Statistical mechanics approaches to high-dimensional survival analysis

by

Guotao Chu

B.S., Nankai University, China, 2013

M.S., University of Connecticut, 2015

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

With the advent of high-dimensional data, variable selection has become a key step in survival data analysis. Recently, a general class of model selection criteria for high-dimensional data, called the generalized information criterion, has been developed. However, the use of the non-convex penalty functions in the generalized information criterion results in high-dimensional non-convex optimization problems. While many works have been proposed, their focus is limited to the application of a convex surrogate approach, which cannot ensure the convergence to the global optimal model with respect to the generalized information criterion.

The objective of this dissertation is to develop new solutions to high-dimensional data challenges of survival analysis. To meet this goal, we develop a powerful framework for high-dimensional survival data analysis using the notion of statistical mechanics, which is one of the pillars of modern physics. The proposed methods in this dissertation are widely applicable to not only model fitting problems but also prediction problems. To investigate the performance of our proposed methods, simulation study and real data analysis are extensively implemented.

In Chapter 1, the background, existing obstacles, rationale, and motivation are discussed. In Chapter 2, we develop a new fast variable selection procedure using the idea of simulated annealing with some modifications. The proposed method allows for rapidly finding the global optimal model with respect to the generalized information criterion. In Chapter 3, we develop a new best predictive model selection method for high-dimensional survival modeling. The proposed method relies on the idea of the optimal Bayesian predictive model, called the median probability model. In Chapter 4, we develop a robust variable selection approach to high-dimensional survival regression models. It is motivated by the "sandwich" estimator and provides a way for finding the global optimal model when the model is misspecified.

Statistical mechanics approaches to high-dimensional survival analysis

by

Guotao Chu

B.S., Nankai University, China, 2013

M.S., University of Connecticut, 2015

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
Gyuhyeong Goh

Copyright

Abstract

With the advent of high-dimensional data, variable selection has become a key step in survival data analysis. Recently, a general class of model selection criteria for high-dimensional data, called the generalized information criterion, has been developed. However, the use of the non-convex penalty functions in the generalized information criterion results in high-dimensional non-convex optimization problems. While many works have been proposed, their focus is limited to the application of a convex surrogate approach, which cannot ensure the convergence to the global optimal model with respect to the generalized information criterion.

The objective of this dissertation is to develop new solutions to high-dimensional data challenges of survival analysis. To meet this goal, we develop a powerful framework for high-dimensional survival data analysis using the notion of statistical mechanics, which is one of the pillars of modern physics. The proposed methods in this dissertation are widely applicable to not only model fitting problems but also prediction problems. To investigate the performance of our proposed methods, simulation study and real data analysis are extensively implemented.

In Chapter 1, the background, existing obstacles, rationale, and motivation are discussed. In Chapter 2, we develop a new fast variable selection procedure using the idea of simulated annealing with some modifications. The proposed method allows for rapidly finding the global optimal model with respect to the generalized information criterion. In Chapter 3, we develop a new best predictive model selection method for high-dimensional survival modeling. The proposed method relies on the idea of the optimal Bayesian predictive model, called the median probability model. In Chapter 4, we develop a robust variable selection approach to high-dimensional survival regression models. It is motivated by the "sandwich" estimator and provides a way for finding the global optimal model when the model is misspecified.

Table of Contents

List of Figures

List of Tables

Acknowledgments

On my way of pursuing my doctoral degree, there have been numerous people who have encouraged and supported me. I would like to take this opportunity to express my sincere appreciation for everything they have done for me.

I would like to convey my special heartfelt gratitude to my major advisor, Dr. Gyuhyeong Goh, whose knowledge and experience have shown to be crucial in formulating my research topics and methodologies. His insightful words inspired me to strengthen my thinking and brought my work to a higher level. He is there at all times to assist me with patient guidance. Without his help, it is impossible to complete this long journey.

I also want to express my profound gratitude to my committee members, Dr. Christopher I. Vahl, Dr. Jingru Mu, Dr. Jisang Yu, and the outside chairperson, Dr. Natalia Cernic-chiaro, for their time and support in serving on my committee, as well as their insightful remarks throughout my preliminary exam and dissertation defense.

Many thanks to the faculty members in the department of statistics, from whom I have taken the fundamental statistics courses, which have equipped me with a distinctive point of view on statistics, and pushed forward my understanding of this discipline. Also, thanks to the faculty and staff members, they have instructed and supported me a lot throughout my time working as a Graduate Teaching Assistant in our department.

