.1, there exists a local $\sqrt{n}$-consistent maximizer, $\hat{\boldsymbol{\beta}}_{1n}$, of $Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T)$ satisfying

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T} = 0 \quad \text{for} \quad j = 1, 2, \cdots, s. \tag{3.19}$$

Set $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T$ and $S_{1n}(\boldsymbol{\beta})$ as the vector consisting of the first s components of $S_n(\boldsymbol{\beta})$, then

$$\begin{aligned} 0 = \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n} &= \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ &= S_{1n}(\boldsymbol{\beta}_0) - n\boldsymbol{b}_{\lambda_n} + \frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) - n\Sigma_{\lambda_n}(\boldsymbol{\beta}_1^*)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \end{aligned} \tag{3.20}$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{T*}, \boldsymbol{\beta}_2^{T*})^T$ lies on the line segment between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$; $\Sigma_{\lambda_n}(\boldsymbol{\beta}_1) = \text{diag}(p''_{\lambda_n}(|\beta_1|), p''_{\lambda_n}(|\beta_2|), \cdots, p''_{\lambda_n}(|\beta_s|))$. From Theorem 3.1, Lemma 3.1 and 3.2, (3.9) holds. This completes the proof. □

Proof of Theorem 3.3 Similar to the proofs of Theorem 3.1 and 2.3, let $\alpha_n = n^{-1/2}$. It is also sufficient to show that for any given $1 - \varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \sup_{||\boldsymbol{u}||=C} Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) < Q(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon. \tag{3.21}$$

Note that

$$\begin{aligned} &\frac{1}{n}[Q(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - Q(\boldsymbol{\beta}_0)] \\ \leq &\frac{1}{n}[\ell(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}_0)] - \lambda_n \sum_{j=1}^{s} [(|\beta_{j0} + \alpha_n u_j| - |\beta_{j0}|)/|\tilde{\beta}_j|^{\gamma}] \\ \leq &\frac{1}{n}[\ell(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}_0)] + \lambda_n \alpha_n \sum_{j=1}^{s} [|u_j|/|\tilde{\beta}_j|^{\gamma}] \end{aligned} \tag{3.22}$$

By Lemmas 3.1 and 3.2, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ \ ||\boldsymbol{u}|| = C\}$, we have

$$\begin{aligned} &\frac{1}{n}[\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0)] \\ = &\frac{1}{n}\left[S_n(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \frac{\partial S_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\{1 + O_P(1)\} \right] \\ = &-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T [V_2(\boldsymbol{\beta}_0) + O_P(1)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_P(n^{-1/2}) \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \\ = &-\frac{1}{2}\alpha_n^2 \boldsymbol{u}^T [V_2(\boldsymbol{\beta}_0) + O_P(1)]\boldsymbol{u} + O_P(n^{-1/2}\alpha_n ||\boldsymbol{u}||) \end{aligned} \tag{3.23}$$

Note that $V_2(\boldsymbol{\beta}_0)$ is a positive definite matrix. The order for first term in the last equality of (3.23) is $C^2\alpha_n^2$ and for second one is $\alpha_n^2 C$. Therefore, for a sufficiently large C, the second term is dominated by the first term in the last equality. On the other hand, by Taylor's expansion and Lemma 3.8,

$$\frac{1}{|\tilde{\beta}_j|^{\gamma}} = \frac{1}{|\beta_{j0}|^{\gamma}} - \frac{\gamma \text{sign}(\beta_{j0})}{|\beta_{j0}|^{\gamma+1}}(\tilde{\beta}_j - \beta_{j0}) + o_P(\tilde{\beta}_j - \beta_{j0}) = \frac{1}{|\beta_{j0}|^{\gamma}} + \frac{O_P(1)}{\sqrt{n}}$$

and hence the second term of (3.22) is bounded by $C\alpha_n^2$, since

$$\begin{aligned} |\alpha_n \lambda_n| \sum_{j=1}^{s} [|u_j|/|\tilde{\beta}_j|^{\gamma}] &= \frac{1}{\sqrt{n}} \lambda_n \sum_{j=1}^{s} \left[\frac{|u_j|}{|\beta_{j0}|^{\gamma}} + \frac{|u_j|}{\sqrt{n}} O_P(1) \right] \\ \leq &Cn^{-1/2}\lambda_n O_P(1) = Cn^{-1}(\sqrt{n}\lambda_n) O_P(1) \leq C\alpha_n^2 O_P(1) \end{aligned}$$

and

$$\sqrt{n}\lambda = O(1).$$

Therefore, the second term of (3.22) is also dominated by the first term of (3.23). Thus, for a sufficiently large C, (3.21) holds, which means that there exists a local maximizer in the ball $\{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}, \ ||\boldsymbol{u}|| \leq C\}$ with probability at least $1 - \varepsilon > 0$. Therefore, there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$. □

Proof of Theorem 3.4 (i)Similar to the proofs of Theorem 3.2 and 2.4, it is sufficient to prove that

$$Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) = \max_{||\boldsymbol{\beta}_2|| \leq Cn^{-1/2}} (Q(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) \tag{3.24}$$

for any given $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_P(n^{-1/2})$ and any constant C.

Since $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$,

$$\begin{aligned} ||S_n(\boldsymbol{\beta})|| &= ||S_n(\boldsymbol{\beta}_0) + \frac{\partial S_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| + O_P(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||^2) \\ &= O_P(n^{1/2}) \end{aligned}$$

So from Lemma 3.3, for $j = s+1, s+2, \cdots, p$, we have

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= S_{nj}(\boldsymbol{\beta}) - n\lambda_n \frac{\text{sign}(\beta_j)}{|\tilde{\beta}_j|^\gamma} \\ &= -(n\lambda_n) n^{\gamma/2} \frac{\text{sign}(\beta_j)}{|\sqrt{n}\tilde{\beta}_j|^\gamma} + O_P(\sqrt{n}) \\ &= \sqrt{n}\left[O_P(1) - (n^{(\gamma+1)/2}\lambda_n)\frac{\text{sign}(\beta_j)}{|O_P(1)|}\right] \end{aligned}$$

where $S_{nj}(\boldsymbol{\beta})$ is jth element of $S_n(\boldsymbol{\beta})$. Since $n^{(\gamma+1)/2}\lambda_n \to \infty$ as $n \to \infty$, the derivative and $-\beta_j$ have the same sign. Therefore (3.24) holds.

(ii) From Theorem 3.3, there exists a local $\sqrt{n}$-consistent maximizer, $\hat{\boldsymbol{\beta}}_{1n}$, of $Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T)$ satisfying

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T} = 0 \quad \text{for} \quad j = 1, 2, \cdots, s. \tag{3.25}$$

Set $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T$ and denote $S_{1n}(\boldsymbol{\beta})$ as a vector consisting of the first s elements of $S_n(\boldsymbol{\beta})$, then

$$\begin{aligned} 0 = \frac{\partial Q(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_1} &= \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \\ &= S_{1n}(\boldsymbol{\beta}_{10}) - n\lambda_n \left(\frac{\text{sign}(\beta_{10})}{|\tilde{\beta}_1|^\gamma}, \frac{\text{sign}(\beta_{20})}{|\tilde{\beta}_2|^\gamma}, \cdots, \frac{\text{sign}(\beta_{s0})}{|\tilde{\beta}_s|^\gamma}\right)^T \cdot \\ &+ \frac{\partial S_{1n}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) - n\lambda_n \Sigma_{\lambda_n}(\boldsymbol{\beta}_1^*)(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \end{aligned} \tag{3.26}$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{T*}, \boldsymbol{\beta}_2^{T*})^T$ lies on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$; $\Sigma_{\lambda_n}(\boldsymbol{\beta}_1) = \text{diag}(p''_{\lambda_n}(|\beta_1|), p''_{\lambda_n}(|\beta_2|), \cdots, p''_{\lambda_n}(|\beta_s|))$. From $\sqrt{n}\lambda_n \to 0$, $\tilde{\boldsymbol{\beta}}_1 \to \boldsymbol{\beta}_{10}$, Theorem 3.3, Lemmas 3.1-3.3, (3.10) holds. This completes the proof. □

3.3 Implementation

From (3.2), the first term of penalized log-marginal likelihood function involves a high-dimensional integration and the integration has no closed form. So it is very difficult to directly maximizing (3.2) with respect to $\boldsymbol{\beta}$. Note that the difference between the integrations (2.1) and (3.1) is only the integration region $\mathcal{D}_n$ because of the different data type. Therefore, by virtue of following Gibbs sampling procedure with right censored data, we can use the three-step MCMC-SA algorithm in Chapter 2 to find PMMLEs for right censored data. In the variable selection procedures, we also use approximated GCV (2.31), to selection proper tuning parameter λ. In the approximated GCV (2.31), $\ell(\hat{\boldsymbol{\beta}}_n)$ for the right censored data can be also approximated by an important sampling method.

Gibbs Sampling with Right Censored Data:

Let $\mathbf{U}_n = (U_1, U_2, \cdots, U_n)^T$ be n independent random variables and U_i has the survival function $\Phi(1-u, Z_i, \boldsymbol{\beta})$. Given $\mathcal{R}_n^*$, define

$$\mathcal{E} = \{(U_1, U_2, \cdots, U_n): \ U_{i_1} < U_{i_2} < \cdots < U_{i_{k_n}}, U_j \geq U_{i_r}, \text{for } j \in \mathcal{L}_{i_r} \text{ and } 0 \leq r \leq k_n\}.$$

Then the density distribution of $\mathbf{U}_n$ condition on $\mathbf{U}_n \in \mathcal{E}$ is (3.7). Moreover, for fixed U_j's $(j \neq i)$, if U_i is not censored and $i = i_k$, U_i has the distribution function $1 - \Phi(1-u, Z_i, \boldsymbol{\beta})$ and it is restricted in $(U_{i_{k-1}}, U_{i_{k+1}})$; Otherwise if U_i is censored in $[U_{i_k}, U_{i_{k+1}})$, U_i also has the distribution function $1 - \Phi(1-u, Z_i, \boldsymbol{\beta})$ but it is restricted in $[U_{i_k}, 1)$.

Given $\mathcal{R}_n^*$ and $\mathbf{Z}_n$, we can generate samples of $\mathbf{U}_n = (U_1, U_2, \cdots, U_n)^T$ from the stationary probability (3.7) through the following Gibbs sampling procedure. Let $\mathbf{U}_n^k = (U_{1,k}, U_{2,k}, \cdots, U_{n,k})^T$ be the current sample of $\mathbf{U}_n$. Then the next sample $\mathbf{U}_n^{k+1}$ can be generated as follows,

0. Set $j = 1$;

1. if $\delta_j = 1$ and $j = i_r$

$$\text{Set } u_j^- = 1 - \Phi(1 - U_{i_{r-1},k}, Z_j, \boldsymbol{\beta}) \text{ and } u_j^+ = 1 - \Phi(1 - U_{i_{r+1},k}, Z_j, \boldsymbol{\beta}),$$

else if $\delta_j = 0$ and $j \in \mathcal{L}_{i_r}$

Set $u_j^- = 1 - \Phi(1 - U_{i_r,k}, Z_j, \boldsymbol{\beta})$ and $u_j^+ = 1$

with $u_0^- = 0$ and $u_{n+1}^+ = 1$;

2. Generate U^* from Unif$[u_j^-, u_j^+]$ and set $U_{j,k+1} = 1 - \Phi^{-1}(1 - U^*, Z_j, \boldsymbol{\beta})$, where $\Phi^{-1}(u, v, w)$ is the inverse function of $\Phi(u, v, w)$ with respect to u.

3. If $j < n$, then $j = j + 1$ and go to step 1. Otherwise stop.

Approximation of $\ell(\boldsymbol{\beta})$ with right censored data:

To approximate $\ell(\boldsymbol{\beta})$, we first approximate $L_n(\boldsymbol{\beta}|\mathcal{R}_n^*, \mathbf{Z}_n)$. Similar to the ranking data case, we can express the marginal likelihood function $L_n(\boldsymbol{\beta}|\mathcal{R}_n^*, \mathbf{Z}_n)$ as an expectation with respect to some one probability and then use important sampling procedure to approximate $L_n(\boldsymbol{\beta}|\mathcal{R}_n^*, \mathbf{Z}_n)$. From (3.7), we can easily find that

$$p_0(\mathbf{u}_n|\mathcal{R}_n^*, \mathbf{Z}_n = \mathbf{0}) = \frac{n!}{c} I(\mathbf{u}_n \in \mathcal{D}_n) \tag{3.27}$$

is the conditional baseline density function of $\mathbf{U}_n$ given $\mathcal{R}_n^*$ and the baseline covariates $\mathbf{Z}_n = \mathbf{0}$, where c is the total number of all possible rankings in $\mathcal{S}_n$. Following Lam and Leung (2001) [48], we can easily find that

$$c = \left\{\prod_{j=1}^{d_0}(n - j + 1)\right\}^{I(d_0>0)} \prod_{l=1}^{k_n}\left\{\prod_{j=1}^{d_l}(c_l - j)\right\}^{I(d_l>0)}$$

where d_l is the number of the event times censored in $[t_{i_l}, t_{i_{l+1}})$ and $c_l = (n-l+1) - \sum_{j=0}^{l-1} d_j$ $(l \geq 1)$ is the number of individuals at risk just prior to t_{i_l}. Note that here we include the case that there may be event times censored before t_{i_1}. So the total number of all possible rankings in $\mathcal{S}_n$ is slightly different from that in Lam and Leung (2001) [48] but they are same when we eliminate the individuals censored before t_{i_1} from our data set. Multiply and divide the integrand in (3.1) by (3.27), the marginal likelihood function can be expressed as

$$L_n(\boldsymbol{\beta}|\mathcal{R}_n^*, \mathbf{Z}_n) = \frac{c}{n!} \mathrm{E}_{p_0}\left[\prod_{i=1}^{n} \phi(1 - U_i, Z_i, \boldsymbol{\beta}) \middle| \mathcal{R}_n^*, \mathbf{Z}_n\right] \tag{3.28}$$

where E_{p_0} means that the conditional expectation (3.28) is respective to the density (3.27) given $\mathcal{R}_n^*$ and $\mathbf{Z}_n$. Hence the important sampling procedure can be used to approximate

$L_n(\beta|\mathcal{R}_n^*, \mathbf{Z}_n)$. Assuming that $\{(U_{i,1}, U_{i,2}, \cdots, U_{i,n})^T\}_{i=1}^{M_0}$ are the M_0 simulated sets of $\mathbf{U}_n$ from (3.27) by above Gibbs sampling procedure with the baseline covariates $\mathbf{Z}_n = \mathbf{0}$, then $L_n(\beta|\mathcal{R}_n^*, \mathbf{Z}_n)$ can be approximated by

$$\hat{L}_n(\beta|\mathcal{R}_n^*, \mathbf{Z}_n) = \frac{c}{n!}\frac{1}{M_0}\sum_{i=1}^{M_0}\left\{\prod_{k=1}^{n}\phi(1-U_{i,k}, Z_k, \beta)\right\} \tag{3.29}$$

Based on the M_0 simulated sets of $\mathbf{U}_n$, we can also similarly approximate the corresponding score function $S_n(\beta)$ and Fisher information matrix $-\frac{\partial}{\partial\beta^T}S_n(\beta)$. Note that the approximations of $S_n(\beta)$ and $-\frac{\partial}{\partial\beta^T}S_n(\beta)$ are independent on c, which can be easily seen from (3.5) and (3.6) in Section 3.2. Based on above approximation procedure, we can further approximate GCV (2.31) for each tuning parameter λ. Given the covariates $\mathbf{Z}_n$, to efficiently assess the selected models, the same M_0 simulated sets of $\mathbf{U}_n$ will be used to compute the approximated GCV (2.31) regardless of the values of β and λ.

3.4 Numeric Studies

In this section, we illustrate our proposed variable selection procedures by three simulation examples and one real data application. With right censored data, we firstly conduct simulations for proportional hazard regression models (PH), proportional odds regression models (PO) and generalized probit transformation models (GP), which are special models of general transformation models (1.3). Then we apply the procedure to the analysis of *Primary Biliary Cirrhosis (PBC) Data* through generalized probit model.

According to Tibshirani (1997) [68], we use median of mean squared error (MMSE) $(\hat{\beta} - \beta_0)^T\Sigma(\hat{\beta} - \beta_0)$ over 100 runs to evaluate the efficiency of the proposed variable selection methods, where Σ is the population covariance matrix of regressors and can be estimated by covariate sample.

3.4.1 Simulation Studies

In these simulation studies, 100 data sets consisting of $n = 100$ and 200 observations are simulated from the models

$$S_Z(t) = \Phi(S_0(t), Z, \beta) \tag{3.30}$$

Table 3.1: Variable Selection results for PH, PO and GP models with right censored data under $C_r = 25\%$

$C_r = 25\%$		$n = 100$			$n = 200$		
			Aver. no. of 0 Coef.			Aver. no. of 0 Coef.	
Models	Penalty	MMSE	correct	incorrect	MMSE	correct	incorrect
PH	HARD	0.091	4.500	0.000	0.054	4.780	0.000
	SCAD	0.094	4.730	0.003	0.048	4.970	0.000
	LASSO	0.620	4.610	0.005	0.594	4.880	0.000
	ALASSO	0.089	4.630	0.010	0.064	4.820	0.000
	Oracle	0.068	5.000	0.000	0.046	5.000	0.000
PO	HARD	0.334	4.450	0.023	0.096	4.880	0.003
	SCAD	0.374	4.730	0.110	0.103	4.830	0.035
	LASSO	0.776	4.640	0.015	0.716	4.770	0.003
	ALASSO	0.277	4.570	0.065	0.095	4.820	0.015
	Oracle	0.190	5.000	0.000	0.084	5.000	0.000
GP	HARD	0.094	4.620	0.000	0.081	4.490	0.000
	SCAD	0.079	4.620	0.000	0.028	4.970	0.000
	LASSO	0.604	4.747	0.005	0.608	4.947	0.000
	ALASSO	0.081	4.880	0.000	0.030	5.000	0.000
	Oracle	0.058	5.000	0.000	0.029	5.000	0.000

Note: 0.000*s indicate that the corresponding values are less than 0.0005.

where $\Phi(1-u,v,w) = h^{-1}(h(u)+v^Tw)$ and $S_0(t) = e^{-t}$. For $h^{-1}(u)$, we consider three cases: (i) standard extreme value survival function, (ii) standard logistic survival function and (iii) standard normal survival function; Then (3.30) correspond to proportional hazards regression models, proportional odds regression models and generalized probit models respectively; $Z \in R^9$ and Z_i follows standard normal distribution independently;

Table 3.2: Variable Selection results for PH, PO and GP models with right censored data under $C_r = 10\%$

$C_r = 10\%$		$n = 100$			$n = 200$		
			Aver. no. of 0 Coef.			Aver. no. of 0 Coef.	
Models	Penalty	MMSE	correct	incorrect	MMSE	correct	incorrect
PH	HARD	0.084	4.550	0.000	0.044	4.800	0.000
	SCAD	0.081	4.730	0.000	0.045	4.980	0.000
	LASSO	0.502	4.600	0.000	0.450	4.730	0.000
	ALASSO	0.095	4.540	0.029	0.064	4.820	0.000
	Oracle	0.065	5.000	0.000	0.034	5.000	0.000
PO	HARD	0.182	4.550	0.010	0.093	4.770	0.003
	SCAD	0.114	4.480	0.000	0.053	4.960	0.000
	LASSO	0.642	4.420	0.005	0.612	4.590	0.000
	ALASSO	0.248	4.700	0.065	0.088	4.900	0.010
	Oracle	0.119	5.000	0.000	0.059	5.000	0.000
GP	HARD	0.072	4.520	0.000	0.029	4.840	0.000
	SCAD	0.066	4.840	0.000	0.026	4.980	0.000
	LASSO	0.535	4.67	0.000	0.515	4.926	0.000
	ALASSO	0.063	4.930	0.000	0.025	5.000	0.000
	Oracle	0.056	5.000	0.000	0.026	5.000	0.000

$\boldsymbol{\beta} = (0.8, 0, -0.8, 0, 0, 0.8, 0, 0, -0.8)^T$; For each special models, we generate censoring times from corresponding models (3.30) with $\Phi(1 - u, v, w) = h^{-1}(h(u) + v^T w + \mu_0)$. By choosing the proper values of μ_0, we consider two censoring ratios (C_r) — 10% and 25%.