Last but not least, I would like to express my heartfelt appreciation to my parents, who are always there to support me both emotionally and financially. Their love, unwavering encouragement, and support have been invaluable in assisting me in completing my Ph.D. degree.

Chapter 1

Introduction

In life science studies, including health science research, survival data analysis plays a particularly important role in modeling the relationship between living status and covariates. Owing to the rapid advances in data processing technologies, high-dimensional data are receiving increasing attention from data scientists. Under the context of high-dimensional regression modeling, how to extract valuable information from a large number of covariates becomes a challenging problem in survival data analysis. For example, genetics data are frequently employed in cancer prognostic research. While the data provides hundreds of thousands of measurement results of genetic marks, how to identify important variables that are closely related to the survival time is technically challenging. To tackle high-dimensional problems, many methodologies have been proposed. Although these innovations are beneficial in certain ways, they possess significant drawbacks from other viewpoints. In this chapter, we will discuss the present hurdles and motivation for our dissertation.

1.1 Challenges of high-dimensionality

One of the major goals of high-dimensional survival analysis is to identify relevant covariates that are closely connected to survival time. In order to determine the best model, various model selection criteria have been developed using a L_0-norm penalty function. For exam-

ple, the Bayesian information criterion (BIC) (Schwarz et al., 1978) and Akaike information criterion (AIC) (Akaike, 1974) are the most popular choices for traditional regression setting where the sample size is much larger than the number of covariates. For high-dimensional model selection, Chen and Chen (2008) propose a modified version of BIC, called the extended BIC, to consistently select the true data generating model over large model spaces. When the number of predictors is relatively small, finding the optimum model can be easily implemented by the best subset selection algorithm, which finds the best model by evaluating all possible candidate models using a model selection criterion. However, when we have high-dimensional data, the best subset selection algorithm can be computationally expensive and time-consuming. For example, when $p = 1000$, we need to evaluate $2^p \approx 10^{301}$ candidate models.

To reduce the heavy computational burden in a high-dimensional variable selection problem, penalized likelihood estimation with a convex surrogate penalty has been proposed in the literature. For example, Goeman (2010); Tibshirani (1997); Zhang and Lu (2007) develop penalized likelihood estimation methods for high-dimensional Cox proportional hazards regression with convex penalties such as lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), and elastic net (Zou and Hastie, 2005). However, these convex approximation approaches cannot ensure the convergence to the global optimum of the model selection criterion since the solution path is generated by only a finite sequence of tuning parameters. In other words, due to the limited coverage of the tuning parameter values, there is a high chance that the solution path has missed the global solution.

1.2 Challenges of non-convexity

Another obstacle to finding the best model comes from the non-convexity in the model selection criterion. The L_0-norm penalty function is commonly included in the model selection criterion, and it has the non-convex property in nature. The non-convexity nature of the L_0-norm part leads to the difficulty of optimization as none of the convex optimization algorithms provides a feasible solution. We notice that the simulated annealing algorithm

(Kirkpatrick et al., 1983) is proposed for the global non-convex optimization problem in thermodynamics, with a key idea of conducting a stochastic search to avoid the chance to get stuck in a local optimum. This algorithm utilizes the Metropolis-Hastings sampling method (Metropolis et al., 1953) for generating a Markov chain with a stationary distribution whose mode is the same as the global optimum of the target function. Despite the fact that simulated annealing assures the convergence to the global optimum, the slow convergence rate and the need to choose the proposal distribution are regarded as major drawbacks. This has resulted in the application of simulated annealing for high-dimensional variable selection being impractical on a computational level due to the fact that its computing efficiency decreases substantially as the number of covariates grows.

1.3 Challenges of predictive model selection

Best predictive model selection is also of great importance for survival regression analysis. Predicting life expectancy can be challenging but beneficial to many disease studies. The problem of best predictive model selection has been extensively studied by researchers based on some parametric models with a relatively small number of covariates. However, in the high-dimensional data scenario, identifying the best predictive model is theoretically and computationally challenging. The critical challenge is that choosing the best fitting model is not the same as finding the best predictive model, since the best fitting model cannot guarantee the best prediction performance. In addition, the computational cost for high-dimensional model selection is a pressing challenge. Hence, there is a strong need for the development of predictive model selection with high-dimensional survival data.