We run the three-step MCMC-SA algorithm in Chapter 2 with Gibbs sampling for right censored data in Section 3.3 to conduct the simulation studies. We will take the same programme parameter setting in the algorithm as the simulation studies in Section

Table 3.3: Summary of estimation results for nonzero effects in PH model with right censored data under $C_r = 25\%$

$C_r = 25\%$		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.0712	0.0464	-0.0546	0.0838	-0.0855	0.0835	-0.0575	0.0671
	SStd	0.1222	0.1389	0.1348	0.1234	0.0914	0.0949	0.0837	0.1046
	MStd	0.1417	0.1416	0.1395	0.1377	0.0914	0.0915	0.0923	0.0914
SCAD	Bias	-0.0731	0.0866	-0.0580	0.0583	-0.0768	0.0819	-0.0924	0.0763
	SStd	0.1316	0.1465	0.1623	0.1322	0.0912	0.0910	0.0847	0.0832
	MStd	0.1347	0.1344	0.1341	0.1353	0.0901	0.0885	0.0894	0.0898
LASSO	Bias	-0.3911	0.3798	-0.4003	0.4000	-0.3791	0.3735	-0.3765	0.3877
	SStd	0.1290	0.1239	0.1275	0.1281	0.0817	0.0755	0.0882	0.0864
	MStd	0.0826	0.0838	0.0819	0.0821	0.0598	0.0605	0.0600	0.0596
ALASSO	Bias	-0.0770	0.0858	-0.0821	0.0877	-0.0960	0.0787	-0.0937	0.0966
	SStd	0.1543	0.1454	0.1397	0.1246	0.0911	0.0963	0.1005	0.1053
	MStd	0.1337	0.1301	0.1316	0.1310	0.0898	0.0887	0.0892	0.0880
Oracle	Bias	-0.0965	0.0919	-0.0887	0.0930	-0.1016	0.0892	-0.1120	0.0946
	SStd	0.1500	0.1605	0.1697	0.1509	0.1131	0.1104	0.1022	0.1171
	MStd	0.1356	0.1386	0.1372	0.1380	0.0927	0.0928	0.0933	0.0930

2.5.1. We use MMSE to assess the efficiency of the proposed variable selection methods. The approximated GCV (2.31) is applied to select tuning parameter λ on a grid of points. For the approximation of approximated GCV (2.31), we choose $M_0 = 20000$.

The MMSE's with $C_r = 25\%$ and $C_r = 10\%$ based on 100 runs are listed in Tables 3.1 and 3.2 respectively. In Tables 3.1 and 3.2, we also report the average number of correctly selected zero coefficients, labeled as "correct", and the average number of coefficients erroneously shrunk to 0, labeled as "incorrect". Based on 100 simulations, Tables 3.3-3.8

Table 3.4: Summary of estimation results for nonzero effects in PH model with right censored data under $C_r = 10\%$

$C_r = 10\%$		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.0291	0.0680	-0.0541	0.0713	-0.0875	0.0631	-0.0688	0.0737
	SStd	0.1236	0.1196	0.1257	0.1272	0.0787	0.0813	0.0779	0.0787
	MStd	0.1259	0.1256	0.1278	0.1259	0.0840	0.0829	0.0835	0.0842
SCAD	Bias	-0.1088	0.0578	-0.0908	0.0622	-0.0715	0.0651	-0.0718	0.0589
	SStd	0.1284	0.1306	0.1580	0.1252	0.0823	0.0818	0.0944	0.0922
	MStd	0.1240	0.1239	0.1225	0.1230	0.0829	0.0822	0.0828	0.0841
LASSO	Bias	-0.3644	0.3527	-0.3580	0.3436	-0.3476	0.3335	-0.3313	0.3426
	SStd	0.1130	0.1087	0.1064	0.0959	0.0761	0.0724	0.0767	0.0654
	MStd	0.0828	0.0838	0.0831	0.0839	0.0590	0.0599	0.0596	0.0597
ALASSO	Bias	-0.0681	0.0732	-0.0695	0.0694	-0.0820	0.1005	-0.0874	0.0972
	SStd	0.1300	0.1298	0.1343	0.1299	0.0910	0.0890	0.0854	0.0843
	MStd	0.1298	0.1259	0.1271	0.1263	0.0818	0.0808	0.0810	0.0808
Oracle	Bias	-0.0821	0.0603	-0.0751	0.0686	-0.0608	0.0641	-0.0753	0.0910
	SStd	0.1078	0.1325	0.1239	0.1313	0.0739	0.0816	0.0834	0.0881
	MStd	0.1306	0.1282	0.1329	0.1295	0.0846	0.0857	0.0850	0.0847

give the estimated bias (Bias), sample standard deviations (SStd) and mean of estimated standard deviation (MStd) (based on the formula (2.28)) for nonzero estimates in $\hat{\beta}_n$. Note that when one covariate among Z_1, Z_3, Z_6 and Z_9 is excluded from selected models, its effect estimate and corresponding estimated standard deviation are set to be 0.

From Tables 3.1 and 3.2, the variable selection methods with SCAD, HARD and ALASSO penalties outperform the method with LASSO penalty and they also perform as well as the oracle estimate in terms of MMSE in all the settings. Moreover, all the

Table 3.5: Summary of estimation results for nonzero effects in PO model with right censored under $C_r = 25\%$

$C_r = 25\%$		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.0611	-0.0647	0.0755	-0.0571	0.0089	-0.0257	0.0398	-0.0217
	SStd	0.2738	0.2789	0.2555	0.2998	0.1492	0.1879	0.1765	0.1880
	MStd	0.2131	0.2127	0.2166	0.2111	0.1414	0.1472	0.1426	0.1468
SCAD	Bias	-0.0910	0.0354	-0.0427	0.0305	-0.0242	0.0022	0.0008	0.0219
	SStd	0.2585	0.2003	0.2392	0.2518	0.1860	0.1989	0.1102	0.1375
	MStd	0.1825	0.2010	0.1896	0.1925	0.1391	0.1434	0.1401	0.1395
LASSO	Bias	-0.4222	0.4332	-0.3847	0.4484	-0.4103	0.4386	-0.4199	0.4240
	SStd	0.1657	0.1870	0.1785	0.1805	0.1297	0.1193	0.1328	0.1467
	MStd	0.0976	0.0947	0.1026	0.0939	0.0743	0.0718	0.0729	0.0717
ALASSO	Bias	-0.0341	0.0347	-0.0255	-0.0018	-0.0054	0.0402	-0.0085	0.0560
	SStd	0.2761	0.2929	0.3005	0.2907	0.1473	0.1954	0.1860	0.1924
	MStd	0.2131	0.2127	0.2166	0.2111	0.1393	0.1367	0.1381	0.1375
Oracle	Bias	-0.0110	0.0136	-0.0424	0.0096	-0.0426	0.0126	-0.0458	0.0357
	SStd	0.2573	0.2763	0.2479	0.2899	0.2019	0.1883	0.1866	0.1753
	MStd	0.2169	0.2216	0.2239	0.2209	0.1477	0.1499	0.1481	0.1496

methods can select about the same correct number of significant covariates in all the models. In addition, we can also find that the values of MMSE in all the settings decrease as the increasing of sample size, which shows that the performance of all the methods will improve when sample size get larger.

Based on Tables 3.3 - 3.8, in all the models, the Biases based on SCAD, HARD and ALASSO penalties are as small as oracle estimates while the Biases based on LASSO are relatively far away from 0 comparing with oracle estimates. This shows that PMMLEs

Table 3.6: Summary of estimation results for nonzero effects in PO model with right censored under $C_r = 10\%$

$C_r = 10\%$		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.0084	-0.0097	0.0180	-0.0020	0.0258	0.0145	-0.0163	-0.0041
	SStd	0.2193	0.2404	0.2204	0.2066	0.1531	0.1572	0.1565	0.1412
	MStd	0.1947	0.1985	0.1924	0.2018	0.1356	0.1340	0.1351	0.1363
SCAD	Bias	0.0071	0.0055	-0.0393	0.0441	-0.0134	0.0024	-0.0087	-0.0020
	SStd	0.2510	0.2662	0.2816	0.2889	0.1675	0.1545	0.1461	0.1393
	MStd	0.1937	0.1909	0.1860	0.1850	0.1303	0.1315	0.1310	0.1343
LASSO	Bias	-0.3849	0.3813	-0.4102	0.3803	-0.3782	0.3824	-0.3808	0.3666
	SStd	0.1823	0.1705	0.1998	0.1507	0.1328	0.1202	0.1245	0.1404
	MStd	0.1020	0.1037	0.0979	0.1055	0.0755	0.0759	0.0758	0.0767
ALASSO	Bias	-0.0779	0.0856	-0.0473	0.0379	-0.0339	0.0409	-0.0228	0.0135
	SStd	0.3095	0.2870	0.2454	0.2932	0.1663	0.1731	0.1817	0.1796
	MStd	0.1711	0.1706	0.1770	0.1765	0.1259	0.1283	0.1272	0.1270
Oracle	Bias	0.0020	-0.0330	0.0176	-0.0007	0.0090	-0.0019	-0.0266	-0.0128
	SStd	0.1956	0.2091	0.2041	0.1963	0.1643	0.1388	0.1272	0.1495
	MStd	0.2041	0.2078	0.2081	0.2000	0.1429	0.1439	0.1458	0.1408

with SCAD, HARD and ALASSO penalties outperform PMMLE with LASSO penalty and they also perform as well as the oracle estimates in terms of estimation. Note that SStd is the sample standard deviation. Without considering Monte Carlo error, it can be seen as the true standard deviation. The Tables display that all the MStds for all the penalties are reasonably close to their corresponding SStd. Moreover the values of SStd, MStd and their differences decrease as the increasing of sample size. This tells us that our proposed standard deviation formula works very well for the right censored data and its

Table 3.7: Summary of estimation results for nonzero effects in GP model with right censored under $C_r = 25\%$

$C_r = 25\%$		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.0194	-0.0426	0.0343	-0.0486	0.0165	-0.0130	0.0280	-0.0279
	SStd	0.1401	0.1590	0.1472	0.1561	0.1432	0.1418	0.1364	0.1460
	MStd	0.1372	0.1409	0.1388	0.1409	0.1452	0.1465	0.1449	0.1464
SCAD	Bias	0.0404	-0.0279	0.0422	-0.0212	0.0309	-0.0183	0.0226	-0.0093
	SStd	0.1495	0.1533	0.1533	0.1538	0.0960	0.1077	0.0938	0.1005
	MStd	0.1386	0.1386	0.1385	0.1379	0.0901	0.0905	0.0901	0.0904
LASSO	Bias	-0.3951	0.4131	-0.4051	0.4154	-0.3897	0.3900	-03979	0.3858
	SStd	0.1217	0.1030	0.1204	0.1141	0.0908	0.0802	0.0753	0.0793
	MStd	0.0749	0.0740	0.0741	0.0748	0.0534	0.0537	0.0534	0.0540
ALASSO	Bias	-0.0209	-0.0023	-0.0199	0.0260	-0.0146	0.0124	-0.0222	0.0052
	SStd	0.1542	0.1517	0.1448	0.1559	0.0979	0.0975	0.0861	0.0987
	MStd	0.1245	0.1282	0.1265	0.1267	0.0843	0.0872	0.0838	0.0859
Oracle	Bias	0.0466	-0.0125	0.0148	-0.0182	0.0155	-0.0223	0.0178	-0.0174
	SStd	0.1444	0.1505	0.1455	0.1156	0.1025	0.0938	0.0928	0.1025
	MStd	0.1380	0.1385	0.1372	0.1355	0.0912	0.0917	0.0909	0.0930

performance will increase when sample size increases. In a word, our proposed procedures with SCAD, Hard thresholding and ALASSO penalties can produce satisfactory results in terms of estimation and variable selection for right censored data.

An interesting finding is that there is almost no difference about MMSE and the number of selected variables for the two censoring ratios. Moreover, Bias, MStd and SStd in all the settings are also similar too. This empirically indicates that our proposed variable selection procedures do not dependent on censoring distribution. Therefore our

Table 3.8: Summary of estimation results for nonzero effects in GP model with right censored data under $C_r = 10\%$

$C_r = 10\%$		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.0248	-0.0300	0.0353	-0.0387	0.0154	0.0019	0.0200	-0.0081
	SStd	0.1471	0.1266	0.1468	0.1360	0.0847	0.0866	0.0852	0.0918
	MStd	0.1269	0.1277	0.1274	0.1309	0.0831	0.0830	0.0825	0.0840
SCAD	Bias	0.0103	-0.0354	-0.0032	-0.0084	0.0191	-0.0064	0.0068	-0.0142
	SStd	0.1342	0.1196	0.1508	0.1291	0.0890	0.0913	0.0935	0.1052
	MStd	0.1307	0.1321	0.1285	0.1306	0.0851	0.0855	0.0847	0.0855
LASSO	Bias	-0.3649	0.3633	-0.3565	0.3622	-0.3443	0.3521	-0.3571	0.3495
	SStd	0.1129	0.1061	0.1225	0.1142	0.0771	0.0803	0.0735	0.0769
	MStd	0.0757	0.0758	0.0762	0.0765	0.0534	0.0537	0.0534	0.0540
ALASSO	Bias	0.0091	0.0123	-0.0011	0.0003	-0.0163	0.0130	-0.0079	0.0247
	SStd	0.1320	0.1462	0.1342	0.1565	0.0914	0.0915	0.0817	0.0860
	MStd	0.1182	0.1200	0.1176	0.1189	0.0797	0.0810	0.0802	0.0804
Oracle	Bias	0.0089	-0.0226	-0.0088	-0.0170	0.0031	-0.0086	-0.0010	-0.0073
	SStd	0.1146	0.1483	0.1237	0.1326	0.0801	0.1003	0.1004	0.0790
	MStd	0.1301	0.1346	0.1322	0.1330	0.0963	0.0975	0.0954	0.0961

proposed procedure may allow informative censoring.

3.4.2 Primary Biliary Cirrhosis Data Application

In this section, we apply our proposed variable selection procedures to analyze *Primary Biliary Cirrhosis (PBC) Data* , gathered in the Mayo Clinical trial in primary biliary cirrhosis of liver conducted between 1949 and 1984. A more detailed account please refer

Table 3.9: Covariates for *PBC* data and their interpretation

Covariates	Interpretation
Z_1	Treatment code, 1 means D-penicillamine and 2 placebo;
Z_2	Age of patients,in years;
Z_3	Sex, 0 means male and 1 female;
Z_4	0 means absence of ascites and 1 presence of ascites;
Z_5	0 means absence of hepatomegaly and 1 presence of hepatomegaly;
Z_6	0 means absence of spiders and 1 presence of spiders;
Z_7	0 means no edema, 0.5 untreated or successfully treated and 1 edema despite diuretic therapy;
Z_8	Serum bilirunbin (mg/dl);
Z_9	Serum cholesterol (mg/dl);
Z_{10}	Serum albumin (mg/dl);
Z_{11}	Urine copper (ug/day);
Z_{12}	Alkaline phosphotase (U/liter);
Z_{13}	Aspartate aminotransferase, once called SGOT (U/ml);
Z_{14}	Triglycerides, (mg/dl);
Z_{15}	Platelet count per cubic ml/10000;
Z_{16}	Prothrombine time, (seconds);
Z_{17}	Histologic stage of disease (needs biopsy), graded 1,2,3,4.

to Dickson *et al.* (1989) [22]. In this data set, 424 *PBC* patients' information was collected and 312 among them agreed to participate in the randomized trial. However, we only consider 276 observations without missing values in our application and 111 patients died before the end of follow-up. The information of clinical, biochemical, serological and histological measurements for each *PBC* patient were collected and they are listed in Table 3.9. Through Cox regression model, Tibshirani (1997) [68] analyzed this data set

Table 3.10: Summary of results for *PBC* data analysis by MMLE and LASSO (I)

	MMLE				LASSO			
Covariate	EST	STD	Z-value	p-value	EST	STD	Z-value	p-value
Z_1	-0.0168	0.1621	-0.1034	0.9176	–	–	–	
Z_2	0.0001	0.0000	2.7203	0.0065	0.0000	0.0000	3.9232	0.0001
Z_3	-0.3290	0.2375	-1.3853	0.1660	–			–
Z_4	0.3858	0.3617	1.0666	0.2861	0.1265	0.0306	4.1324	0.0000
Z_5	0.0321	0.1874	0.1714	0.8639	–	–	–	
Z_6	0.2903	0.1891	1.5347	0.1249	0.0000	0.0000	3.5133	0.0004
Z_7	0.7132	0.3551	2.0082	0.0446	0.5663	0.0982	5.7643	0.0000
Z_8	0.0524	0.0226	2.3230	0.0202	0.0468	0.0071	6.5807	0.0000
Z_9	0.0003	0.0004	0.9009	0.3676	–	–		–
Z_{10}	-0.3704	0.2269	-1.6325	0.1026	-0.1969	0.0411	-4.7853	0.0000
Z_{11}	0.0021	0.0010	2.1874	0.0287	0.0014	0.0003	5.2391	0.0000
Z_{12}	0.0000	0.0000	0.3523	0.7246	–			
Z_{13}	0.0036	0.0015	2.3637	0.0181	0.0000	0.0000	3.6742	0.0002
Z_{14}	-0.0005	0.0013	-0.3587	0.7198	–			
Z_{15}	0.0001	0.0009	0.1200	0.9045	–		–	
Z_{16}	0.1726	0.0877	1.9684	0.0490	0.0545	0.0125	4.3563	0.0000
Z_{17}	0.3555	0.1230	2.8904	0.0038	0.1533	0.0282	5.4443	0.0000

by the variable selection methods with LASSO and best-subset penalties. In his paper, the method with best-subset penalty selected all the significant variables except *Serum albumin* in maximum partial likelihood estimation. Although the variable selection procedure with LASSO penalty shrunk most of non-significant effects to 0, it also shrunk some significant effects to some extent. Zhang and Lu (2007) [79] considered variable selection for this data set using Cox regression models and Adaptive-Lasso penalty. They

Table 3.11: Summary of results for *PBC* data analysis by HARD and SCAD (II)

	HARD				SCAD			
Covariate	EST	STD	Z-value	p-value	EST	STD	Z-value	p-value
Z_1	–	–	–	–	–	–	–	–
Z_2	0.0001	0.0000	3.2323	0.0012	0.0001	0.0000	3.3460	0.0008
Z_3	-0.1653	0.2288	-0.7223	0.4701	–	–	–	–
Z_4	0.7873	0.3017	2.6096	0.0091	–	–	–	–
Z_5	–	–	–	–	–	–	–	–
Z_6	–	–	–	–	–	–	–	–
Z_7	–	–	–	–	0.9275	0.2974	3.1189	0.0018
Z_8	0.0692	0.0201	3.4414	0.0006	0.0601	0.0180	3.3403	0.0008
Z_9	–	–	–	–	–	–	–	–
Z_{10}	–	–	–	–	-0.4067	0.2136	-1.9037	0.0570
Z_{11}	0.0027	0.0010	2.8183	0.0048	0.0028	0.0009	3.0896	0.0020
Z_{12}	–	–	–	–	–	–	–	–
Z_{13}	0.0042	0.0015	2.8257	0.0047	0.0037	0.0015	2.5124	0.0120
Z_{14}	-0.0007	0.0012	-0.5979	0.5499	–	–	–	–
Z_{15}	–	–	–	–	–	–	–	–
Z_{16}	0.2153	0.0805	2.6760	0.0075	0.2028	0.0815	2.4877	0.0129
Z_{17}	0.4263	0.1085	3.9298	0.0001	0.3979	0.1048	3.7975	0.00010

also selected all the significant variables except *Serum albumin* in maximum partial likelihood estimation. The aim of this application is to study the dependence of survival time on the seventeen covariates listed in Table 3.9 and to select important variables through our proposed procedures using the generalized probit model given by (1.3). In the model, $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_{17})$, $\Phi(1-u, v, w) = h^{-1}(h(u)+v^T w)$ and $h^{-1}(\cdot)$ is the standard normal survival function.