1.4 Problem of model misspecification

For survival analysis, a variety of semi-parametric and parametric models, such as the Cox proportional hazards model (Cox, 1972) and the accelerated failure time (AFT) model (Wei, 1992), are extensively developed. In spite of the fact that these models are widely used,

they require some structural assumptions to guarantee some important properties, including consistency and asymptotic normality. However, in the real world, such model assumptions can be easily violated, often referred to as the problem of model misspecification. To address this issue, the robust inference methods for survival analysis have been proposed by using the idea of "sandwich" estimator, which serves as a proper variance estimator for misspecified models (Gail et al., 1984; Lagakos et al., 1984; Lagakos, 1988; Morgan et al., 1986; O'neill, 1986; Solomon, 1984; Struthers and Kalbfleisch, 1986). Although the model misspecification problems have been extensively studied for inference, there has been no attempt to address model misspecification issues for model selection.

1.5 Motivation and outline of the dissertation

Our motivation stems from our desire to address the challenges that have been mentioned above. This dissertation covers a wide range of model selection problems for high-dimensional survival data analysis. In this dissertation, we aim to develop innovative strategies that can assist us in achieving the following specific objectives. First, we aim to develop a fast algorithm that provides an effective way of finding the global optimum model for the generalized information criterion. Second, we aim to introduce a new method of best predictive model selection for high-dimensional survival data. Third, we aim to develop a robust model selection procedure in the presence of model misspecification. The structure for the remainder of the dissertation is as follows.

In Chapter 2, we develop a global optimal model selection method for determining the model that optimizes the generalized model selection criterion. The proposed method is originally inspired by simulated annealing (Kirkpatrick et al., 1983), which is widely used for energy optimization in the field of statistical physics. The key idea of our proposed method is to incorporate Gibbs sampling into the framework of simulated annealing by utilizing the notion of Boltzmann distribution in statistical mechanics (Gibbs, 1902). The proposed algorithm enables us to perform a faster and more stable probabilistic search than the traditional simulated annealing algorithm. The simulation study shows that our proposed

algorithm works well with high-dimensional survival data. In addition, the proposed method is applied to blood cancer data.

In Chapter 3, we propose a new way of determining the best predictive model with high-dimensional survival data. Our proposed method is inspired by the median probability model, which is originally proposed by Barbieri et al. (2004) for optimal predictive Bayesian model selection. The key idea of our proposed method is to incorporate the median probability model into a frequentist framework via the concept of Boltzmann distribution. Simulation study and real data analysis are implemented to demonstrate the performance of our proposed method.

In Chapter 4, we extend the problem of model selection to a more general scenario, in which a model misspecification problem occurs, and develop a robust variable selection approach to high-dimensional survival regression models. The proposed method is motivated by the so-called "sandwich" estimator, which is a robust variance estimator under a misspecified model. The key idea is to use the sandwich estimator to construct a robust model selection criterion. Using the simulated annealing scheme, we introduce a new algorithm that finds the global optimal model for the proposed model selection criterion.

In Chapter 5, concluding remarks including extensions and limitations are discussed.

Chapter 2

Global optimal model selection for high-dimensional survival analysis

2.1 Introduction

In high-dimensional survival analysis, a primary goal is to identify relevant covariates that are related to the survival time. Various model selection criteria have been developed using a L_0-norm penalty function. For example, the Bayesian information criterion (BIC) (Schwarz et al., 1978) and Akaike information criterion (AIC) (Akaike, 1974) are the most popular choices for classical regression modeling. In a high-dimensional regression setting, Chen and Chen (2008) propose a modified version of BIC, called the extended BIC, to consistently select the true data-generating model over large model spaces.

Since the use of L_0-norm penalty yields a non-convex optimization problem, finding the best model, which is the global optimum of the model selection criterion, is computationally expensive and time-consuming in a high-dimensional data setting in which the number of candidate covariates is large. To reduce the heavy computational burden in a high-dimensional variable selection problem, penalized partial-likelihood estimation with a convex surrogate penalty has been proposed in the literature. For example, penalized likelihood estimation methods with convex penalties (e.g. lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006),

and elastic net (Zou and Hastie, 2005)) are developed for high-dimensional Cox proportional hazards regression (Goeman, 2010; Tibshirani, 1997; Zhang and Lu, 2007). However, these approaches do not ensure the convergence to the global optimum of the model selection criterion since the solution path is generated by only a finite sequence of tuning parameters. In other words, due to the limited coverage of the tuning parameter values, there is a high chance that the solution path has missed the global solution to the model selection criterion.

In thermodynamics, Kirkpatrick et al. (1983) proposes a global optimization algorithm, called simulated annealing. The key idea of simulated annealing is to perform a stochastic search so that we can avoid the chance of getting stuck in a local optimum. Using the Metropolis-Hastings sampling (Metropolis et al., 1953), the move of simulated annealing generates a Markov chain with a stationary distribution whose mode is the same as the global optimum of the target function. Although simulated annealing assures the convergence to the global optimum in a non-convex optimization framework, the slow convergence and the choice of the proposal distribution are major drawbacks. As a result, the application of simulated annealing for high-dimensional variable selection is computationally infeasible since its computational efficiency drops dramatically as the number of covariates increases.

In this chapter, we propose a new global optimization method for high-dimensional survival model selection with a general class of model selection criteria, often referred to as generalized information criterion (Atkinson, 1980; Kim et al., 2012; Zhang et al., 2010). The key idea of the proposed method is to incorporate Gibbs sampling into a simulated annealing framework via the concept of Boltzmann distribution in statistical mechanics. The proposed method enables us to perform a probabilistic search using the Gibbs sampler, which leads to the fast and stable convergence to the target distribution (Casella and George, 1992). In addition, the use of the Gibbs sampler yields that the probability of accepting the move to a new model always becomes one so that it automatically eliminates the issue of proposal distribution selection in the traditional simulated annealing algorithm. The Cox proportional hazards model is employed as the assumed model for explanation, which is discussed in Chapter 2.2. In Chapter 2.3, the penalized likelihood methods, which serve as ways of the convex surrogate are explained with their drawbacks. The technical details of the proposed

method are given in Section 2.4. As shown in Section 2.5, our proposed method outperforms many existing methods. The real data analysis in Section 2.6 also demonstrates the applicability of the proposed method to a blood cancer study.

2.2 Basic setup and generalized information criterion

For subject $i \in \{1, \ldots, n\}$, let $T_i = \min(T_i^*, C_i)$ be the observed failure time and $\mathbf{x}_i$ be the p-dimensional vector of possible covariates, where T_i^* is the actual death time of the i-th individual and C_i is the censoring time. Denote by $\delta_i = I\{T_i < C_i\}$ the indicator of the occurrence of the event, where $I\{\cdot\}$ represents an indicator function.

In survival analysis, the Cox proportional hazards model (Cox, 1972) is a widely used semiparametric regression model. The Cox model provides a way to examine how the covariates are associated with the rate of a particular event happening (e.g., death) at time t, where the rate is referred to as the hazard rate. Specifically, the Cox model explains the relationship between the hazard function and the covariates by assuming the form

$$h(t) = h_0(t) \times \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \tag{2.1}$$

where $h(t)$ is the hazard function at time t, $h_0(t)$ is the baseline hazard function, which reflects the underlying hazard for subjects with all covariates equal to 0, $\mathbf{x}$ is the p-dimensional vector of covariates, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the p-dimensional coefficient vector, which is of our primary interest.

Under the right-censored scenario with observations $\{(T_i, \mathbf{x}_i, \delta_i), i = 1, ..., n\}$, the regression parameter $\boldsymbol{\beta}$ in the Cox model (2.1) can be estimated by constructing the partial likelihood without imposing a distributional assumption on the data. In a survival analysis

framework, the likelihood function is given as

$$\begin{aligned} L(\boldsymbol{\beta}) &= \left\{ \prod_{i:\delta_i=1} f(T_i \mid \boldsymbol{\beta}) \right\} \times \left\{ \prod_{i:\delta_i=0} [1 - F(T_i \mid \boldsymbol{\beta})] \right\} \\ &= \prod_{i=1}^{n} f(T_i \mid \boldsymbol{\beta})^{\delta_i} S(T_i \mid \boldsymbol{\beta})^{1-\delta_i}, \\ &= \prod_{i=1}^{n} \left[\frac{f(T_i \mid \boldsymbol{\beta})}{S(T_i \mid \boldsymbol{\beta})} \right]^{\delta_i} S(T_i, \mid \boldsymbol{\beta}), \end{aligned}$$

where $f(\cdot \mid \boldsymbol{\beta})$ is the probability density function for time T_i given parameter $\boldsymbol{\beta}$, $F(\cdot \mid \boldsymbol{\beta})$ is the cumulative density function, and $S(\cdot \mid \boldsymbol{\beta})$ is the survival function. Under the Cox proportional hazards assumption (2.1), the likelihood function can be expressed as