Table 3.12: Summary of results for *PBC* data analysis by ALASSO (III)

Covariate	EST	STD	Z-value	p-value	Covatiate	EST	STD	Z-value	p-value
Z_1	-	-	-	-	Z_{10}	-0.3400	0.1676	-2.0285	0.0425
Z_2	0.0001	0.0000	3.1793	0.0015	Z_{11}	0.0024	0.0008	2.9752	0.0029
Z_3	-	-	-	-	Z_{12}	-	-	-	-
Z_4	-	-	-	-	Z_{13}	0.0031	0.0013	2.4576	0.0140
Z_5	-	-	-	-	Z_{14}	-	-	-	-
Z_6	-	-	-	-	Z_{15}	-	-	-	-
Z_7	0.8590	0.2713	3.1667	0.0015	Z_{16}	0.1634	0.0704	2.3227	0.0202
Z_8	0.0587	0.0166	3.5423	0.0004	Z_{17}	0.3672	0.1015	3.6162	0.0003
Z_9	-	-	-	-					

We conduct variable selection for this data set using our proposed variable selection procedures with SCAD, HARD Thresholding, LASSO and ALASSO penalties. The programme parameter setting also follows the simulation studies in Section 2.5. We use approximated GCV to select proper tuning parameter λ. We summary the results in Tables 3.10, 3.11 and 3.12 including estimate (EST), standard deviation (STD), Z-value (Ratio of estimate and corresponding standard deviation) and p-value. To compare the variable selection results, we also present rank-based maximum marginal likelihood estimate (MMLE) in the Tables.

Based on MMLE, we can see that *Age*, *Edema*, *Serum bilirunbin*, *Urine copper*, *Aspartate aminotransferase*, *Prothrombine time* and *Histologic stage of disease* have significant impacts on survival time. LASSO with $\lambda = 0.01$ shrinks most nonsignificant effects to 0, for example, *Treatment*, *Sex* etc. At the same time, it also shrinks some significant effects to 0, for example, *Age* and *SGOT*. This result is in line with the method with LASSO penalty in Tibshirani (1997) [68] and Zhang and Lu (2007) [79]. Hard thresholding with $\lambda = 0.08$ can also select all the significant variables except *edema*. While it also

select some non-significant MMLEs, they are also non-significant in selected model except *ascites*. SCAD with $\lambda = 0.53$ and ALASSO with $\lambda = 0.0024$ select all the significant MMLEs except *Serum albumin*, which agree with the best-subset results of Tibshirani (1997) [68] and the ALASSQ results of Zhang and Lu (2007) [79].

From the Tables, we can see that *treatment* and *sex* do not influence the patient's survival probability. We also can get an obvious conclusion that the older a patient is and the later stage he is at, lower survival probability he suffers from.

□ **End of chapter.**

Chapter 4

Variable Selection for General Transformation Models with Interval Censored Data

Interval censored failure time is another type of survival data. This type data often occur in medical studies, financial, epidemiological, demographical and sociological studies. Its main feature is that we only know that the failure time falls in an interval but we can not observe it exactly. A typical example of interval censored data is the human immunodeficiency virus (HIV) infection times. In this case, the determination of HIV infection time is usually based on regular blood testing. Therefore, for a patient, who was HIV negative at the beginning of this study, if we found that his blood indicates positivity at one testing, then his exact infection time was censored by an interval bracketed by the last HIV negative testing date and this HIV positive testing date. However, we can not observe his exact HIV infection time. When the right bound of the censoring interval is infinity, right censorship can be seen as a special case of interval censorship. Sun (1998) [62] and Zhang and Sun (2010) [80] summarized analysis methods of interval censored data. Sun (2006) [63] discussed the statistical inference for interval censored data systematically.

There is a large volume literature on the study of regression analysis for interval censored failure times and many efficient inference procedures have been proposed. Finkelstein and Wolfe (1985) [33] developed a semiparametric regression model and a maximum likelihood approach for the analysis of interval censored failure times; Finkelstein (1986)

[32] studied the interval censored data through proportional hazards Cox's models and maximum likelihood; Satten (1996) [58] proposed a rank-based analysis method for proportional hazard regression models with interval censored data. Huang (1996) [44] proposed an efficient estimate of parameters in proportional hazards regression models with interval censored data; Huang and Wellner (1997) [46] established asymptotic properties of the proportional hazards regression models with interval censored data. Some other regression models are also proposed for the analysis of interval censored data, for example, proportional odds models (Huang and Rossini, 1997 [45]; Sun, *et al.*, 2007 [64]), Accelerated failure time models (Rabinowitz *et al.* 1995, [57]), linear spline models (Kooperberg and Clarkson, 1997[47]), logistic regression models (Sun, 1997 [61]), simple transformation models (Younes and Lachin, 1997 [73]; Zhang *et al.*, 2005 [81]), additive hazards models (Bacchetti and Quale, 2002 [5]; Zeng *et al.*, 2006 [74]; Chen and Sun, 2010 [14]), general transformation models (Gu, *et al.*, 2005[40]) etc. Gu, *et al.* (2005)[40] proposed an efficient three-stage Monte Carlo Markov chain stochastic approximation (MCMC-SA) algorithm to find rank-based maximum marginal likelihood estimate of parameters in general transformation models with interval censored data. Although the maximum marginal likelihood estimate is very satisfactory, its large sample properties are open. This problem will be one of our interests in this Chapter.

In our knowledge, we found very few discussions on the dimension reduction or variable selection for regression model with interval censored data except for model selection considered by Sinha *et al.* (1999) [60]. Since the generality of interval censored data, we will study variable selection for general transformation models with interval censored data through rank-based penalized maximum marginal likelihood approach, in which we also consider Hard, SCAD, LASSO and ALASSO penalties.

In this Chapter, we first prove the asymptotic properties of rank-based maximum marginal likelihood estimate and then based on it, we further study the oracle properties of rank-based penalized maximum marginal likelihood estimate for unstratified interval censored data. Similar to the discussion in Section 2.4, we can easily extend the variable selection procedure to analyze stratified interval censored data. For the ease of proof, we also assume that the interval censoring is non-informative. Some simulation studies will be given to illustrate the proposed variable selection procedures.

4.1 Marginal Likelihood and Penalized Log-marginal Likelihood

Denote $T \in R^+$ as a failure time and $Z \in R^p$ as the corresponding covariate vector. Let $(L, V]$ be the censoring interval. Then we have $L < T \leq V$. Obviously, when $V = \infty$, the interval censoring reduces to right censoring, when $L = 0$, it reduces to left censoring and when $L = V$, the censoring disappears and the failure time T can be observed exactly. It is also assumed that T and the censoring interval $(L, V]$ are independent given Z, which means that

$$P(T \leq t|L = l, V = v, L < T \leq V, Z) = P(T \leq t|l < T \leq v, Z) \tag{4.1}$$

This independence assumption is to say that an interval $(L, V]$ gives no more than the information that T is simply bracketed by two observed values ([63]). This assumption has been used by Self and Grossman (1986) [59], Oller *et al.* (2004) [54], Zhang *et al.* (2005) [81] and Sun (2006)[63]. In this study, we assume that L and V have continuous cumulative density function and satisfy $P(L \leq V) = 1$.

We assume that T and Z are modeled by models (1.3). In the variable selection procedures, we also assume that $\boldsymbol{\beta}$ only contains regression parameters not any model transformation parameters. Otherwise if there are model transformation parameters in $\boldsymbol{\beta}$, such parameters should be not penalized in penalized log-marginal likelihood. Let $\boldsymbol{\beta}_0$ be the true values of $\boldsymbol{\beta}$ and we partition $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ such that $\boldsymbol{\beta}_{10}$ contains all the nonzero effects and $\boldsymbol{\beta}_{20} = \mathbf{0}$. In this chapter, we also consider the four penalty functions (1.5), (1.6), (1.7) and (1.8) defined in Section 1.3.

Suppose that$\{Z_i, (L_i, V_i]\}_{i=1}^n$ is a sample of size n from $\{Z, (T, V]\}$. Denote

$$A_i = \{j : V_i \leq L_j, j = 1, \cdots, n, j \neq i\} \tag{4.2}$$

as the set of indices of observations that must occur after the ith observations and

$$B_i = \{j : V_j \leq L_i, j = 1, \cdots, n, j \neq i\} \tag{4.3}$$

as the set of indices of observations that must occur before the ith observations. Note that if $L_i = V_j$ and $i \neq j$, the ith observation must occur after the jth observation. Denote

$$C_n = \{\mathbf{t} = (t_1, t_2, \cdots, t_n)' : t_j < t_i, j \in B_i, \quad t_i < t_k, k \in A_i, i = 1, 2, \cdots, n\}$$

as the set of times consistent with the order restrictions in the observed data. Let $\mathcal{R}_n = (r_1, r_2, \cdots, r_n)^T$ be the complete ranking of underlying failure times T_i's. It is obvious that $r_i < r_j$ if $j \in A_i$ and $r_i > r_j$ if $j \in B_i$.

Denote $\mathbf{T}_n = (T_1, T_2, \cdots, T_n)$, $\mathbf{Z}_n = (Z_1, Z_2, \cdots, Z_n)^T$ and $\mathcal{S}_n$ as a ranking set that consists of all possible ranking $\mathcal{R}_n$ of $\mathbf{T}_n$. Then the rank-based marginal likelihood based on interval censored data is given by

$$\begin{aligned} L_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n) &= P(\mathcal{R}_n \in \mathbf{S}_{\tilde{n}}|\mathbf{Z}_n) = P(\mathbf{T}_n \in C_n|\mathbf{Z}_n) \\ &= (-1)^n \int_{C_n} \prod_{i=1}^{n} \phi(S_0(t_i), Z_i, \boldsymbol{\beta}) \prod_{i=1}^{n} dS_0(t_i) \\ &= \int_{\mathcal{D}_n} \prod_{i=1}^{n} \phi(1-u_i, Z_i, \boldsymbol{\beta}) \prod_{i=1}^{n} du_i \end{aligned} \tag{4.4}$$

where $\phi(u, v, w)$ is the same one in (3.1); the third equality holds because the simple transformation $u_i = 1 - S_0(t_i)$; $\mathcal{D}_n$ is the corresponding collection of uniform (0,1) vectors consistent with the order restrictions in C_n, namely

$$\mathcal{D}_n = \{\mathbf{u}_n = (u_1, u_2, \cdots, u_n)' : u_j < u_i, j \in B_i, \quad u_i < u_k, k \in A_i, i = 1, 2, \cdots, n\}.$$

Denote $U_i = 1 - S_0(T_i)$, $\phi_3(u, v, w) = \frac{\partial \phi(u,v,w)}{\partial w}$, $\psi(u, v, w) = \frac{\phi_3(u,v,w)}{\phi(u,v,w)}$, $\ell(\boldsymbol{\beta}) = \log(L_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n))$ and $\mathscr{F}_n$ as the smallest σ-algebra generated by $\mathcal{S}_n$ and $\mathbf{Z}_n$. Differentiating (4.4) with respect to $\boldsymbol{\beta}$, we can easily get the score function as follows

$$\begin{aligned} S_n(\boldsymbol{\beta}) &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \int_{\mathcal{D}_n} \sum_{i=1}^{n} \psi(1-u_i, Z_i, \boldsymbol{\beta}) p(\mathbf{u}_n; \mathcal{S}_n, \mathbf{Z}_n, \boldsymbol{\beta}) \prod_{i=1}^{n} du_i \\ &= \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\beta}}\{\psi(1-U_i, Z_i, \boldsymbol{\beta})|\mathscr{F}_n\} \end{aligned} \tag{4.5}$$

where

$$p(\mathbf{u}_n; \boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n) = \frac{I(\mathbf{u}_n \in \mathcal{D}_n)}{L_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n)} \prod_{i=1}^{n} \phi(1-u_i, Z_i, \boldsymbol{\beta}). \tag{4.6}$$

is the conditional density of $\mathbf{U}_n = (U_1, U_2, \cdots, U_n)^T$ given $\mathscr{F}_n$. Then we can obtain the rank-based maximum marginal likelihood estimate (MMLE) of $\boldsymbol{\beta}$, denoted by $\tilde{\boldsymbol{\beta}}_n$, by solving

$$S_n(\boldsymbol{\beta}) = 0. \tag{4.7}$$

Gu *et al* (2005) [40] proposed a three-stage Monte Carlo Markov Chain stochastic approximation (MCMC-SA) algorithm to find $\tilde{\boldsymbol{\beta}}_n$ and corresponding covariance matrix. Under some regular conditions, we will prove large sample properties in next Section.

Based on (4.4), we have the rank-based penalized marginal likelihood function as follows

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{i=1}^{p} p_\lambda(|\beta_i|) \tag{4.8}$$

where $p_\lambda(\cdot)$ is a penalty function, which should be irregular at original for the variable selection procedure (Fan and Li, 2002 [25]); λ is a penalty tuning parameter. The penalized maximum marginal likelihood estimate (PMMLE), denoted by $\hat{\boldsymbol{\beta}}_n$, of $\boldsymbol{\beta}$ can be obtained by maximizing the penalized log-marginal likelihood function (4.8) with respect to $\boldsymbol{\beta}$. With a proper penalty function $p_\lambda(\cdot)$, some components of $\hat{\boldsymbol{\beta}}_n$ will be zero and they will not appear in selected models, which achieves the purpose of variable selection for the model (1.3) with interval censored data.

Remark 4.1 From (4.4), the marginal likelihood function only depends on the ranking of underlying survival times instead of "nuisance" parameter $S_0(t)$ and censoring distribution. So the rank-based maximum marginal likelihood estimate and penalized maximum marginal likelihood estimate are also baseline-free and of censoring-distribution-free.

Remark 4.2 Denote $\Phi_3(u,v,w) = \partial\Phi(u,v,w)/\partial w$ and define

$$f(u_1,u_2,v,w) = \begin{cases} \frac{\Phi_3(1-u_1,v,w)-\Phi_3(1-u_2,v,w)}{\Phi(1-u_1,v,w)-\Phi(1-u_2,v,w)} & u_1 \neq u_2 \\ \psi(u,v,w) & u_1 = u_2 = u \end{cases}$$

$$U_i^- = \begin{cases} \max\{U_j: \quad j \in B_i\} & B_i \neq \varnothing \\ 0 & B_i = \varnothing \end{cases}$$

and

$$U_i^+ = \begin{cases} \min\{U_j: \quad j \in A_i\} & A_i \neq \varnothing \\ 1 & A_i = \varnothing \end{cases}$$

then we have

$$\begin{aligned}
&\mathrm{E}[\psi(1-U_i,Z_i,\boldsymbol{\beta})|\ U_i^- < U_i < U_i^+, Z_i] \\
=&\frac{\int_{U_i^-}^{U_i^+}\psi(1-u,Z_i,\boldsymbol{\beta})\phi(1-u,Z_i,\boldsymbol{\beta})du}{\int_{U_i^-}^{U_i^+}\phi(1-u,Z_i,\boldsymbol{\beta})du} \\
=&\frac{\Phi_3(1-U_i^-,Z_i,\boldsymbol{\beta})-\Phi_3(1-U_i^+,Z_i,\boldsymbol{\beta})}{\Phi(1-U_i^-,Z_i,\boldsymbol{\beta})-\Phi(1-U_i^+,Z_i,\boldsymbol{\beta})} \\
=&f(U_i^-,U_i^+,Z_i,\boldsymbol{\beta}),
\end{aligned} \tag{4.9}$$

Therefore the score function $S_n(\boldsymbol{\beta})$ can be expressed as follows,

$$\begin{aligned}
S_n(\boldsymbol{\beta}) =& \sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{\beta}}[\psi(1-U_i,Z_i,\boldsymbol{\beta})|\mathscr{F}_n] \\
=& \sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{\beta}}\left\{\mathrm{E}[\psi(1-U_i,Z_i,\boldsymbol{\beta})|U_i^- < U_i < U_i^+, Z_i]|\mathscr{F}_n\right\} \\
=& \sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{\beta}}\left[f(U_i^-,U_i^+,Z_i,\boldsymbol{\beta})|\mathscr{F}_n\right]
\end{aligned} \tag{4.10}$$

4.2 Consistency and Oracle Properties

In this section, we first study the asymptotic properties of rank-based maximum marginal likelihood estimate, denoted by $\widetilde{\boldsymbol{\beta}}_n$, and then we further explore the consistency and oracle properties of rank-based penalized maximum marginal likelihood estimates, $\hat{\boldsymbol{\beta}}_n$. Before discussing them, let's give some regular conditions.

(A0) Given covariate vector Z, T and the censoring interval $(L, V]$ are independent (or the interval censoring is non-informative), that is, the equality (4.1) holds.

(A1) Both the censoring interval bounds L and V are continuous random variables with support on $[0,\infty)$ satisfying

$$P(L<V)=1 \ \text{ and } \ P(|V-L|<\varepsilon \ \mid \ L<T\le V, Z)>0 \ \text{ for any } \ \varepsilon>0.$$

(A2) Suppose that $\boldsymbol{\beta}\in\Theta$, a compact subset of the Euclidean space R^p and the true value $\boldsymbol{\beta}_0$ is an interior of Θ. The covariate vector Z is exogenous, and assumed to be bounded, equivalently, there exists a constant $M_1>0$, such that $P(||Z||\le M_1)=1$. And for all $u\in(0,1)$, $||v||\le M_1$ and $w\in\Theta$.

$$\phi_3(u,v,w)=\frac{\partial\phi(u,v,w)}{\partial w} \quad \phi_{33}(u,v,w)=\frac{\partial^2\phi(u,v,w)}{\partial w\partial w^T}$$

exist and are continuous with respect to $w \in \Theta$. The condition holds with $\phi(u, v, w)$ replaced by $\Phi(u, v, w)$

(A3) For any v satisfying $||v|| \leq M_1$, there are functions $F_1(u, v)$ and $F_2(u, v)$, integrable with respect to u over $(0, 1)$, such that

$$||\phi_3(u, v, w)|| < F_1(u, v), \quad ||\phi_{33}(u, v, w)|| < F_2(u, v), \quad \text{for all} \quad w \in \Theta.$$

This condition holds when $\phi(u, v, w)$ is replaced by $\Phi(u, v, w)$.

(A4) Denote $U = 1 - S_0(T)$, for any $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, an neighborhood of $\boldsymbol{\beta}_0$ in Θ, $\mathrm{E}_{\boldsymbol{\beta}_0}[\psi(1 - U, Z, \boldsymbol{\beta})]$, $\mathrm{E}_{\boldsymbol{\beta}_0}[\psi^2(1 - U, Z, \boldsymbol{\beta})]$ and $\mathrm{E}_{\boldsymbol{\beta}_0}[-\frac{\partial \psi(1-U,Z,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}]$ exist.

(A5) Denote U_1 and U_2 as any two copies of a sample from U. For any $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, $\mathrm{E}_{\boldsymbol{\beta}_0}[f(U_1, U_2, Z, \boldsymbol{\beta})]$, $\mathrm{E}_{\boldsymbol{\beta}_0}[f^2(U_1, U_2, Z, \boldsymbol{\beta})]$ and $\mathrm{E}_{\boldsymbol{\beta}_0}\left[-\frac{\partial}{\partial \boldsymbol{\beta}^T} f(U_1, U_2, Z, \boldsymbol{\beta})\right]$ exist. The assumption also holds when U_1 and U_2 are replaced by $F_0(L)$ and $F_0(V)$ respectively. Denote

$$V_1(\boldsymbol{\beta}) = \mathrm{Var}_{\boldsymbol{\beta}_0}\left\{f(F_0(L), F_0(V), Z_i, \boldsymbol{\beta})\right\} \tag{4.11}$$

and it is assumed that $V_1(\boldsymbol{\beta}_0)$ is positive definite.

Remark 4.3 Define

$$V_2(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{\beta}_0}\left\{-\frac{\partial}{\partial \boldsymbol{\beta}^T} f(F_0(L), F_0(V), Z, \boldsymbol{\beta})\right\}. \tag{4.12}$$

Then under the conditions (A2)-(A5), it is easily to show that $V_1(\boldsymbol{\beta}_0) = V_2(\boldsymbol{\beta}_0)$.

(A6) There exists a constant M_2 such that

$$\max\left\{\mathrm{E}_{\boldsymbol{\beta}_0}[||\psi(1 - U, Z, \boldsymbol{\beta}_0)||^2], \ \mathrm{E}_{\boldsymbol{\beta}_0}[||f(U_1, U_2, Z, \boldsymbol{\beta}_0)||^2], \ \mathrm{E}_{\boldsymbol{\beta}_0}[||f(F_0(L), F_0(V), Z, \boldsymbol{\beta}_0)||^2]\right\} < M_2$$

(A7) The function $\psi(u, v, w)$ is continuous for $u \in (0, 1)$ and satisfies Lipschitz condition with respect to u in any closed subset of (0,1). That is, for any $[\mathcal{L}, \mathcal{R}] \subset (0, 1)$, there exists a constant $M_3(\mathcal{L}, \mathcal{R})$, dependent on $[\mathcal{L}, \mathcal{R}]$, such that for all $u_1, u_2 \in [\mathcal{L}, \mathcal{R}]$, $|v| \leq M_1$ and $w \in \mathcal{O}_{\boldsymbol{\beta}_0}$,

$$||\psi(u_1, v, w) - \psi(u_2, v, w)|| \leq M_3|u_1 - u_2|.$$

This condition also holds when $\psi(u, v, w)$ is replaced by $f(u_3, u_4, v, w)$ with respect to one in $[\mathcal{L}, \mathcal{R}]$ and another fixed in (0,1) among u_3 and u_4 $(u_3 \neq u_4)$.

(A8) The function $\frac{\partial\psi(u,v,w)}{\partial w^T}$ is continuous for $u \in (0,1)$ and satisfies Lipschitz condition with respect to u. That is, there exists a constant M_4 such that for all $u_1, u_2 \in (0,1)$, $|v| \leq M_1$ and $w \in \mathcal{O}_{\beta_0}$,

$$\left\| \frac{\partial\psi(u_1,v,w)}{\partial w^T} - \frac{\partial\psi(u_2,v,w)}{\partial w^T} \right\| \leq M_4|u_1 - u_2|.$$

This condition also holds when $\psi(u,v,w)$ is replaced by $f(u_3,u_4,v,w)$ with respect to one in $[\mathcal{L},\mathcal{R}]$ and another fixed in (0,1) among u_3 and u_4 ($u_3 \neq u_4$).

(A9) For any discretization (refer to next Section), there exists N such that when $n > N$,

$$\mathrm{E}_{\beta_0}\left[-\frac{\partial}{\partial\beta^T}S_{\acute{n}}(\beta)\right] \geq \mathrm{E}_{\beta_0}\left[-\frac{\partial}{\partial\beta^T}S_{n,m}(\beta)\right]$$

where $S_{n,m}(\beta)$ is given by (4.21) in the next Section.

Conditions (A0) and (A1) are the restrictions for the interval censored data considered in this Chapter. (A2) is the regular condition for the models (1.3) while (A3) allows the interchangeability of order for differentiation and integration or sum. Conditions (A4)-(A9) are used to prove large sample properties of maximum marginal likelihood estimate and penalized maximum marginal likelihood estimate. The inequality

$$\mathrm{E}_{\beta_0}\left[-\frac{\partial}{\partial\beta}S_n(\beta)\right] \geq \mathrm{E}_{\beta_0}\left[-\frac{\partial}{\partial\beta}S_{n,m}(\beta)\right]$$

should hold for $\beta = \beta_0$, otherwise, use of discretized data is better than use of original data.

4.2.1 Asymptotic Properties of MMLE

Theorem 4.1 *Under conditions (A0)-(A9), there exists a root $\widetilde{\beta}_n$ of $S_n(\beta)$ such that*

(i) (Consistency)

$$\widetilde{\beta}_n \xrightarrow{P} \beta_0 \quad as \quad n \to \infty;$$

(ii) (Asymptotic Normality and Efficiency)

$$\sqrt{n}\left(\widetilde{\beta}_n - \beta_0\right) \xrightarrow{D} N(0, \mathcal{V}^{-1}),$$

where $\mathcal{V} = V_1(\beta_0) = V_2(\beta_0)$.

This theorem will be proved in Section 4.2.4.

4.2.2 Results of Variable Selection Methods with Non-concave Penalties

Denote

$$a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|): \quad \beta_{j0} \neq 0\}, \quad b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)|: \quad \beta_{j0} \neq 0\},$$

$$\Sigma_{\lambda_n}(\boldsymbol{\beta}_1) = \text{diag}\{p''_{\lambda_n}(|\beta_1|), p''_{\lambda_n}(|\beta_2|), \cdots, p''_{\lambda_n}(|\beta_s|)\}$$

and

$$\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_1) = (p'_{\lambda_n}(|\beta_1|)\text{sign}(\beta_1), p'_{\lambda_n}(|\beta_2|)\text{sign}(\beta_2), \cdots, p'_{\lambda_n}(|\beta_s|)\text{sign}(\beta_s))^T,$$

where $p_\lambda(\cdot)$ is one of (1.5), (1.7) and (1.8). Then we have,

Theorem 4.2 (Consistency) *Under conditions (A0)-(A9), if $b_n \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ of $Q(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$.*

Remark 4.4 From this theorem, we can see that with a proper tuning parameter λ_n, as long as $a_n = O(n^{-1/2})$, there exists a $\sqrt{n}$-consistent rank-based penalized maximum marginal likelihood estimate.

Theorem 4.3 (Oracle property) *Assume that the penalty function $p_{\lambda_n}(\cdot)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\beta\to 0+} p'_{\lambda_n}(\beta)/\lambda_n > 0.$$

If $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem 4.2, with probability tending to 1, the $\sqrt{n}$-consistent local maximizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^T, \hat{\boldsymbol{\beta}}_{2n}^T)^T$ in Theorem 4.2 must satisfy:

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$;

(ii) (Asymptotic normality)

$$\sqrt{n}(\mathcal{V} + \Sigma_{\lambda_n}(\boldsymbol{\beta}_{10}))\{\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10} + (\mathcal{V} + \Sigma_{\lambda_n}(\boldsymbol{\beta}_{10}))^{-1}\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_{10})\} \to N(\mathbf{0}, \mathcal{V}) \qquad (4.13)$$

where $\mathcal{V}$ is the upper leading $s \times s$ submatrix of $V_2(\boldsymbol{\beta}_0)$ defined by (4.12).

Remark 4.5 Note that with Hard and SCAD penalties, if $\lambda_n \to 0$, then for a sufficiently large n, $a_n = 0$, $\Sigma_{\lambda_n}(\boldsymbol{\beta}_{10}) = \mathbf{0}$ and $\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_{10}) = \mathbf{0}$. Therefore, when $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \to N(\mathbf{0}, \mathcal{V}^{-1}) \text{ and } \hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}.$$

That is, with proper tuning parameter, PMMLEs with HARD and SCAD penalties enjoy the oracle property and they perform very well as if we have known $\boldsymbol{\beta}_2 = \mathbf{0}$ in advance when we estimate $\boldsymbol{\beta}_1$. However, for LASSO penalty, $a_n = \lambda_n$. The conditions $\lambda_n = O(n^{-1/2})$ and $\sqrt{n}\lambda_n \to \infty$ in Theorem 4.3 will contradict. So PMMLE with LASSO penalty can not enjoy oracle properties.

4.2.3 Results of Variable Selection Method with Adaptive-LASSO Penalty

Theorem 4.4 (Consistency) *Under the conditions (A0)-(A9), if* $\sqrt{n}\lambda_n = O(1)$, *then there exists a local maximizer* $\hat{\boldsymbol{\beta}}_n$ *of* $Q(\boldsymbol{\beta})$ *with ALASSO penalty such that* $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$.

This theorem tells us that with proper λ_n, the PMMLEs $\hat{\boldsymbol{\beta}}_n$ with Adaptive-LASSO penalty is $\sqrt{n}$-consistent.

Theorem 4.5 (Oracle) *If* $\sqrt{n}\lambda_n \to 0$ *and* $n^{(1+\gamma)/2}\lambda_n \to \infty$ *for some* $\gamma > 0$, *then under conditions of Theorem 4.4, with probability tending to 1, the local maximizer* $\boldsymbol{\beta}_n$ *with ALASSO penalty must satisfy:*

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$;

(ii) (Asymptotic Normality)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(\mathbf{0}, \mathcal{V}^{-1}) \quad as \quad n \to \infty, \tag{4.14}$$

where $\mathcal{V}$ *is the same as one in Theorem 4.2.*

Based on Theorem 4.5, we can see that, with proper tuning parameter λ_n, PMMLEs with Adaptive-LASSO penalty also enjoys oracle properties as if we have known which effects are equal to 0 in advance when we estimate $\boldsymbol{\beta}_1$.

4.2.4 Proof of Theorems

In this Section, we give the proofs of Theorems in Sections 4.2.1-4.2.3 by discretization technique. Without loss of generality, we assume that the baseline cumulative density function $F_0(t) = 0$ for $t < 0$ and is strictly increasing in $(0, \infty)$. Firstly, we describe the discretization technique for interval censored data.

Based on the condition (A6), there exist two positive real numbers $\mathcal{L}$ and $\mathcal{R}$ such that

$$\begin{aligned}
&\mathrm{E}_{\beta_0}[||\psi(1-U,Z,\beta_0)||^2 \cdot I(U<\mathcal{L})] < \varepsilon/9,\\
&\mathrm{E}_{\beta_0}[||\psi(1-U,Z,\beta_0)||^2 \cdot I(U>\mathcal{R})] < \varepsilon/9,\\
&\mathrm{E}_{\beta_0}[||f(F_0(L),F_0(V),Z,\beta_0)||^2 \cdot I(V<\mathcal{L})] < \varepsilon/9,\\
&\mathrm{E}_{\beta_0}[||f(F_0(L),F_0(V),Z,\beta_0)||^2 \cdot I(L>\mathcal{R})] < \varepsilon/9.
\end{aligned} \tag{4.15}$$

For above ε, $\mathcal{L}$ and $\mathcal{R}$, there exists a positive integer number m_0 such that when $m > m_0$, we can have a partition of $[0,\infty)$: $0 = t_0 < t_1 < t_2, \cdots, t_m < t_{m+1} = \infty$ such that there exist two numbers l and r satisfying

$$F_0^{-1}(\mathcal{L}) = t_l \quad F_0^{-1}(\mathcal{R}) = t_r$$

and

$$\max_{0\le j<m} |F_0(t_{j+1}) - F_0(t_j)| < \frac{\sqrt{\varepsilon}}{3M_3}. \tag{4.16}$$

Define

$$T^{(m)} = t_j \quad \text{if} \quad t_j \le T < t_{j+1}, \quad j = 0,1,\cdots,m.$$

Then we get discretized versions of T, L and V with the common distribution support at $t_0, t_1, \cdots, t_m$. We denote the discretized versions of L and V as $L^{(m)}$ and $V^{(m)}$. Then $T^{(m)}$, $L^{(m)}$ and $V^{(m)}$ satisfy

$$L^{(m)} \le T^{(m)} \le V^{(m)}. \tag{4.17}$$

Define $U^{(m)} = F_0(T^{(m)})$, then we can also obtain the discretized version of $U = F_0(T)$ with the distribution support at $u_0, u_1, \cdots, u_m$, where $u_i = F_0(t_i)$. At the same time, we also get the following conclusions based on the partition:

$$\begin{aligned}
&|U - U^{(m)}| = |F_0(T) - F_0(T^{(m)})| \le \max_{0\le j\le m}\{|F_0(t_j) - F_0(t_{j+1})|\} < \frac{\sqrt{\varepsilon}}{3M_3},\\
&|F_0(L) - F_0(L^{(m)})| \le \max_{0\le j\le m}\{|F_0(t_j) - F_0(t_{j+1})|\} < \frac{\sqrt{\varepsilon}}{3M_3},\\
&|F_0(V) - F_0(V^{(m)})| \le \max_{0\le j\le m}\{|F_0(t_j) - F_0(t_{j+1})|\} < \frac{\sqrt{\varepsilon}}{3M_3}.
\end{aligned} \tag{4.18}$$

The probability mass function of $U^{(m)}$ can be expressed as

$$\phi^*(1-u_j,Z,\boldsymbol{\beta}) = P(U^{(m)} = u_j|Z) = \Phi(1-u_j,Z,\boldsymbol{\beta}) - \Phi(1-u_{j+1},Z,\boldsymbol{\beta}).$$

Suppose $\{(L_i, V_i]\}_{i=1}^n$ are n i.d.d. copies of $(L, V]$, then $\{[L_i^{(m)}, V_i^{(m)}]\}_{i=1}^n$ are also n i.d.d. copies of $[L^{(m)}, V^{(m)}]$.

Let $\mathcal{R}_{m,n} = (r_1^{(m)}, r_2^{(m)}, \cdots, r_n^{(m)})^T$ $(1 \le r_i^{(m)} \le \min(m, n))$ be the complete ranking of underlying $T_i^{(m)}$'s. From the form of the discretized censoring intervals (4.17), the ranking of $T_i^{(m)}$'s have the following properties:

(1) If both L_i and V_i are in the same discretized interval $[t_k, t_{k+1})$, then $T_i^{(m)} = L_i^{(m)} = V_i^{(m)} = t_k$. Hence the discretized failure time $T_i^{(m)}$ is observed exactly at t_k;

(2) If both the censoring intervals $(L_i, V_i]$ and $(L_j, V_j]$ $(i \neq j)$ fall in the same discretized interval $[t_k, t_{k+1})$, then $L_i^{(m)} = V_i^{(m)} = L_j^{(m)} = V_j^{(m)} = t_k$, that is, both $T_i^{(m)}$ and $T_j^{(m)}$ occur exactly at t_k, which means $r_i^{(m)} = r_j^{(m)}$.

(3) If L_i and V_j $(V_j \le L_i$, that is, $r_j < r_i)$ fall in $[t_k, t_{k+1})$, then $L_i^{(m)} = V_j^{(m)} = t_k$. So $r_j^{(m)} \le r_i^{(m)}$;

(4) If L_i and V_j $(V_j < L_i$, that is, $r_j < r_i)$ fall in different discretized intervals, then $r_j^{(m)} < r_i^{(m)}$, that is, their relative ranking will be the same in both $\mathcal{R}_n$ and $\mathcal{R}_{n,m}$.

Based on above properties, when $n > m$, some discretized failure times may be observed exactly and some discretized failure times may tie with each other. When some discretized failure times are tied, we will assign them the same rank.

Similar to (4.2) and (4.3), define $A_i^{(m)} = \{j : V_i^{(m)} \le L_j^{(m)}, j = 1, \cdots, n, j \neq i\}$ and $B_i^{(m)} = \{j : V_j^{(m)} \le L_i^{(m)}, j = 1, \cdots, n, j \neq i\}$. If $k \in A_i^{(m)}$,the discretized failure time $T_k^{(m)}$ occur after the occurrence of $T_i^{(m)}$ and at most they occur simultaneously at their common bounds of their discretized censoring intervals. Similarly, if $k \in B_i^{(m)}$, $T_k^{(m)}$ fails before the occurrence of $T_i^{(m)}$ and at most they occur simultaneously at their common bounds of their discretized censoring intervals. Denote

$$U_i^{(m)-} = \begin{cases} \max\{U_j^{(m)} : \quad j \in B_i^{(m)}\} & B_i^{(m)} \neq \varnothing \\ 0 & B_i^{(m)} = \varnothing \end{cases}$$

and

$$U_i^{(m)+} = \begin{cases} \min\{U_j^{(m)} : \quad j \in A_i^{(m)}\} & A_i^{(m)} \neq \varnothing \\ 1 & A_i^{(m)} = \varnothing, \end{cases}$$

Consequently, if $T_i^{(m)}$ can be observed exactly, we have $U_i^{(m)-} = U_i^{(m)+} = F_0(L_i^{(m)}) = F_0(V_i^{(m)})$.

Define the event $\mathcal{E}_n$ as

$$\begin{aligned}\mathcal{E}_n = \{&\text{For any } \ i, \ \ i = 0, 1, \cdots, m, \ \ \text{there exists at least one of } T_j^{(m)}\text{'s} \\ &(1 \le j \le n) \ \text{ observed exactly at } t_i, \text{ equivalently, } \ T_j^{(m)} = t_i\}.\end{aligned} \tag{4.19}$$

Following the condition (A1), for fixed m, when n is large enough, the probability of $\mathcal{E}_n$ will be 1. Specifically,

Let $p_j = P(t_j \le L < T \le V < t_{j+1})$ and $p = \min_{0 \le j \le m} p_j$. Then following the condition (A1), $p_j > 0$ and hence $0 < p < 1$. Denote e_j as the event that there is no observed death at t_j for discretized failure times, equivalently, that there is no censoring interval $(L, V]$ falling in $[t_j, t_{j+1})$. Consequently, $P(e_j) = (1 - p_j)^n$. Note that $\mathcal{E}_n^c = \cup_{1 \le j \le m} e_j$. Then we have that

$$P(\mathcal{E}_n^c) \le \sum_{i=1}^{m} P(e_j) = \sum_{i=1}^{m} (1 - p_j)^n \le m(1 - p)^n = m \exp\{-n\delta\}, \tag{4.20}$$

where $\delta = \ln \frac{1}{1-p}$. Thus $P(\mathcal{E}_n^c)$ convergences to 0 with the rate $O(e^{-n})$ as n goes to infinity. Therefore, for fixed m, when the event $\mathcal{E}_n$ occurs, knowing the ranking of observed discretized failure times is equivalent to the knowledge of their exact discretized failure times. For the interval censored discretized failure time $T_i^{(m)}$, there must exist two positive integers i_1 and i_2 ($1 \le i_1, i_2 \le n$ and $i_1, i_2 \ne i$) such that $L_i^{(m)} = T_{i_1}^{(m)}$ and $V_i^{(m)} = T_{i_2}^{(m)}$ since there is at least one discretized failure time occurring at any one point among $\{t_i\}_{i=0}^m$, that is, the rank of the discretized interval censored failure time satisfies that $r_{i_1}^{(m)} \le r_i^{(m)} \le r_{i2}^{(m)}$. Hence $U_i^{(m)-} = F_0(T_{i_1}^{(m)})$ and $U_i^{(m)+} = F_0(V_{i_2}^{(m)})$ under the occurrence of $\mathcal{E}_n$.