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left[\frac{f(T_i \mid \boldsymbol{\beta})}{S(T_i \mid \boldsymbol{\beta})} \right]^{\delta_i} S(T_i, \mid \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \left[\frac{h(T_i \mid \mathbf{x}_i, \boldsymbol{\beta})}{\sum_{l \in R(T_i)} h(T_i \mid \mathbf{x}_l, \boldsymbol{\beta})} \right]^{\delta_i} \left[\sum_{l \in R(T_i)} h(T_i \mid \mathbf{x}_l, \boldsymbol{\beta}) \right]^{\delta_i} S(T_i \mid \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \left[\frac{h_0(T_i) \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})}{\sum_{l \in R(T_i)} h_0(T_i) \exp(\mathbf{x}_l^{\mathrm{T}} \boldsymbol{\beta})} \right]^{\delta_i} \left[\sum_{l \in R(T_i)} h(T_i \mid \mathbf{x}_l, \boldsymbol{\beta}) \right]^{\delta_i} S(T_i \mid \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})}{\sum_{l \in R(T_i)} \exp(\mathbf{x}_l^{\mathrm{T}} \boldsymbol{\beta})} \right]^{\delta_i} \left[\sum_{l \in R(T_i)} h(T_i \mid \mathbf{x}_l, \boldsymbol{\beta}) \right]^{\delta_i} S(T_i \mid \boldsymbol{\beta}), \end{aligned} \tag{2.2}$$

where $R(T_i)$ is the risk set at time T_i, which represents the number of individuals who survived at least until time T_i. To complete the specification of the likelihood function, we need to further assume a parametric form of the baseline hazard function $h_0(\cdot)$. However, without any prior knowledge about the data, it is dangerous to assume a specific form of the baseline hazard function. As an alternative, Cox (1972) employs a partial likelihood approach in which the baseline hazard function can be completely unspecified. Note that the first term in the product in (2.2) contains the most information about $\boldsymbol{\beta}$, while the last two terms have the information about the baseline hazard function. Hence, by treating the last two terms as a constant with respect to $\boldsymbol{\beta}$, the partial likelihood function can be defined

by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\sum_{l \in R(T_i)} \exp(\mathbf{x}_l^{\mathrm{T}}\boldsymbol{\beta})} \right]^{\delta_i},$$

and the corresponding partial log-likelihood function can be derived as

$$\begin{aligned} l(\boldsymbol{\beta}) &= \log L(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \log \left\{ \sum_{l \in R(T_i)} \exp(\mathbf{x}_l^{\mathrm{T}}\boldsymbol{\beta}) \right\} \right]. \end{aligned} \tag{2.3}$$

Under a low-dimensional regression setting (i.e., $p \ll n$), it is well known that the asymptotic normality holds for the maximum partial likelihood estimator, which is obtained by maximizing (2.3). However, when the number of covariates p is large, variable selection is necessary to eliminate irrelevant covariates from the model so that the useful asymptotic property can be achieved under the reduced model.

In a high-dimensional Cox regression setting, the best model can be identified by minimizing the penalized partial log-likelihood as follows:

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta}),$$

where $\mathrm{pen}(\boldsymbol{\beta})$ is a penalty function which increases as the number of parameters increases. In the model selection literature, the penalty function is commonly assumed to be a linear function of L_0-norm, that is,

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_0, \tag{2.4}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} I\{\beta_j \neq 0\}$ denotes the L_0-norm and λ is a prespecified tuning parameter controlling the degrees of penalization. The model selection criterion in (2.4) is often referred to as generalized information criterion (GIC). According to the choice of λ, GIC reduces to a well-known model selection criterion. For example, when we set $\lambda = \ln(n_0)$, GIC becomes

BIC for censored survival models (Volinsky and Raftery, 2000), where $n_0 = \sum_{i=1}^{n} \delta_i$, which denotes the total number of uncensored failure events. With $\lambda = 2$, GIC reduces to AIC.