Denote $\mathcal{S}_{n,m}$ as a set that consists of all possible rankings $\mathcal{R}_{n,m}$ consistent with $\{[L_i^{(m)}, V_i^{(m)}]\}_{i=1}^n$ and $\mathscr{F}_{n,m}$ as the smallest σ-algebra generated by $\mathcal{S}_{n,m}$ and $\mathbf{Z}_n$. Similar to $S_n(\boldsymbol{\beta})$, the score function based on discretized failure times is given by

$$S_{n,m}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\beta}}[\psi^*(1 - U_i^{(m)}, Z_i, \boldsymbol{\beta}) | \mathscr{F}_{n,m}], \tag{4.21}$$

where $\psi^*(u, v, w) = \frac{\partial \phi^*(u,v,w)/\partial w}{\phi^*(u,v,w)}$.

Under the condition (A3), for the true value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$, we have

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}[\psi(1-U,Z,\boldsymbol{\beta}_0)] \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\mathrm{E}[\psi(1-U,Z,\boldsymbol{\beta}_0)|Z]\right\} \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\int_0^1 \psi(1-u,Z,\boldsymbol{\beta}_0)\phi(1-u,Z,\boldsymbol{\beta}_0)du\right\} \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\int_0^1 \phi_3(1-u,Z,\boldsymbol{\beta}_0)du\right\} \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{\partial}{\partial\boldsymbol{\beta}}\int_0^1 \phi(1-u,Z,\boldsymbol{\beta})du\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\} \\
=&0
\end{aligned} \tag{4.22}$$

and

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}[\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)] \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\mathrm{E}[\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)|Z]\right\} \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\sum_{i=0}^{m}[\psi^*(1-u_i,Z,\boldsymbol{\beta}_0)\phi^*(1-u_i,Z,\boldsymbol{\beta}_0)]\right\} \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\sum_{i=0}^{m}[\Phi_3(1-u_i,Z,\boldsymbol{\beta}_0)-\Phi_3(1-u_{i+1},Z,\boldsymbol{\beta}_0)]\right\} \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i=0}^{m}[\Phi(1-u_i,Z,\boldsymbol{\beta}_0)-\Phi(1-u_{i+1},Z,\boldsymbol{\beta})]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\} \\
=&0.
\end{aligned} \tag{4.23}$$

Thus both $\{\psi(1-U_i,Z_i,\boldsymbol{\beta}_0)\}_{i=1}^n$ and $\{\psi^*(1-U_i^{(m)},Z_i,\boldsymbol{\beta}_0)\}_{i=1}^n$ are i.i.d. random variables with mean being 0.

Consequently, we can conclude that $\mathrm{E}_{\boldsymbol{\beta}_0}[S_n(\boldsymbol{\beta}_0)]=0$ and $\mathrm{E}_{\boldsymbol{\beta}_0}[S_{n,m}(\boldsymbol{\beta}_0)]=0$, which means that we can get the rank-based maximum marginal likelihood estimate of $\boldsymbol{\beta}$ by solving $S_n(\boldsymbol{\beta})=0$ or $S_{n,m}(\boldsymbol{\beta})=0$. In this Chapter, we denote $\tilde{\boldsymbol{\beta}}_n$ as the rank-based maximum marginal likelihood estimate of $\boldsymbol{\beta}$. We can prove that $\tilde{\boldsymbol{\beta}}_n$ is consistent and distributed asymptotically normally later.

Denote $\delta_i = I(T_i^{(m)-} = T_i^{(m)+})$ as the indicator whether the discretized failure time of ith individual was observed exactly and $\delta = I(L^{(m)} = V^{(m)})$ as the indicator whether the discretized failure time $T^{(m)}$ can be observed exactly. Similar to (4.10), $S_{n,m}(\boldsymbol{\beta})$ can be

also rewritten as

$$\begin{aligned} & S_{n,m}(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\beta}} \left[\psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta})\delta_i + f(U_i^{(m)-}, \triangle U_i^{(m)+}, Z_i, \boldsymbol{\beta})(1-\delta_i) | \mathscr{F}_{n,m} \right] \end{aligned} \tag{4.24}$$

where $\triangle U^{(m)}$ for a discretized variable $U^{(m)}$ means that $\triangle U^{(m)} = u_{j+1}$ if $U^{(m)} = u_j$. Under the occurrence of $\mathcal{E}_n$, $S_{n,m}(\boldsymbol{\beta})$ will become a sum of i.i.d. random variables.

Denote

$$\bar{S}_{n,m}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta})\delta_i + f(F_0(L_i^{(m)}), \triangle F_0(V_i^{(m)}), Z_i, \boldsymbol{\beta})(1-\delta_i) \right]$$

$$\widetilde{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} f(F_0(L_i), F_0(V_i), Z_i, \boldsymbol{\beta}).$$

Next, we introduce some Lemmas based on above discretization technique.

Lemma 4.1 If the assumptions (A0)-(A7) hold, then for any $\varepsilon > 0$, there exists a constant m_0 independent of n such that for any $m > m_0$

$$\mathrm{E}_{\boldsymbol{\beta}_0} \left[\frac{1}{n} ||S_n(\boldsymbol{\beta}_0) - S_{n,m}(\boldsymbol{\beta}_0)||^2 \right] < \varepsilon \tag{4.25}$$

uniformly for all n.

Proof: Following (4.22) and (4.23), for any i ($i = 1, 2, \cdots, n$),

$$\mathrm{E}_{\boldsymbol{\beta}_0} \left[\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) \right] = 0 \quad \text{and} \quad \mathrm{E}_{\boldsymbol{\beta}_0} \left[\psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0) \right] = 0.$$

And then we have

$$\begin{aligned} & \frac{1}{n} \mathrm{E}_{\boldsymbol{\beta}_0} \left[||S_n(\boldsymbol{\beta}_0) - S_{n,m}(\boldsymbol{\beta}_0)||^2 \right] \\ = \; & \frac{1}{n} \mathrm{E}_{\boldsymbol{\beta}_0} \left\{ \left\| \mathrm{E}_{\boldsymbol{\beta}_0} \left[\sum_{i=1}^{n} (\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)) \middle| \mathscr{F}_n \vee \mathscr{F}_{n,m} \right] \right\|^2 \right\} \\ \leq \; & \frac{1}{n} \mathrm{E}_{\boldsymbol{\beta}_0} \left\{ \mathrm{E}_{\boldsymbol{\beta}_0} \left[\left\| \sum_{i=1}^{n} (\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)) \right\|^2 \middle| \mathscr{F}_n \vee \mathscr{F}_{n,m} \right] \right\} \\ = \; & \frac{1}{n} \mathrm{E}_{\boldsymbol{\beta}_0} \left\{ \left\| \sum_{i=1}^{n} [\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)] \right\|^2 \right\} \\ = \; & \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathrm{E}_{\boldsymbol{\beta}_0} ||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 \right\} \end{aligned}$$

The last equality is due to that $\left\{\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)\right\}_{i=1}^n$ are independent with mean zero.

On the other hand,

$$\begin{aligned}
& \mathrm{E}_{\beta_0}\left\{||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2\right\} \\
= \; & \mathrm{E}_{\beta_0}\left\{||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 I_{(U_i\in[\mathcal{L},\mathcal{R}])}\right\} \\
& + \mathrm{E}_{\beta_0}\left\{||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 I_{(U_i<\mathcal{L})}\right\} \\
& + \mathrm{E}_{\beta_0}\left\{||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 I_{(U_i>\mathcal{R})}\right\} \\
= \; & A_{1i} + A_{2i} + A_{3i}
\end{aligned}$$

When $U_i \in [\mathcal{L}, \mathcal{R}]$, from condition (A7), we can easily get that

$$||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0) - \psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 < M_3^2\left(\frac{\sqrt{\varepsilon}}{3M_3}\right)^2 = \frac{\varepsilon}{9}$$

So $A_{1i} < \frac{\varepsilon}{9}$. By (4.15) and the property of Riemann integral, when m is large enough, it is easily obtained that

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left\{||\psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 I_{(U_i<\mathcal{L})}\right\} < \frac{\varepsilon}{9}$$

Thus

$$\begin{aligned}
A_{2i} \;\le\; & 2\mathrm{E}_{\boldsymbol{\beta}_0}\left\{||\psi(1-U_i, Z_i, \boldsymbol{\beta}_0)||^2 I_{(U_i<\mathcal{L})}\right\} \\
& + 2\left\{||\psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta}_0)||^2 I_{(U_i<\mathcal{L})}\right\} < \frac{4\varepsilon}{9}
\end{aligned} \tag{4.26}$$

Similar to (4.26), we can reach $A_{3i} < \frac{4\varepsilon}{9}$. Hence when m_0 is large enough and for any $m > m_0$,

$$\mathrm{E}_{\beta_0}\left[\frac{1}{n}||S_n(\boldsymbol{\beta}_0) - S_{n,m}(\boldsymbol{\beta}_0)||^2\right] < \varepsilon$$

holds uniformly for all n. □

Lemma 4.2 If the assumptions (A0)-(A8) hold, then when m_0 is large enough, for any fixed $m > m_0$,

$$\mathrm{E}_{\beta_0}\left[\frac{1}{\sqrt{n}}||S_{n,m}(\boldsymbol{\beta}_0) - \bar{S}_{n,m}(\boldsymbol{\beta}_0)||\right] \to 0 \quad \text{as} \quad n \to \infty \tag{4.27}$$

Proof: Since when the event $\mathcal{E}_n$ happens, we will know the exact values of observed discretized failure times $U_i^{(m)}$'s and the bounds of censoring intervals for censored discretized

failure times. Moreover $U_i^{(m)-} = L_i^{(m)}$ and $U_i^{(m)+} = V_i^{(m)}$ for any censored discretized failure times $U_i^{(m)}$'s. Thus following the form (4.24) of $S_{n,m}(\boldsymbol{\beta})$, we have

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{\sqrt{n}}\left|\left|S_{n,m}(\boldsymbol{\beta}_0)-\bar{S}_{n,m}(\boldsymbol{\beta})\right|\right|\right] \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{\sqrt{n}}\left|\left|S_{n,m}(\boldsymbol{\beta}_0)-\bar{S}_{n,m}(\boldsymbol{\beta})\right|\right| I_{\mathcal{E}_n^c}\right] \\
\leq&\left\{\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{\sqrt{n}}\left|\left|S_{n,m}(\boldsymbol{\beta}_0)-\bar{S}_{n,m}(\boldsymbol{\beta})\right|\right|\right]^2\right\}^{1/2}[P(\mathcal{E}_n^c)]^{1/2} \\
\leq&\left\{2\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{n}\left|\left|S_{n,m}(\boldsymbol{\beta}_0)(\boldsymbol{\beta})\right|\right|^2\right]+2\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{n}\left|\left|\bar{S}_{n,m}(\boldsymbol{\beta}_0)(\boldsymbol{\beta})\right|\right|^2\right]\right\}^{1/2}[P(\mathcal{E}_n^c)]^{1/2}
\end{aligned} \tag{4.28}$$

Based on the proof of Lemma 4.1, the form (4.21) of $S_{n,m}(\boldsymbol{\beta})$ and the condition (A6), when m_0 is large enough, for any fixed $m > m_0$, we have

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{n}\left|\left|S_{n,m}(\boldsymbol{\beta}_0)(\boldsymbol{\beta})\right|\right|^2\right] \\
\leq&\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)||^2\right] \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)-\psi(1-U,Z,\boldsymbol{\beta}_0)+\psi(1-U,Z,\boldsymbol{\beta}_0)||^2\right] \\
\leq&2\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)-\psi(1-U,Z,\boldsymbol{\beta}_0)||^2\right]+2\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi(1-U,Z,\boldsymbol{\beta}_0)||^2\right] \\
\leq&2\varepsilon+2M_2
\end{aligned} \tag{4.29}$$

in which the first inequality holds because $\{\psi^*(1-U_i^{(m)},Z_i,\boldsymbol{\beta}_0)\}_{i=1}^n$ are i.i.d. with mean being 0.

Obviously, $\left\{\psi(1-U_i^{(m)},Z_i,\boldsymbol{\beta}_0)\delta_i+f(F_0(L_i^{(m)}),\triangle F_0(V_i^{(m)}),Z_i,\boldsymbol{\beta}_0)(1-\delta_i)\right\}_{i=1}^n$ are i.i.d. random variables with mean being 0, we have

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{n}\left|\left|\bar{S}_{n,m}(\boldsymbol{\beta}_0)\right|\right|^2\right] \\
\leq&\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}),\triangle F_0(V^{(m)}),Z,\boldsymbol{\beta}_0)||^2(1-\delta)\right]+\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)||^2\delta\right] \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}),\triangle F_0(V^{(m)}),Z,\boldsymbol{\beta}_0)-f(F_0(L),F_0(V),Z,\boldsymbol{\beta}_0)\right. \\
&\left.+f(F_0(L),F_0(V),Z,\boldsymbol{\beta}_0)||^2(1-\delta)\right]+\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)||^2\delta\right] \\
\leq&2\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}),\triangle F_0(V^{(m)}),Z,\boldsymbol{\beta}_0)-f(F_0(L),F_0(V),Z,\boldsymbol{\beta}_0)||^2(1-\delta)\right] \\
&+2\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L),F_0(V),Z,\boldsymbol{\beta}_0)||^2(1-\delta)\right]+\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1-U^{(m)},Z,\boldsymbol{\beta}_0)||^2\delta\right]
\end{aligned}$$

Based on the condition (A8), when $L^{(m)} < V^{(m)}$, similar to the proof of Lemma 4.1, we

have

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}), \triangle F_0(V^{(m)}), Z, \boldsymbol{\beta}_0) - f(F_0(L), F_0(V), Z, \boldsymbol{\beta}_0)||^2\right] \\
=&\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}), \triangle F_0(V^{(m)}), Z, \boldsymbol{\beta}_0) - f(F_0(L^{(m)}), F_0(V), Z, \boldsymbol{\beta}_0)\right. \\
&\left.+f(F_0(L^{(m)}), F_0(V), Z, \boldsymbol{\beta}_0) - f(F_0(L), F_0(V), Z, \boldsymbol{\beta}_0)||^2\right] \qquad (4.30) \\
\leq&2\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}), \triangle F_0(V^{(m)}), Z, \boldsymbol{\beta}_0) - f(F_0(L^{(m)}), F_0(V), Z, \boldsymbol{\beta}_0)||^2\right] \\
&+ 2\mathrm{E}_{\boldsymbol{\beta}_0}\left[||f(F_0(L^{(m)}), F_0(V), Z, \boldsymbol{\beta}_0) - f(F_0(L), F_0(V), Z, \boldsymbol{\beta}_0)||^2\right] \leq 4\varepsilon.
\end{aligned}$$

When $L^{(m)} = V^{(m)}$, from the arguments of (4.29), we have

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left[||\psi^*(1 - U^{(m)}, Z, \boldsymbol{\beta}_0)||^2\right] < 2\varepsilon + 2M_2$$

Thus under the condition (A6) and (4.15), we have

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{n}\left|\left|\bar{S}_{n,m}(\boldsymbol{\beta}_0)\right|\right|^2\right] \leq 8\varepsilon + 2M_2 \qquad (4.31)$$

Combining (4.29) and (4.31), we have

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{\sqrt{n}}\left|\left|S_{n,m}(\boldsymbol{\beta}_0) - \bar{S}_{n,m}(\boldsymbol{\beta})\right|\right|\right]^2 \leq 20\varepsilon + 8M_2$$

From that $P(\mathcal{E}_n)$ goes to 1 for fixed $m > m_0$ as n goes to infinity, the proof is completed. □

Lemma 4.3 Under the conditions (A0)-(A8) and a large enough number m_0 for any $m > m_0$, we have

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left[\frac{1}{n}\left|\left|\bar{S}_{n,m}(\boldsymbol{\beta}_0) - \widetilde{S}_n(\boldsymbol{\beta}_0)\right|\right|^2\right] < \varepsilon \qquad (4.32)$$

uniformly for all n.

Proof: Since both $\bar{S}_{n,m}(\boldsymbol{\beta}_0)$ and $\widetilde{S}_n(\boldsymbol{\beta}_0)$ are the sum of i.i.d. random variables with

mean being 0, we have

$$
\begin{aligned}
&\mathrm{E}_{\beta_0}\left[\frac{1}{\sqrt{n}}\left(\bar{S}_{n,m}(\beta_0)-\widetilde{S}_n(\beta_0)\right)\right]^2 \\
=&\frac{1}{n}\mathrm{E}_{\beta_0}\left\{\left|\left|\sum_{i=1}^{n}\left[\left(\psi^*(1-U_i^{(m)},Z_i,\beta_0)-f(F_0(L_i),F_0(V_i),Z_i,\beta_0)\right)\delta_i\right.\right.\right. \\
&\left.\left.\left.+\left(f(F_0(L_i^{(m)}),\triangle F_0(V_i^{(m)}),Z_i,\beta_0)-f(F_0(L_i),F_0(V_i),Z_i,\beta_0)\right)(1-\delta_i)\right]\right|\right|^2\right\} \\
=&\frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{E}_{\beta_0}\left|\left|\left(\psi^*(1-U_i^{(m)},Z_i,\beta_0)-f(F_0(L_i),F_0(V_i),Z_i,\beta_0)\right)\delta_i\right|\right|^2\right. \\
&\left.+\mathrm{E}_{\beta_0}\left|\left|\left(f(F_0(L_i^{(m)}),\triangle F_0(V_i^{(m)}),Z_i,\beta_0)-f(F_0(L_i),F_0(V_i),Z_i,\beta_0)\right)(1-\delta_i)\right|\right|^2\right\} \\
=&\mathrm{E}_{\beta_0}\left[\left|\left|\psi^*(1-U^{(m)},Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2\delta\right] \\
&+\mathrm{E}_{\beta_0}\left[\left|\left|f(F_0(L^{(m)}),\triangle F_0(V^{(m)}),Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2(1-\delta)\right]
\end{aligned}
\tag{4.33}
$$

When $L^{(m)}=V^{(m)}$ and a large enough number m_0, similar to the proof of Lemma 4.1, for any $m>m_0$, we have

$$
\begin{aligned}
&\mathrm{E}_{\beta_0}\left[\left|\left|\psi^*(1-U^{(m)},Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2\right] \\
=&\mathrm{E}_{\beta_0}\left[\left|\left|\psi^*(1-U^{(m)},Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2 I(V<\mathcal{L})\right] \\
&+\mathrm{E}_{\beta_0}\left[\left|\left|\psi^*(1-U^{(m)},Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2 I(L>\mathcal{R})\right] \\
&+\mathrm{E}_{\beta_0}\left[\left|\left|\psi^*(1-U^{(m)},Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2 I(\mathcal{L}\le L<V\le\mathcal{R})\right] \\
<&\varepsilon
\end{aligned}
$$

When $L^{(m)}<V^{(m)}$, from (4.30) we have

$$
\mathrm{E}_{\beta_0}\left[\left|\left|f(F_0(L^{(m)}),\triangle F_0(V^{(m)}),Z,\beta_0)-f(F_0(L),F_0(V),Z,\beta_0)\right|\right|^2\right]<4\varepsilon
$$

This completes the proof of the Lemma. □

Since $\widetilde{S}_n(\beta_0)$ is a sum of i.i.d. random variables with mean 0 and variance matrix $V_1(\beta_0)$, from Lemmas 4.1-4.3, we can easily obtain the following Lemma.