Let $\boldsymbol{s} = (s_1, \ldots, s_p)$ represent a reduced Cox model such that $\beta_j \neq 0$ if $s_j = 1$ and $\beta_j = 0$ if $s_j = 0$ for $j = 1, \ldots, p$. Given $\boldsymbol{s}$, let $\boldsymbol{\beta}(\boldsymbol{s})$ be the sub-vector of $\boldsymbol{\beta}$ corresponding to one elements in $\boldsymbol{s}$. Then, the form of GIC in (2.4) can be further generalized as follows:

$$\text{GIC}(\boldsymbol{s}) = -2l(\hat{\boldsymbol{\beta}}(\boldsymbol{s})) + \text{pen}(|\boldsymbol{s}|), \tag{2.5}$$

where $|\boldsymbol{s}| = \sum_{j=1}^{p} s_j$ denotes the number of parameters under model $\boldsymbol{s}$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ is the maximum partial likelihood estimate of $\boldsymbol{\beta}(\boldsymbol{s})$ under model $\boldsymbol{s}$. Note that, when $\text{pen}(|\boldsymbol{s}|) = \lambda|s|$, (2.5) reduces to the original form of GIC in (2.4). Throughout this chapter, GIC refers to the form of (2.5).

Due to the non-convexity of the penalty function in (2.5), model selection with GIC should be performed by a brute-force algorithm (i.e., by comparing all possible models). However, when p is large, finding the best model using a brute-force algorithm, often called best subset selection in the statistical literature (James et al., 2013), becomes a NP-hard (non-deterministic polynomial-time hard) problem. For example, when $p = 100$, we need to compare $2^{100} \approx 1.27^{30}$ candidate models.

2.3 Convex surrogate

In recent high-dimensional regression research, sparse estimation with convex penalties has been extensively studied. For example, using the L_1-norm penalty, Tibshirani (1996) proposes the lasso (least absolute shrinkage and selection operator),

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1,$$

where $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta|_j$ and $\lambda \geq 0$. To improve the statistical efficiency of lasso, Zou (2006) develops the adaptive lasso by using a weighted L_1-norm as follows:

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $w_1, \ldots, w_p$ are data-driven weights. For high-dimensional and correlated data, Zou and Hastie (2005) propose the elastic net,

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||^2,$$

where $||\boldsymbol{\beta}||^2 = \sum_{j=1}^{p} \beta_j^2$ and λ_1 and λ_2 are the non-negative tuning parameters. Note that the elastic net includes the lasso as a special case with $\lambda_2 = 0$. Since these penalties lead to not only the convexity of the objective function but also the sparse estimates of $\boldsymbol{\beta}$, they have been considered as a solution to high-dimensional variable selection. However, unlike GIC, the tuning parameters are unknown in the penalized likelihood estimation framework, and they must be chosen by a model selection criterion. From this aspect, the penalized likelihood estimation with GIC tuning parameter selection can be considered as a convex surrogate of the GIC model selection. This model selection procedure can be summarized as follows:

- Step 1: Define a sequence of values for tuning parameter λ, Λ. For example, $\Lambda = \{\lambda_t = \epsilon(t-1) : t = 1, \ldots, T\}$ for small $\epsilon > 0$.
- Step 2: For each value of $\lambda \in \Lambda$, compute the sparse estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}(\lambda)$, by minimizing the penalized likelihood objective function given λ. For example, $\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1$ for lasso.
- Step 3: Let $\boldsymbol{s}_\lambda = \{(s_1, \ldots, s_p) : s_j = I\{\hat{\beta}_j(\lambda) \neq 0\}, j = 1, \ldots, p\}$ be the reduced model given $\hat{\boldsymbol{\beta}}(\lambda)$. Compute $\text{GIC}(\boldsymbol{s}_\lambda)$ for $\lambda \in \Lambda$.
- Step 4: Find the best model by $\min_{\lambda \in \Lambda} \text{GIC}(\boldsymbol{s}_\lambda)$.

Note that although the above procedure is computationally efficient and fast, there is a main limitation that the best model is generally a local optimum, not the global optimum, because the solution path has been generated by a finite sequence of λ-values. In the following section, we introduce our proposed solution to global optimum model selection with GIC.

2.4 Model selection via stochastic search

In this section, we introduce a new method to find the global optimum model using GIC. Our proposed method is motivated by the idea of simulated annealing, which is a popular global optimization algorithm in statistical mechanics.

2.4.1 Simulated annealing

Simulated annealing (SA), originally proposed by Kirkpatrick et al. (1983), is a stochastic optimization method for finding the global optimum in a non-convex optimization problem. The technique mimics the process of annealing in metallurgy, which is a technique involving the heating and cooling of a material to increase the size of its crystals and reduce its defects. Let $E(\boldsymbol{s})$ be a energy function at state $\boldsymbol{s}$. In general, SA is used to find the state that leads to a global minimum energy.

In statistical thermodynamics, the probability of a physical system being in the state $\boldsymbol{s}$ with energy $E(\boldsymbol{s})$ at temperature τ can be described by the Boltzmann distribution (Gibbs, 1902),

$$p_\tau(\boldsymbol{s}) \propto \exp\left\{-\frac{E(\boldsymbol{s})}{\kappa\tau}\right\} \tag{2.6}$$

where κ is the Boltzmann's constant, which is usually a known constant in SA. Using the Boltzmann distribution, SA always converges to the global optimum by performing a stochastic search with annealing. The detailed SA algorithm is given below:

- Step 1: Set an initial state $\boldsymbol{s} = \boldsymbol{s}_0$ and an initial temperature $\tau = \tau_0$.
- Step 2:

(a) Draw a new state $\boldsymbol{s}^*$ from a proposal distribution $q(\boldsymbol{s}^* \mid \boldsymbol{s})$, which represents the conditional probability of $\boldsymbol{s}^*$ given the current value of $\boldsymbol{s}$.

(b) Move to the new state $\boldsymbol{s}^*$ with probability

$$\min\left\{1, \frac{p_\tau(\boldsymbol{s}^*)q(\boldsymbol{s} \mid \boldsymbol{s}^*)}{p_\tau(\boldsymbol{s})q(\boldsymbol{s}^* \mid \boldsymbol{s})}\right\}. \tag{2.7}$$

(c) Repeat step (a) and step (b) until the chain is reached an equilibrium state.

- Step 3: Decrease the temperature by $\tau = \tau - \epsilon$ for small $\epsilon > 0$. If $\tau \leq 0$, then terminate. Otherwise, go to Step 2.

In the context of GIC optimization, we can employ SA by replacing the energy function $E(\boldsymbol{s})$ with $\mathrm{GIC}(\boldsymbol{s})$. Then, the Boltzmann distribution can be obtained by

$$p_\tau(\boldsymbol{s}) \propto \exp\left\{-\frac{\mathrm{GIC}(\boldsymbol{s})}{\kappa\tau}\right\}. \tag{2.8}$$

For the remainder of this chapter, without loss of generality, we assume $\kappa = 2$. It is important to note that Step 2 in SA comes from the idea of Metropolis-Hastings sampling, which is a Markov chain Monte Carlo (MCMC) method for sampling from a probability distribution when direct sampling is difficult. However, the specification of proposal distribution $q(\cdot \mid \cdot)$ is cumbersome in our setting. An even more serious problem is that the acceptance probability in (2.7) tends to decrease exponentially as the number of covariates, p, increases. As a result, the probability of moving to a new state can be extremely small in a high-dimensional variable selection case.

To demonstrate, we perform a simulation study as follows: First, we generate artificial survival times of 100 subjects, $T_1^*, \ldots, T_{100}^*$, by $T_i^* = -\frac{\log(U_i)}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}$, where $U_i \overset{iid}{\sim} \mathrm{Unif}(0,1)$, $\mathbf{x}_i \overset{iid}{\sim} \mathrm{N}_p(0,\Sigma)$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$, and $\boldsymbol{\beta} = (0, -0.7, -0.7, 0, \ldots, 0)^{\mathrm{T}}$. Second, we create censoring time C_i by generating a sample from $\mathrm{Exp}(0.58)$, which produces about 40% censoring rate. Then, given the simulated data, we proceed with the SA algorithm with $\tau = 1$ and count the acceptance rate of moving to a new state as follows:

- Step 1: Set $\boldsymbol{s} = (0, \ldots, 0)$, $\tau = 1$, and $a = 0$ (count for a new move).
- Step 2:
  (a) Generate $\boldsymbol{s}^* = (s_1^*, s_2^*, ..., s_p^*)$ with $s_j^* \overset{iid}{\sim} \text{Ber}(0.5)$ for $j = 1, \ldots, p$.
  (b) Calculate $\Delta(\boldsymbol{s}, \boldsymbol{s}^*) = \exp\left\{-\frac{\text{GIC}(\boldsymbol{s}^*) - \text{GIC}(\boldsymbol{s})}{\kappa\tau}\right\}$, where GIC is chosen to be EBIC defined in (2.10).
  (c) Generate $u \sim \text{Unif}(0, 1)$.
  (d) If $\Delta(\boldsymbol{s}, \boldsymbol{s}^*) \geq u$, set $\boldsymbol{s} = \boldsymbol{s}^*$ and $a = a + 1$. Otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.
- Step 3: Repeat Step 2 for 5000 times.