Lemma 4.4 Under conditions (A0)-(A8),

$$
\frac{1}{\sqrt{n}}S_n(\beta_0)\ \xrightarrow{D}\ N(0,V_1(\beta_0)),\quad \text{as}\quad n\longrightarrow\infty.
$$

where $V_1(\boldsymbol{\beta})$ is defined in (4.11). □

It is very easy to obtain that the Fisher information matrix based on the interval censored data $\{(L_i, V_i]\}_{i=1}^n$ is given by

$$\begin{aligned} -\frac{\partial}{\partial \boldsymbol{\beta}} S_n(\boldsymbol{\beta}) =& \sum_{i=1}^n \mathrm{E}_{\boldsymbol{\beta}} \left[-\frac{\partial}{\partial \boldsymbol{\beta}^T} f(U_i^-, U_i^+, Z_i, \boldsymbol{\beta}) \Big| \mathscr{F}_n \right] \\ & - \mathrm{Var}_{\boldsymbol{\beta}} \left[\sum_{i=1}^n f(U_i^-, U_i^+, Z_i, \boldsymbol{\beta}) \Big| \mathscr{F}_n \right] \\ =& D_{1n}(\boldsymbol{\beta}) - D_{2n}(\boldsymbol{\beta}) \end{aligned} \tag{4.34}$$

and the Fisher information matrix based on discretized interval censored data $\{[L_i^{(m)}, V_i^{(m)}]\}_{i=1}^n$ is given by

$$\begin{aligned} & -\frac{\partial}{\partial \boldsymbol{\beta}} S_{n,m}(\boldsymbol{\beta}) \\ =& \sum_{i=1}^n \mathrm{E}_{\boldsymbol{\beta}} \left\{ -\frac{\partial}{\partial \boldsymbol{\beta}^T} \left[\psi^*(1 - U_i^{(m)}, Z_i, \boldsymbol{\beta}) \delta_i + f(U_i^{(m)-}, \triangle U_i^{(m)+}, Z_i, \boldsymbol{\beta})(1 - \delta_i) \right] \Big| \mathscr{F}_{n,m} \right\} \\ & - \mathrm{Var}_{\boldsymbol{\beta}} \left[\sum_{i=1}^n \left[\psi^*(1 - U_i^{(m)}, Z_i, \boldsymbol{\beta}) \delta_i + f(U_i^{(m)-}, \triangle U_i^{(m)+}, Z_i, \boldsymbol{\beta})(1 - \delta_i) \right] \Big| \mathscr{F}_{n,m} \right] \\ =& D_{1n}^{(m)}(\boldsymbol{\beta}) - D_{2n}^{(m)}(\boldsymbol{\beta}) \end{aligned} \tag{4.35}$$

Lemma 4.5 Under the conditions (A0)-(A9),

$$-\frac{1}{n} \cdot \frac{\partial}{\partial \boldsymbol{\beta}} S_n(\boldsymbol{\beta}) \xrightarrow{P} V_2(\boldsymbol{\beta}) \quad \text{as} \quad n \to \infty$$

holds uniformly for all $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$, where $V_2(\boldsymbol{\beta})$ is defined by (4.12).

Proof: For the proof of this Lemma, we will do it by two steps, that is, the first step is to prove that $\frac{1}{n} D_{1n}(\boldsymbol{\beta}) \xrightarrow{P} V_2(\boldsymbol{\beta})$ as $n \to \infty$ and the next step is to prove that $\frac{1}{n} D_{2n}(\boldsymbol{\beta}) \xrightarrow{P} 0$ as $n \to \infty$.

(*i*) $\frac{1}{n} D_{1n}(\boldsymbol{\beta}) \xrightarrow{P} V_2(\boldsymbol{\beta})$ as $n \to \infty$.

Under condition (A8) and the inequality (4.18), we can easily obtain that

$$\mathrm{E}_{\boldsymbol{\beta}_0} \left\{ \frac{1}{n} \left\| D_{1n}(\boldsymbol{\beta}) - D_{1n}^{(m)}(\boldsymbol{\beta}) \right\| \right\} \to 0 \quad \text{as} \quad m \to \infty \tag{4.36}$$

holds uniformly for all n and any $\boldsymbol{\beta} \in \mathcal{O}_{\boldsymbol{\beta}_0}$.

For the fixed m and a large enough integer n, $P(\mathcal{E}_n) = 1$, that is, the exact values of $U_i^{(m)}$'s for observed discretized failure times and the bounds of censoring intervals for

censored discretized failure times will be known. Therefore, under conditions (A7)-(A8) and (4.18), we have that

$$\frac{1}{n}D_{1n}^{(m)}(\boldsymbol{\beta})-\frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial\boldsymbol{\beta}^T}\left[\psi^*(1-U_i^{(m)},Z_i,\boldsymbol{\beta})\delta_i+f(L_i^{(m)-},\triangle V_i^{(m)+},Z_i,\boldsymbol{\beta})(1-\delta_i)\right]\right\}\xrightarrow{P}0 \tag{4.37}$$

holds uniformly for any $\boldsymbol{\beta}\in\mathcal{O}_{\boldsymbol{\beta}_0}$ and fixed m as $n\to\infty$.

Under conditions (A7)-(A8) and (4.18), it can be easily obtained that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial\boldsymbol{\beta}^T}\left[\psi^*(1-U_i^{(m)},Z_i,\boldsymbol{\beta})\delta_i+f(L_i^{(m)-},\triangle V_i^{(m)+},Z_i,\boldsymbol{\beta})(1-\delta_i)\right]\right\}\\&-\frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial\boldsymbol{\beta}^T}\left[f(L_i,V_i,Z_i,\boldsymbol{\beta})\right]\right\}\xrightarrow{P}0\end{aligned} \tag{4.38}$$

holds uniformly for any $\boldsymbol{\beta}\in\mathcal{O}_{\boldsymbol{\beta}_0}$ and all n as $m\to\infty$.

Note that $\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial\boldsymbol{\beta}^T}\left[f(L_i,V_i,Z_i,\boldsymbol{\beta})\right]\right\}$ is a sum of i.i.d. random variables. By Weak Convergence Laws of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial\boldsymbol{\beta}^T}\left[f(L_i,V_i,Z_i,\boldsymbol{\beta})\right]\right\}\xrightarrow{P}V_2(\boldsymbol{\beta})\ \text{ as }\ n\to\infty \tag{4.39}$$

holds uniformly for any $\boldsymbol{\beta}\in\mathcal{O}_{\boldsymbol{\beta}_0}$. Combing (4.36) - (4.39), $\frac{1}{n}D_{1n}(\boldsymbol{\beta})\xrightarrow{P}V_2(\boldsymbol{\beta})$ as $n\to\infty$.

(ii) $\frac{1}{n}D_{2n}(\boldsymbol{\beta})\xrightarrow{P}0$ as $n\to\infty$

This can be done by proving

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}D_{2n}(\boldsymbol{\beta})\right\}\to 0\ \text{ as }\ n\to\infty$$

Following condition (A9), for fixed positive integer m, there exists a positive integer N, when $n>N$, we have

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}D_{2n}(\boldsymbol{\beta})\right\}\le\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}D_{2n}^{(m)}(\boldsymbol{\beta})\right\}+\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}\left|D_{1n}(\boldsymbol{\beta})-D_{1n}^{(m)}(\boldsymbol{\beta})\right|\right\} \tag{4.40}$$

holds for any $\boldsymbol{\beta}\in\mathcal{O}_{\boldsymbol{\beta}_0}$.

From (4.36), it is obvious that

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}\left|D_{1n}(\boldsymbol{\beta})-D_{1n}^{(m)}(\boldsymbol{\beta})\right|\right\}\to 0\ \text{ as }\ m\to\infty \tag{4.41}$$

holds for all n and any $\boldsymbol{\beta}\in\mathcal{O}_{\boldsymbol{\beta}_0}$.

From the definition of $\mathcal{E}_n$ in (4.19), we can see that

$$\left\{D_{2n}^{(m)}\neq 0\right\}=\mathcal{E}_n^c,$$

that is,

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}D_{2n}^{(m)}(\boldsymbol{\beta})\right\} = \mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}D_{2n}^{(m)}(\boldsymbol{\beta})I_{\mathcal{E}_n^c}\right\} \\
\leq &\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}\mathrm{E}_{\boldsymbol{\beta}}\left[\left(\sum_{i=1}^{n}\Big(\psi^*(1-U_i^{(m)}, Z_i, \boldsymbol{\beta})\delta_i\right.\right.\right. \\
&\left.\left.\left. + \ f(F_0(L_i^{(m)-}), \triangle F_0(V_i^{(m)+}), Z_i, \boldsymbol{\beta})(1-\delta_i)\Big)\right)^2 \middle| \mathscr{F}_{n,m}\right] I_{\mathcal{E}_n^c}\right\}
\end{aligned}$$

Note that for all $u, l, v \in \{u_0, u_1, u_2, \cdots, u_m\}$, $||Z|| < M_1$ and $\boldsymbol{\beta} \in \mathcal{I}_{\boldsymbol{\beta}_0}$ with $\mathcal{I}_{\boldsymbol{\beta}_0}$ being a bounded closed neighborhood contained in $\mathcal{O}_{\boldsymbol{\beta}_0}$, there exists a finite number $M(m)$, which is related to m, such that

$$\psi^*(u, Z, \boldsymbol{\beta}) \leq M(m) \ \text{ and } \ f(l, v, Z, \boldsymbol{\beta}) \leq M(m).$$

Following the inequality (4.20), we have that

$$\mathrm{E}_{\boldsymbol{\beta}_0}\left\{\frac{1}{n}D_{2n}(\boldsymbol{\beta})\right\} \leq nM^2(m)P(\mathcal{E}_n^c) \leq nM^2(m)me^{-n\delta} \to 0 \ \text{ as } \ n \to \infty \tag{4.42}$$

holds uniformly for all fixed m and any $\boldsymbol{\beta} \in \mathcal{I}_{\boldsymbol{\beta}_0}$.
Combing (4.41) and (4.42), we have $\frac{1}{n}D_{2n}(\boldsymbol{\beta}) \xrightarrow{P} 0$ as $n \to \infty$. □

Proof of Theorem 4.1 We can show (i) by a straightforward extension of Foutz (1977) [34], considered by Ni (2008) [53] and Wu (2008) [71]. The proof of (ii) can be easily completed by Taylor's expansion of $S_n(\boldsymbol{\beta})$ around the true value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$.

(i) Note that the Lemmas 4.1- 4.3 still hold when $\frac{1}{\sqrt{n}}$ is replaced by $\frac{1}{n}$. That is, $\frac{1}{n}S_n(\boldsymbol{\beta}_0)$ can be asymptotically approximated by $\frac{1}{n}\widetilde{S}_{n,m}(\boldsymbol{\beta}_0)$. By Weak Convergence Laws of Large Number, we have

$$\frac{1}{n}S_n(\boldsymbol{\beta}_0) \xrightarrow{P} 0 \ \text{ as } \ n \to \infty. \tag{4.43}$$

By Condition (A2), we know that $\frac{1}{n} \cdot \frac{\partial}{\partial \boldsymbol{\beta}^T}S_n(\boldsymbol{\beta})$ exists and is continuous with respect to $\boldsymbol{\beta}$ in an open neighborhood of $\boldsymbol{\beta}_0$;

Following Lemma 4.5, $\frac{1}{n} \cdot \frac{\partial}{\partial \boldsymbol{\beta}^T}S_n(\boldsymbol{\beta})$ is negative definite with probability going to 1 as $n \to \infty$ and

$$-\frac{1}{n} \cdot \frac{\partial}{\partial \boldsymbol{\beta}^T}S_n(\boldsymbol{\beta}) \xrightarrow{P} V_2(\boldsymbol{\beta}) \ \text{ as } \ n \to \infty$$

holds uniformly for any $\boldsymbol{\beta}$ in an open neighborhood of $\boldsymbol{\beta}_0$. Following Ni (2008) [53] and Wu (2008) [71], this completes the proof of (i).

(*ii*) By Taylor's expansion of $S_n(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$, we have that

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = -\left[\frac{1}{n}\cdot\frac{\partial}{\partial\boldsymbol{\beta}^T}S_n(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right]^{-1}\left[\frac{1}{\sqrt{n}}S_n(\boldsymbol{\beta}_0)\right], \tag{4.44}$$

where $\boldsymbol{\beta}^*$ is between $\widetilde{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Following the proof of (*i*), we also have that $\boldsymbol{\beta}^* \xrightarrow{P} \boldsymbol{\beta}_0$ as n goes to infinity. Combing Lemmas 4.4-4.5 and (4.44), the asymptotic normality of $\widetilde{\boldsymbol{\beta}}_n$ is obtained. □

Discussion of proofs for Theorems 4.1-4.4

Based on Lemmas 4.1-4.5 and Theorem 4.1, the proofs of Theorems 4.2-4.5 are vary similar to that of Theorems 3.1-3.4 and Theorems 2.1-2.4. Thus we omit them here.

4.3 Implementation

From (4.4), we can see that the marginal likelihood function is a high-dimensional integration. It is impossible to get the closed expression for $L_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n)$. Therefore, it is difficult to directly maximize (4.8) with respect to $\boldsymbol{\beta}$. On the other hand, the marginal likelihood function (4.4) has the same form as (2.1) and (3.1) except the integration region $\mathcal{D}_n$ because of the rank restriction in $\mathbf{T}_n$. Therefore for the interval censored data, we can also use the three-step MCMC-SA algorithm in Chapters 2 with the Gibbs sampling procedure for interval censored data in Appendix B to maximize (4.4) .

Theoretically, the approximated GCV (2.31) can be also used to select tuning parameter λ in the variable selection procedures for interval censored data. The similar approximation methods for rank-based log-marginal likelihood function, score function and Fisher information matrix for right censored data is also used to approximate $L_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n)$, $S_n(\boldsymbol{\beta})$ and $-\frac{\partial}{\partial\boldsymbol{\beta}}S_n(\boldsymbol{\beta})$ with interval censored data. The approximation, $\hat{L}_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n)$, of $L_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n)$ with interval censored data has the form of (3.29) except the constant c and the samples $U_{i,j}$. In this case, we can sample $U_{i,j}$ from (4.6) by virtue of Gibbs Sampling procedure for interval censored data in Appendix B. c is total number of all the possible rankings consistent with the censoring intervals $\{(L_i, V_i]\}_{i=1}^n$. However, c also increases dramatically with the increasing of sample size n. Moreover, the computation of c is very time-consuming especially when many censoring intervals are overlapped with each other. Since the complication of interval censoring, it is very difficult to develop a general

procedure to find c. To avoid calculating c and motivated by Wang *et al.* (2007) [70], we propose another criterion, BIC, to select the proper tuning parameter λ. The BIC criterion is defined by

$$BIC = -2\log(\hat{L}_n(\boldsymbol{\beta}|\mathcal{S}_n, \mathbf{Z}_n)) + e(\lambda)\log(n) \tag{4.45}$$

where $e(\lambda)$ is the degree of freedom for the selected model and has the same form of (2.30) with $\nabla^2\ell(\boldsymbol{\beta})$ replaced by $-\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$ based on (4.4). We can select the proper tuning parameter λ by minimizing (4.45). Note that given the censoring intervals $\{(L_i, V_i]\}_{i=1}^n$, $\frac{c}{n!}$ is a constant for all λ and $\boldsymbol{\beta}$. Consequently, the logarithm of $\frac{c}{n!}$ has the additive form in (4.45). On the other hand, following the approximation method in Section 3.1, the approximation of $-\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$ is independent on the constant $\frac{c}{n!}$. So in the approximation of BIC, we can omit the term $-2\log(\frac{c}{n!})$ and this will not affect the selection of a proper tuning parameter λ. The same samples of $\mathbf{U}_n$ are also used to approximate $\ell(\boldsymbol{\beta})$ regardless of the values of $\boldsymbol{\beta}$ and λ.

4.4 Numeric Studies

In this section, we will conduct some simulation studies to illustrate our proposed variable selection procedures for general transformation models with interval censored data. In the simulation studies, we also consider three special models of general transformation models (1.3)– proportional hazards Cox's regression model (PH), proportional odds regression model (PO) and generalized probit model (GP).

We use the average proportion of overlapped censoring intervals to measure the censoring degree and the average proportion of overlapped censoring intervals is defined as follows. For ith censoring interval, the proportion of censoring intervals overlapped with it is given by

$$p_i = 1 - \frac{\#(A_i) + \#(B_i)}{n},$$

where $\#(\cdot)$ means the number of elements in the set A_i or B_i defined by (4.2) and (4.3). Then the average proportion of overlapped censoring intervals is given by $p = \frac{\sum_{i=1}^n p_i}{n}$. Following Satten (1996) [58], we will generate a censoring interval for each failure time by an independent renewal process which begun at time 0 with log-normally distributed

Table 4.1: Variable selection results for PH, PO and GP models with Light interval censoring

Light		$n = 100$			$n = 200$		
			Aver. no. of 0 Coef.			Aver. no. of 0 Coef.	
Models	Penalty	MMSE	correct	incorrect	MMSE	correct	incorrect
PH	HARD	0.067	4.600	0.000	0.053	4.853	0.000
	SCAD	0.063	4.446	0.000	0.050	4.958	0.000
	LASSO	0.456	4.070	0.000	0.389	4.550	0.000
	ALASSO	0.081	4.382	0.000	0.040	4.826	0.000
	Oracle	0.060	5.000	0.000	0.035	5.000	0.000
PO	HARD	0.196	4.717	0.040	0.071	4.641	0.019
	SCAD	0.277	4.308	0.056	0.095	4.333	0.006
	LASSO	0.660	4.160	0.015	0.552	4.480	0.000
	ALASSO	0.207	4.297	0.005	0.177	4.780	0.003
	Oracle	0.124	5.000	0.000	0.060	5.000	0.000
GP	HARD	0.090	4.450	0.000	0.032	4.740	0.000
	SCAD	0.086	4.460	0.000	0.028	4.890	0.000
	LASSO	0.563	4.293	0.000	0.441	4.910	0.000
	ALASSO	0.077	4.021	0.000	0.029	4.890	0.000
	Oracle	0.054	5.000	0.000	0.025	5.000	0.000

Note: 0.000*s indicate that the corresponding values are less than 0.0005; ALASSO means Adaptive-LASSO.

increments. We consider two censoring cases –light censoring and heavy censoring – by proper choice of mean and coefficient of variation for the increment. The average proportion of overlapped intervals for light censored data is about 11% and it is 56% for heavy censoring. The BIC criterion (4.45) without the term $-2\log(\frac{c}{n!})$ is used to select tuning parameter λ. In the approximation of BIC criterion, $M_0 = 20000$.

Table 4.2: Variable selection results for PH, PO and GP models with Heavy interval censoring

Heavy		$n = 100$			$n = 200$		
			Aver. no. of 0 Coef.			Aver. no. of 0 Coef.	
Models	Penalty	MMSE	correct	incorrect	MMSE	correct	incorrect
PH	HARD	0.103	4.417	0.000	0.060	4.815	0.000
	SCAD	0.074	4.657	0.000	0.044	4.963	0.000
	LASSO	0.692	4.050	0.000	0.583	4.744	0.000
	ALASSO	0.160	4.135	0.000	0.113	4.757	0.000
	Oracle	0.085	5.000	0.000	0.067	5.000	0.000
PO	HARD	0.533	4.130	0.033	0.161	4.740	0.038
	SCAD	0.506	4.600	0.053	0.125	4.738	0.051
	LASSO	0.808	4.310	0.028	0.680	4.640	0.000
	ALASSO	0.378	4.000	0.017	0.173	4.640	0.003
	Oracle	0.230	5.000	0.000	0.099	5.000	0.000
GP	HARD	0.130	4.883	0.002	0.051	4.992	0.000
	SCAD	0.112	4.470	0.000	0.057	4.960	0.000
	LASSO	0.464	3.900	0.000	0.436	4.560	0.000
	ALASSO	0.196	4.080	0.000	0.065	4.110	0.000
	Oracle	0.097	5.000	0.000	0.048	5.000	0.000

In this simulations, we generate 100 data sets consisting of $n = 100$ and 200 censoring interval observations from the general transformation models (1.3) with $\Phi(u, v, w) = g^{-1}(g(u) + v^T w)$. In the models, $1 - g^{-1}(u)$ takes standard exponential distribution function, standard logistic distribution function and standard normal distribution function, which correspond to the proportional hazards regression models, proportional odds regression models and generalized probit models. All the covariates in $Z \in R^9$ are generated from standard normal distribution independently and the baseline survival function is

Table 4.3: Summary of results for nonzero effects in PH model with Light interval censoring

Light		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.0578	0.0501	-0.0517	0.0665	-0.0854	0.0893	-0.0980	0.0948
	SStd	0.1248	0.1217	0.1147	0.1224	0.0731	0.0726	0.0769	0.0712
	MStd	0.1165	0.1174	0.1150	0.1158	0.0799	0.0794	0.0800	0.0797
SCAD	Bias	-0.0671	0.0368	-0.0594	0.0484	-0.1013	0.0870	-0.0958	0.0941
	SStd	0.1103	0.1164	0.1310	0.0958	0.0833	0.0682	0.0841	0.0723
	MStd	0.1178	0.1169	0.1156	0.1182	0.0795	0.0784	0.0798	0.0795
LASSO	Bias	-0.3177	0.3347	-0.3279	0.3463	-0.3197	0.3196	-0.3089	0.3187
	SStd	0.1073	0.1083	0.0982	0.1010	0.0926	0.0807	0.0812	0.0835
	MStd	0.0830	0.0819	0.0823	0.0815	0.0593	0.0590	0.0594	0.0595
ALASSO	Bias	-0.0802	0.0806	-0.0815	0.0812	-0.0911	0.0756	-0.0897	0.0721
	SStd	0.1038	0.1064	0.1009	0.1120	0.0756	0.0734	0.0777	0.0805
	MStd	0.1154	0.1146	0.1147	0.1153	0.0780	0.0785	0.0786	0.0786
Oracle	Bias	-0.0832	0.0649	-0.0660	0.0598	-0.0649	0.0617	-0.0709	0.0710
	SStd	0.1167	0.1106	0.1041	0.1263	0.0807	0.0824	0.0686	0.0789
	MStd	0.1164	0.1144	0.1174	0.1162	0.0802	0.0799	0.0794	0.0796

$S_0(t) = \exp\{-t\}$. We take the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (0.8, 0.0, -0.8, 0.0, 0.0, 0.8, 0.0, -0.8)^T$, that is, there are five zero effects included in the models.

We run the three-step MCMC-SA algorithm in Chapter 2 and Gibbs sampling for interval censored data to conduct the simulation studies. We will adopt the same programme parameter setting as the simulation studies in Section 2.5.1. We use median of mean squared error (MMSE), defined in Section (2.5) or (3.4), to assess the efficiency of the proposed variable selection methods and BIC (4.45) to select tuning parameter λ on a grid of points.

Table 4.4: Summary of results for nonzero effects in PH model with Heavy interval censoring

Heavy		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.0755	0.0515	-0.0324	0.0515	-0.0924	0.0707	-0.0778	0.0857
	SStd	0.1465	0.1275	0.1597	0.1430	0.0914	0.0956	0.1073	0.0903
	MStd	0.1449	0.1440	0.1491	0.1465	0.0981	0.0967	0.0975	0.0976
SCAD	Bias	-0.0408	0.0498	-0.0577	0.0595	-0.0854	0.0773	-0.0696	0.0809
	SStd	0.1207	0.1272	0.1373	0.1456	0.0942	0.0919	0.0812	0.0927
	MStd	0.1436	0.1456	0.1443	0.1449	0.0968	0.0974	0.0980	0.0976
LASSO	Bias	-0.3828	0.3592	-0.3989	0.3818	-0.4116	0.4055	-0.4217	0.4040
	SStd	0.1059	0.1003	0.1092	0.0908	0.0730	0.0700	0.0830	0.0834
	MStd	0.0910	0.0924	0.0897	0.0913	0.0602	0.0607	0.0596	0.0603
ALASSO	Bias	-0.1333	0.1336	-0.1461	0.1428	-0.1397	0.1426	-0.1435	0.1434
	SStd	0.1247	0.1333	0.1321	0.1413	0.0948	0.0833	0.0893	0.0911
	MStd	0.1443	0.1447	0.1452	0.1440	0.0976	0.0983	0.0968	0.0967
Oracle	Bias	-0.0748	0.0816	-0.0984	0.1174	-0.1076	0.1067	-0.1015	0.1070
	SStd	0.1351	0.1250	0.1343	0.1329	0.0824	0.0826	0.0917	0.0832
	MStd	0.1448	0.1432	0.1433	0.1423	0.0954	0.0969	0.0959	0.0955

MMSE, the average number of zero effects correctly detected, labeled by "correct" and the average number of nonzero effects wrongly excluded from models, labeled by "incorrect" are reported in Tables 4.1 and 4.2. The results of nonzero effect estimates are given in Tables 4.3 - 4.4 for PH, in Tables 4.5 - 4.6 for PO and in Tables 4.7 - 4.8 for GP, including bias (Bias), sample standard deviation (SStd) and mean of estimated standard deviation (MStd). In the Tables, MMstds for PMMLEs are calculated based on the covariance matrix formula (2.28) while MStds for oracle estimates are the inverse of Fisher information matrix at MMLE.

Table 4.5: Summary of results for nonzero effects in PO model with Light interval censoring

Light		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	-0.0090	-0.0570	0.0529	-0.0234	-0.0034	-0.0117	0.0030	0.0006
	SStd	0.2687	0.2840	0.2473	0.2588	0.1881	0.1271	0.1839	0.2014
	MStd	0.1779	0.1841	0.1875	0.1856	0.1283	0.1319	0.1298	0.1274
SCAD	Bias	-0.0101	-0.0492	-0.0012	-0.0268	0.0412	-0.0212	0.0135	-0.0387
	SStd	0.2761	0.2511	0.3061	0.2993	0.1394	0.1304	0.1804	0.1754
	MStd	0.1796	0.1866	0.1773	0.1806	0.1331	0.1321	0.1302	0.1311
LASSO	Bias	-0.3731	0.3621	-0.3936	0.3703	-0.3450	0.3679	-0.3443	0.3606
	SStd	0.2264	0.1815	0.1943	0.1970	0.1429	0.1425	0.1603	0.1534
	MStd	0.1047	0.1083	0.1037	0.1059	0.0792	0.0778	0.0787	0.0778
ALASSO	Bias	-0.0809	0.0555	-0.0889	0.0663	-0.1583	0.1161	-0.1439	0.1552
	SStd	0.1920	0.2214	0.2536	0.2140	0.1662	0.1607	0.1646	0.1843
	MStd	0.1668	0.1687	0.1626	0.1677	0.1081	0.1414	0.1391	0.1571
Oracle	Bias	0.0302	-0.0099	0.0393	-0.0154	0.0136	-0.0186	0.0133	-0.0143
	SStd	0.2181	0.2189	0.1991	0.1862	0.1359	0.1391	0.1291	0.1466
	MStd	0.1948	0.1918	0.1925	0.1940	0.1326	0.1327	0.1322	0.1322

From Table 4.1, for the light interval censoring case, we can see that the four variable selection methods can correctly select about the same number of significant covariates for each setting. Moreover, we also find that the number of covariates selected by all the methods will get closer to the true number 5 as sample size increases. According to MMSE, the variable selection methods with SCAD, HARD and Adaptive-LASSO penalties outperform method with LASSO penalty in all the settings. PMMLEs with SCAD, HARD and Adaptive-LASSO penalties also perform as well as Oracle estimates as if we

Table 4.6: Summary of results for nonzero effects in PO model with Heavy interval censoring

Heavy		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.1090	-0.0589	0.0315	-0.0910	0.0128	-0.0527	0.0740	0.0197
	SStd	0.3398	0.3989	0.3905	0.2801	0.2477	0.2249	0.2306	0.2753
	MStd	0.2511	0.2393	0.2364	0.2479	0.1635	0.1666	0.1660	0.1576
SCAD	Bias	0.1390	-0.0148	0.0926	-0.0639	-0.0074	-0.0442	0.0152	-0.0016
	SStd	0.3206	0.3709	0.2910	0.3208	0.2671	0.2028	0.2425	0.2818
	MStd	0.2453	0.2263	0.2447	0.2398	0.1560	0.1655	0.1601	0.1601
LASSO	Bias	-0.4686	0.4089	-0.4213	0.4208	-0.3987	0.4052	-0.3953	0.4066
	SStd	0.2029	0.2073	0.1926	0.1776	0.1378	0.1313	0.1443	0.1422
	MStd	0.1001	0.1113	0.1100	0.1114	0.0857	0.0858	0.0857	0.0847
ALASSO	Bias	-0.0800	0.1570	-0.0957	0.0890	-0.1452	0.1276	-0.1072	0.1242
	SStd	0.2826	0.2897	0.2920	0.2653	0.2167	0.1968	0.1610	0.2000
	MStd	0.2049	0.1938	0.2046	0.2045	0.1257	0.1269	0.1295	0.1265
Oracle	Bias	0.1364	-0.0963	0.0677	-0.0699	0.0660	-0.0390	0.0549	-0.0776
	SStd	0.3081	0.2983	0.2566	0.2702	0.1897	0.1656	0.2001	0.2039
	MStd	0.2575	0.2542	0.2555	0.2524	0.1705	0.1696	0.1717	0.1714

have known the significant variables in advance. Based on the Table 4.2, the variable selection procedures for heavy interval censoring case can perform as well as that for light interval censoring case.

In Tables 4.3 - 4.8, Bias, SStd and MStd based on 100 simulations are reported for each setting. In our simulations, if a significant covariate is excluded from the models, its estimate and estimated standard deviation will be set to be 0. SStd is the sample standard deviation of estimates based on 100 runs, which can be seen as the true value of

Table 4.7: Summary of results for nonzero effects in GP model with Light interval censoring

Light		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.0521	-0.0641	0.0581	-0.0594	0.0233	-0.0200	0.0108	-0.0323
	SStd	0.1606	0.1657	0.1395	0.1512	0.0859	0.0872	0.1103	0.0921
	MStd	0.1216	0.1238	0.1235	0.1217	0.0824	0.0823	0.0816	0.0827
SCAD	Bias	0.0499	-0.0369	0.0620	-0.0422	0.0121	-0.0097	0.0167	-0.0077
	SStd	0.1480	0.1346	0.1667	0.1910	0.1012	0.0899	0.1027	0.0886
	MStd	0.1206	0.1213	0.1202	0.1191	0.0822	0.0814	0.0813	0.0810
LASSO	Bias	-0.3234	0.3334	-0.3389	0.3159	-0.3614	0.3704	-0.3574	0.3621
	SStd	0.0900	0.0978	0.1010	0.1012	0.0810	0.0786	0.0825	0.0847
	MStd	0.0816	0.0810	0.0808	0.0818	0.0583	0.0578	0.0583	0.0581
ALASSO	Bias	0.0236	-0.0470	0.0212	-0.0373	-0.0082	0.0019	-0.0167	-0.0001
	SStd	0.1512	0.1424	0.1423	0.1470	0.0934	0.0958	0.0961	0.0955
	MStd	0.1247	0.1239	0.1244	0.1237	0.0805	0.0815	0.0802	0.0807
Oracle	Bias	0.0328	-0.0371	0.0211	-0.0254	0.0213	-0.0236	0.0255	-0.0216
	SStd	0.1450	0.1344	0.1373	0.1182	0.0884	0.0915	0.0922	0.1000
	MStd	0.1210	0.1240	0.1220	0.1214	0.0838	0.0830	0.0845	0.0838

standard deviation of $\hat{\beta}_n$. MStd is the sample mean of 100 estimated standard deviation based on the variance formula (2.28). So the difference between SStd and MStd should be reasonably small if the variance formula works very well. From Tables 4.3, 4.5 and 4.7, in the light interval censoring, all the Biases of PMMLEs with HARD, SCAD and ALASSO penalties are reasonably as small as Oracle estimates for each situation. Since LASSO method is a shrinkage one, the estimates based on LASSO penalty suffer from relatively large bias and they can not be used as an efficient estimate. MStd in all the settings based on 100 runs are very reasonably close to their corresponding SStd. Moreover, the

Table 4.8: Summary of results for nonzero effects in GP model with Heavy interval censoring

Heavy		$n = 100$				$n = 200$			
Penalty		β_1	β_3	β_6	β_9	β_1	β_3	β_6	β_9
HARD	Bias	0.1154	-0.1126	0.1022	-0.1086	0.0664	-0.0799	0.0492	-0.0545
	SStd	0.1974	0.1867	0.1886	0.1977	0.1152	0.1207	0.1258	0.1160
	MStd	0.1554	0.1577	0.1582	0.1562	0.1031	0.1030	0.1028	0.1030
SCAD	Bias	0.0750	-0.1030	0.1038	-0.1182	0.0657	-0.0687	0.0547	-0.0592
	SStd	0.1763	0.1834	0.1898	0.1922	0.1164	0.1254	0.1134	0.1447
	MStd	0.1571	0.1597	0.1567	0.1597	0.1032	0.1039	0.1032	0.1037
LASSO	Bias	-0.3029	0.3255	-0.3203	0.3080	-0.3183	0.3180	-0.3102	0.3251
	SStd	0.1219	0.1258	0.1178	0.1310	0.1025	0.0875	0.0939	0.1015
	MStd	0.0981	0.0965	0.0973	0.0975	0.0671	0.0674	0.0674	0.0669
ALASSO	Bias	0.1173	-0.1244	0.0947	-0.1034	0.0141	-0.0351	0.0298	-0.0050
	SStd	0.2366	0.2031	0.2009	0.2122	0.1409	0.1280	0.1460	0.1203
	MStd	0.1538	0.1562	0.1560	0.1578	0.1199	0.1185	0.1186	0.1187
Oracle	Bias	0.1196	-0.0907	0.1012	-0.0843	0.0820	-0.0650	0.0711	-0.0620
	SStd	0.2027	0.1916	0.1709	0.1691	0.1310	0.1315	0.1180	0.1151
	MStd	0.1587	0.1584	0.1575	0.1573	0.1037	0.1032	0.1036	0.1041

values of SStd, MStd and their difference decrease when sample size increases. Hence, Tables 4.3, 4.5 and 4.7 suggest that the variance formula can also perform very well for interval censored data and its performance will improve when sample size increase. In a word, PMMLEs with SCAD, HARD and Adaptive-LASSO penalties can perform as well as Oracle estimate in terms of estimation and variable selection for the general transformation models with light interval censoring. From Tables 4.4, 4.6 and 4.8, we can also find the similar conclusions for heavy interval censoring case as that for light interval censoring case.

Therefore the simulation results are consistent with oracle properties given in Section 4.2. An interesting finding is that although the proposed variable selection procedures for light interval censoring performs better than heavy interval censoring case in terms of variable selection and estimation, the difference is very small. This may be the reason that the proposed procedures are only dependent on the ranking of observations and have nothing with censoring distribution and baseline distribution. So we believe that the proposed procedure should be efficient for the informative censoring case.

□ **End of chapter.**

Chapter 5

Conclusions and Further Studies

In this thesis, we considered the variable selection for general transformation models with ranking data, right censored data and interval censored data. The variable selection procedures are done by maximizing rank-based penalized log-marginal likelihood function. We mainly considered four penalty functions –HARD, SCAD, LASSO and ALASSO. We also proposed a three-step MCMC-SA algorithm by developing the three-stage MCMC-SA algorithm in Gu, *et al.* (2005) [40] and MCMC-SA algorithm in Gu and Kong (1998) [39]. Through the variable selection procedure, we not only can select important variables but also can estimate the corresponding effects simultaneously except method with LASSO. The approach with LASSO penalty would not give a consistent estimate for large coefficients when regularization parameter λ is large although it can produce sparse solution. When λ is too small, LASSO can not produce sufficiently sparse models. So PMMLEs with LASSO penalty can be not seen as an efficient parameter estimate. One advantage of the proposed procedures is independent of baseline distribution and censoring distribution, which is enjoyed by partial likelihood method [25]. Therefore our proposed procedures may allow the informative censoring. This guess has been empirically illustrated by simulation studies.

In Chapter 2, we studied the variable selection for general transformation models with ranking data. With proper penalty function, we established $\sqrt{n}$-consistency and oracle properties of penalized maximum marginal likelihood estimate under some regular conditions. We proposed a three-step MCMC-SA algorithm for the variable selection procedures with ranking data by developing the three-stage MCMC-SA algorithm in Gu,

et al. (2005) [40] and the MCMC-SA algorithm in Gu and Kong (1998) [39]. Based on Newton-Raphson iterative formula, we proposed a covariance matrix formula. Simulation studies showed that this formula is very effective. We also extend the procedures into stratified ranking data. At last, we applied the variable selection procedures to analyze Hong Kong Horse Racing data. In Chapter 3, we considered variable selection for general transformation models with right censored data. We also present the consistency and oracle properties. Simulation studies illustrated that the proposed variable selection procedures can also perform very well for right censored data. We also applied the procedure to the analysis of *PBC* data and reach to the similar conclusions as done by others previously. In Chapter 4, we considered the variable selection for general transformation models with interval censored data. We also proved the asymptotic properties of rank-based maximum marginal likelihood estimate with interval censored data by discretization technique, based on which we further gave the $\sqrt{n}$-consistency and oracle properties for the rank-based penalized maximum marginal likelihood estimate.

From (3.1) and (4.4), we can see that marginal likelihood function is the probability of complete ranking given the censored data. That is, the marginal likelihood is independent of censoring distribution. So the proposed variable selection procedure can also deal with the right informative censored or interval informative censored survival data. In microarray gene data, the dimension of covariate increases as the increasing of sample size and generally $p >> n$. In this setting, our proposed procedure can not work. Thus variable selection for high-dimensional data should be also considered. All the aspects will be studied in our subsequent research.

□ **End of chapter.**

Appendix A

Three-stage MCMC-SA Algorithm

The three stage Monte Carlo Markov Chain stochastic approximation (MCMC-SA) algorithm is given in Gu *et al.* (2005) [40] and we present it again as follows, Its ideas are: In stage I, a large gain constant sequence are used in stochastic approximation to force the estimates to move quickly into a small neighborhood of the estimate $\hat{\boldsymbol{\beta}}$. In stage II, the off-line average method of Polyak and Juditski (1992) [56] to obtain the MLE. In stage III, a new Markov chain is run based on the parameter obtained in stage II so that the variance estimate can be obtained. Specifically,

Stage I Choose a positive integer m and K_0 and set an initial value $\boldsymbol{\beta}_0$, an initial matrix Γ_0, an initial data $\mathbf{U}_{0,m}$ and $k = 1$, then iterate the following step 1 and step 2 until $k = K_1$.

Step 1 For fixed k, set $\mathbf{U}_{k,0} = \mathbf{U}_{k-1,m}$. For $i = 1, 2, \cdots, m$, generate $\mathbf{U}_{k,i}$ from the transition probability $\Pi_{\boldsymbol{\beta}_{k-1}}(U_{k,i-1},.)$ with stationary distribution $p(\mathbf{u}_n; \boldsymbol{\beta}|\mathcal{R}_n, \mathbf{Z}_n)$;

Step 2 Update the estimate $\hat{\boldsymbol{\beta}}$ by

$$\Gamma_k = \Gamma_{k-1} + \gamma_{1k}(\overline{I}_0(\boldsymbol{\beta}_{k-1}; \mathbf{U}_k) - \Gamma_{k-1}),$$

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \gamma_{1k}\Gamma_k^{-1}\overline{H}(\boldsymbol{\beta}_{k-1}; \mathbf{U}_k),$$

where $\mathbf{U}_k = (\mathbf{U}_{k,1}, \mathbf{U}_{k,2}, \cdots, \mathbf{U}_{k,m})$, $\mathbf{U}_{k,i} = (U_{k,i,1}, U_{k,i,2}, \cdots, U_{k,i,n})$ with $i =$

$1, 2, \cdots, m$,

$$\overline{H}(\boldsymbol{\beta}_{k-1}; \mathbf{U}_k) = \frac{1}{m}\sum_{i=1}^{m} H(\boldsymbol{\beta}_{k-1}; \mathbf{U}_{k,i}),$$

$$\overline{I}_0(\boldsymbol{\beta}_{k-1}; \mathbf{U}_k) = -\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\beta}^T} H(\boldsymbol{\beta}_{k-1}; \mathbf{U}_{k,i}),$$

$$H(\boldsymbol{\beta}; \mathbf{u}_n) = \sum_{j=1}^{n} \psi(u_j, Z_j, \boldsymbol{\beta}),$$

$\gamma_{1k} = 10/(k^{c_1} + 10)$, c_1 is arbitrary real number in the range of $(0, 1/2)$ and K_1 is determined by

$$K_1 = \inf\left\{ K \geq K_0 : \left|\left| \sum_{s=K-K_0+1}^{K} \text{sign}(\boldsymbol{\beta}_s - \boldsymbol{\beta}_{s-1})/K_0 \right|\right| \leq 0.1 \right\}$$

Stage II Following the stage I, take the final values $\boldsymbol{\beta}, \Gamma$ and $\mathbf{U}$ of the stage I as the initial values of the stage II and iterate the same step 1 and step 2 of the stage I with $k = 1, 2, \cdots, K_2$ and replacing γ_{1k} by $\gamma_{2k} = 10/(k^{c_2} + 10)$ with $c_2 \in (1/2, 1)$. K_2 is determined by

$$K_2 = \inf\left\{ k : \hat{\Delta}_k \leq 0.0001 \right\} \tag{A.1}$$

where

$$\hat{\Delta}_k = \tilde{H}_k^T \tilde{\Gamma}_{k-1}^{-1} \tilde{H}_k + \text{trace}(\tilde{\Gamma}_{k-1}^{-1} \tilde{\Sigma}_k)/k$$

$$\tilde{\Gamma}_k = \tilde{\Gamma}_{k-1} + (\Gamma_k - \tilde{\Gamma}_{k-1})/k$$

$$\tilde{\boldsymbol{\beta}}_k = \tilde{\boldsymbol{\beta}}_{k-1} + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_{k-1})/k$$

$$\tilde{H}_k = \tilde{H}_{k-1} + (\overline{H}(\boldsymbol{\beta}_k; \mathbf{U}_k) - \tilde{H}_{k-1})/k$$

with $\tilde{\Gamma}_0 = \mathbf{0}$, $\tilde{H}_0 = \mathbf{0}$ and $\tilde{\boldsymbol{\beta}}_0 = \mathbf{0}$. $\hat{\Sigma}_k$ is an estimate of the Monte Carlo error and it can be estimated by the sample covariance of $\{\overline{H}(\boldsymbol{\beta}_{j-1}; \mathbf{U}_j), j = 1, 2, \cdots, k\}$. After K_2th iteration, the off-line average $\tilde{\boldsymbol{\beta}}_{K_2}$ is used our final estimate of $\hat{\boldsymbol{\beta}}$.

Stage III Choose a large M and run the same Markov chain M iterations with the fixed value $\tilde{\boldsymbol{\beta}}_{K_2}$. The variance estimate of $\hat{\boldsymbol{\beta}}$ is taken to be the inverse of $\frac{1}{M}\sum_{i=1}^{M} I_n(\tilde{\boldsymbol{\beta}}_{K_2}; \mathbf{U}_i)$, where

$$I_n(\boldsymbol{\beta}; \mathbf{u}_n) = -\frac{\partial}{\partial \boldsymbol{\beta}^T} H(\boldsymbol{\beta}; \mathbf{u}_n) - H^T(\boldsymbol{\beta}; \mathbf{u}_n) H(\boldsymbol{\beta}; \mathbf{u}_n).$$

Remark $\boldsymbol{\beta}_0$, Γ_0, m and K_0 in stage I can be chosen rather arbitrary. Only $c_1 \in (0, 1/2)$, will $\boldsymbol{\beta}$ go quickly towards $\hat{\boldsymbol{\beta}}$. An initial data $\mathbf{U}_{0,m}$ can be obtained as follows for the three type of data:

- For ranking data:

  We firstly generate a sample of size n from uniform distribution over (0,1) then order it such that it is consistent with $\mathcal{R}_n$. Then the ordered sample can be taken as the initial values of $\mathbf{U}$;

- For right censored data:

  Suppose we have observed event times $Y = (Y_1, Y_2, \cdots, Y_n)^T$, then the initial data $\mathbf{U}_{0,m}$ can be taken as

  $$\mathbf{U}_{0,m} = \frac{Y}{\max_{i=1}^{n}\{Y_i\} + 1/n};$$

- For interval censored data:

  Suppose we have censoring interval observations $\{(L_i, V_i]\}_{i=1}^{n}$, then the initial data $\mathbf{U}_{0,m}$ can be taken as

  $$U_{0,m,i} = \frac{\tilde{U}_i}{\max_{i=1}^{n}\{\tilde{U}_i\} + 1/n}, \quad i = 1, 2, \cdots, n,$$

  where $\tilde{U}_i = L_i + \alpha_i(V_i - L_i)$ and α_i is a random number from uniform distribution over $(0, 1)$.

□ **End of chapter.**

Appendix B

Gibbs Sampling Procedure for Interval Censored Data

This procedure is given in the appendix of Gu *et al.* (2005). We present it here again. Let $Y = (Y_1, Y_2, \cdots, Y_n)$ be n independent random variables each is restricted to $(0, 1)$. Assume Y_i has survival function $\Phi(1 - y, Z_i, \beta)$. Define

$$\varepsilon = \{(Y_1, Y_2, \cdots, Y_n)^T : \quad Y_i^- < Y_i < Y_i^+, \quad i = 1, 2, \cdots, n\},$$

where

$$Y_i^- = \begin{cases} \max\{Y_j : \quad j \in B_i\} & B_i \neq \varnothing \\ 0 & B_i = \varnothing \end{cases}$$

and

$$Y_i^+ = \begin{cases} \min\{Y_j : \quad j \in A_i\} & A_i \neq \varnothing \\ 1 & A_i = \varnothing \end{cases}$$

where A_i and B_i are defined by (4.2) and (4.3). It is easy to see that the distribution of Y conditional on $Y \in \varepsilon$ is the conditional distribution $p(\cdot, \beta)$ defined by (4.6). Moreover, conditional on all other $Y_j (j \neq i)$ fixed, Y_i follows the distribution function $\Phi(1 - y, Z_i, \beta)$ restricted to the interval (Y_i^-, Y_i^+).

Following the Gibbs sampling idea, the following three steps give an irreducible Markov chain with stationary distribution $p(, \beta)$. Let $Y_k = (Y_1^k, Y_2^k, \cdots, Y_n^k)^T$ be the current values of Y. Then, to generate next value Y_{k+1} from $\Pi_\beta(Y_k, \cdot)$, the sampling procedure proceeds as follows.

Step 0. Set $i = 1$;

Step 1. Set $V_i^- = 1 - \Phi(1 - Y_i^-, Z_i, \boldsymbol{\beta})$ and $V_i^+ = 1 - \Phi(1 - Y_i^+, Z_i, \boldsymbol{\beta})$;

Step 2. Generate U_i^{k+1} from $Unif[V_i^-, V_i^+]$. Set $Y_i^{k+1} = 1 - \Phi^{-1}(1 - U_i^{k+1}, Z_i, \boldsymbol{\beta})$, where $\Phi^{-1}(u, v, w)$ is the inverse function of $\Phi(u, v, w)$ in terms of the first argument.

Step 3. If $i < n$, then $i = i + 1$ and go to Step 2. Otherwise stop.

□ **End of chapter.**

Bibliography

[1] D. M. Allen. The relatoinship between variable selection and data augumentation and a mthod for prediction. *Thehnometrics*, 16(1):125–127, 1974.

[2] A. Antoniadis. Wavelets in statistics: A review (with discussion). *Journal of the Italian Statistical Association*, 6(2):97–130, 1997.

[3] A. Antoniadis and J. Q. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–955, 2001.

[4] A. Antoniadis, P. Fryzlewicz, and L. Frédérique. The dantzig selector in coxs proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):531–552, 2010.

[5] P. Bacchetti and C. Quale. Generlized additive models with interval-censored data and time-varying covariates: application to human immunodeficiency virus infection in hemophiliacs. *Biometrics*, 58(2):443–447, 2002.

[6] A. Barron, L. Birge, and P. Massar. Risk bounds for model selection via penalization. *Theory Related Fields*, 113(3):301–413, 1999.

[7] S. Bennett. Analysis of survival data by proportional odds model. *Statistics in Medicine*, 2(2):273–277, 1983.

[8] W. Benter. *Computer based horse race handicapping and wagering aystems: a report*, pages 183–198. New York: Academic Press, 1994.

[9] R. N. Bolton and R. G. Chapman. Searching for positive returns at the track: a multinomial logit model for handicapping horse races. *Management Science*, 32(8):1040–1060, 1986.

[10] L. Breiman. Better subset regression using the nonnagative garrote. *Technometricx*, 37(4):373–384, 1995.

[11] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[12] H. Y. Chen. Fitting semiparametric transformation regression models to data from a modified case cohort design. *Biometrika*, 88(1):255–268, 2001.

[13] K. Chen, Z. Jin, and Z. L. Ying. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668, 2002.

[14] L. Chen and J. G. Sun. A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. *Conputatin Statistics and Data Analysis*, 54(4):1109–1116, 2010.

[15] Z. Chen and R. Little. Ra profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime Data Analysis*, 7(3):207–224, 2001.

[16] S. C. Cheng, L. J. Wei, and Z. Ying. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

[17] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34(2):187–220, 1972.

[18] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[19] P. Craven and G. Wahba. Smoothing noisy data with spline functions estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):383–393, 1979.

[20] J. Cuzick. Rank regression. *The Annals of Statistics*, 16(4):1369–1389, 1988.

[21] D. Dabrowska and K. Doksum. Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics*, 15(1):1–22, 1988.

[22] E. Dickson, P. Grambsch, T. Fleming, L. Fisher, and A. Langworthy. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*, 10(1):1–7, 1989.

[23] D. L. Donoha and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

[24] J. Fan. Comments on wavelets instatistics: a review by a. antoniadis. *Journal of the Italian Statistical Association*, 6(2):131–138, 1997.

[25] J. Fan and R. Li. Variable selection for cox proportional hazards model and frailty model. *The Annals of Statistics*, 30(2):74–99, 2002.

[26] J. Fan and R. Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians*, III:595–622, 2006.

[27] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

[28] J. Q. Fan and R. Z. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[29] D. Faraggi and R. Simon. Bayesian variable selection method for censored survival data. *Biometrics*, 54(4):1475–1485, 1998.

[30] J. P. Fine. Analyzing competing risks data with transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(4):817–830, 1999.

[31] J. P. Fine, Z. Ying, and L. G. Wei. On the linear transformation model for censored data. *Biometrika*, 85(4):980–986, 1998.

[32] D. M. Finkelstein. A proportional hazards model for interval-censored failure time data. *Biometrics*, 42(4):845–854, 1986.

[33] D. M. Finkelstein and R. A. Wolfe. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41(4):933–945, 1985.

[34] R. V. Foutz. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, 72(357):147–148, 1977.

[35] I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression. *Technometrics*, 35(2):109–148, 1993.

[36] W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[37] P. E. Gabriel and J. R. Marsden. An examination of market efficiency in british racetrack betting. *Journal of Political Economy*, 98(4):874–885, 1990.

[38] P. E. Gabriel and J. R. Marsden. An examination of efficiency in british racetrack betting: errata and corrections. *Journal of Political Economy*, 99(3):657–659, 1991.

[39] M. G. Gu and F. H. Kong. A stochastic approximation algorithm with markov chain monte-carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences of the United States of American*, 95:7270–7274, 1998.

[40] M. G. Gu, L. Q. Sun, and G. X. Zuo. A baseline-free procedure for transformation models under interval censorship. *Lifetime Data Analysis*, 11(4):473–488, 2005.

[41] D. B. Hausch, W. T. Ziemba, and M. Rubinstein. Efficiency of the market for racetrack betting. *Management Science*, 27(12):1435–1452, 1981.

[42] F. Hsieh. On heteroscedastic hazards regression models: theory and application. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):63–79, 2001.

[43] B. Huang. *Asymptotic properties of general transformation models.* PhD thesis, The Chinese University of Hong Kong, 2005.

[44] J. Huang. Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568, 1996.

[45] J. Huang and J. A. Rossini. Sieve estimation for the proportional-odds failure-time regression model with interval. *Journal of the American statistical Association*, 92(439):960–967, 1997.

[46] J. Huang and J. A. Wellner. Interval censored survival data: A review of recent progress. *Proceddings of the First Seattle Symposium in Biostatistics: Survival Analysis*, pages 123–169, 1997.

[47] C. Kooperberg and D. B. Clarkson. Harzard regression with interval censored data. *Biometrics*, 53(4):1485–1494, 1997.

[48] K. F. Lam and T. L. Leung. Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis*, 7(1):39–54, 2001.

[49] D. V. Lindley. The choice of variables in multiple regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1):31–66, 1968.

[50] R. L. Losey and J. C. Talbott. Back on the track with the efficient markets hypothesisauthor. *The Journal of Finance*, 35(4):1039–1043, 1980.

[51] W. B. Lu and H. Zhang. Variable selection for proportional odds model. *Statistics in Medicine*, 26(20):3771–3781, 2007.

[52] S. Murphy, A. Rossini, and A. van der Vaart. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):187–201, 1997.

[53] Z. X. Ni. *Misspecified general transformation model and general transformation model with mixed-effects.* PhD thesis, The Chinese University of Hong Kong, 2008.

[54] R. Oller, G. Gomez, and M. L. Calle. Interval censoring: model characterizations for the validity of the simplified likelihood. *The Canadian Journal of Statistics*, 32(3):315–326, 2004.

[55] A. N. Pettitt. Proportional odds models for survival data and estimates using ranks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(2):169–175, 1984.

[56] B. T. Polyak and A. B. Juditski. Acceleration of stochastic approximation by averaging. *Journal of Control Optimization*, 30(4):838–855, 1992.

[57] D. Rabinowitz, A. Tsiatis, and J. Aragon. Regression with interval censored data. *Biometrika*, 82(3):501–513, 1995.

[58] G. A. Satten. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, 83(2):355–370, 1996.

[59] S. G. Self and E. A. Grossman. Linear rank test for interval censored data with applicationto pbc levels in adipose tissue of transformationer repaire workers. *Biometrics*, 42(3):521–530, 1986.

[60] D. Sinha, M. H. Chen, and S. K. Ghosh. Bayesian analysis andmodel selection for interval-censored survival data. *Biometrics*, 55(2):585–590, 1999.

[61] J. G. Sun. Regression analysis of interval censored failure time data. *Statistics in Medicine*, 16(5):497–504, 1997.

[62] J. G. Sun. *Interval Censoring*, pages 2090–2095. John Wiley, first edition, 1998.

[63] J. G. Sun. *The statistical analysis of interval censored failure time data.* New York: Springer, 2006.

[64] J. G. Sun, L. Q. Sun, and C. Zhu. Testing the proportional odds model for interval-censored data. *Lifetime Data Analysis*, 13(1):37–50, 2007.

[65] M. C. Sung, J. E. V. Johnson, and A. C. Bruce. Searching for semi-strong form inefficiency in the uk racetrack betting market. In L. VaughanWilliams, editor, *Information Efficiency in Financial and Betting Markets*, pages 179–192. Cambridge: Cambridge University, 2005.

[66] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

[67] L. L. Thurstone. Rank order as a psychological method. *Journal of Experimental Psychology*, 14:187–201, 1927.

[68] R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1996.

[69] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),*, 58(1):267–288, 1996.

[70] H. S. Wang, R. Z. Li, and C. L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

[71] Y. Q. Wu. *General transformation model with censoring, time-varying covariates and covariates with measurement errors.* PhD thesis, The Chinese University of Hong Kong, 2008.

[72] S. Yang and R. L. Prentice. Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, 94(445):125–136, 1999.

[73] N. Younges and J. Lachin. Link-based models for survival data with interval censored data. *Biometrics*, 53(4):1199–1211, 1997.

[74] D. L. Zeng, J. W. Cai, and Y. Shen. Semiparametric additive risks model for interval-censored data. *Statistica Sinica*, 16:287–302, 2006.

[75] D. L. Zeng, Q. X. Chen, and J. Ibrahim. Gamma frailty transformation models for multivariate survival times. *Biometrika*, 96(2):277–291, 2009.

[76] D. L. Zeng and D. Y. Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):507–564, 2007a.

[77] D. L. Zeng and D. Y. Lin. Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102(477):167–180, 2007b.

[78] D. L. Zeng and D. Y. Lin. A generalized asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica*, 20(2):871–910, 2010.

[79] H. Zhang and W. B. Lu. Adaptive lasso for coxs proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

[80] Z. G. Zhang and J. G. Sun. Interval censoring. *Statistical Methods in Medical Reseach*, 19(1):53–70, 2010.

[81] Z. G. Zhang, L. Q. Sun, X. Q. Zhao, and J. G. Sun. Regression analysis of interval-censored failure time data with linear transformation models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 33(1):61–70, 2005.

[82] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.