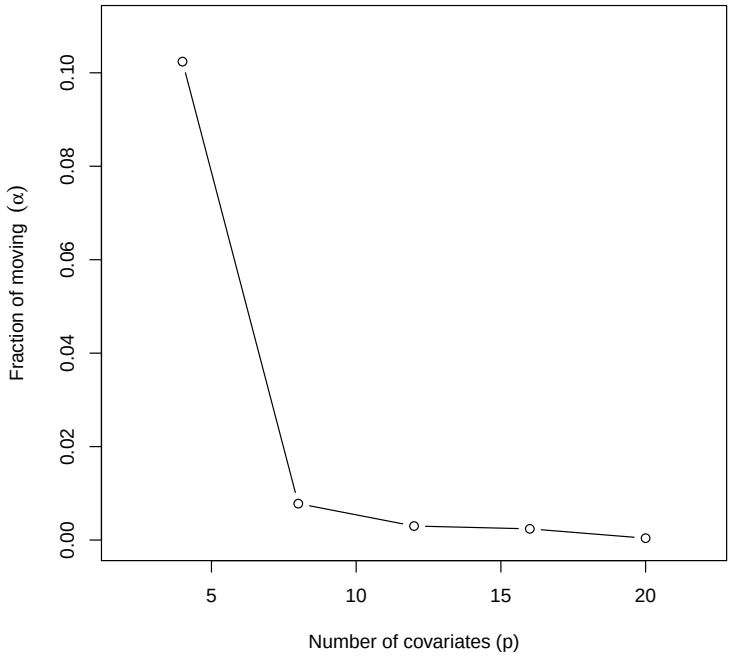

Figure 2.1: *Changes in the acceptance rate of moving to a new state in SA when the number of covariates p increases.*

Figure 2.1 displays our simulation result. It clearly shows that the acceptance rate of moving to a new state, $\alpha = a/5000$, drops dramatically as p increases. In particular, when

$p = 20$, there is almost no chance that the current state moves to a new state. In this case, SA cannot converge to the global optimum even if the iteration number is extremely large. To address the limitations of SA for high-dimensional variable selection, we propose a new stochastic search algorithm using the idea of Gibbs sampler. The details are discussed in the next section.

2.4.2 Proposed method

Motivated by the idea of Gibbs sampling, we propose to generate a candidate model for the next move in SA by adding a new predictor to or deleting one from the current model. To this end, let $\boldsymbol{s} = (s_1, ..., s_p)^{\mathrm{T}} \in \mathbb{R}^p$ be the current state. For a given j, we define a candidate model by $\boldsymbol{s}^* = (s_1^*, \ldots, s_p^*)$ such that $s_k^* = s_k$ if $k \neq j$. Then, we can obtain the following important property.

Lemma 1. *Let $p_\tau(\cdot)$ be the Boltzmann distribution defined in (2.8). Assume that the proposal distribution, $q(\boldsymbol{s}^* \mid \boldsymbol{s})$, in SA is proportional to $p_\tau(\boldsymbol{s}^*)$ with respect to s_j^*, that is, $q(\boldsymbol{s}^* \mid \boldsymbol{s}) \propto p_\tau(\boldsymbol{s}_j^*)$ with respect to s_j^*. Then, in Step 2 of SA, $\boldsymbol{s}^*$ is accepted with probability one.*

Proof of Lemma 1. Since $q(\boldsymbol{s}^* \mid \boldsymbol{s}) \propto p_\tau(\boldsymbol{s}_j^*)$ with respect to s_j^*, it can be viewed as $q(\boldsymbol{s}^* \mid \boldsymbol{s}) \propto p_\tau(s_j^* \mid \boldsymbol{s}_{-j}^*)$, where $\boldsymbol{s}_{-j}$ is obtained by deleting the j-th component of $\boldsymbol{s}$. Recall that by the definition of $\boldsymbol{s}^*$, we have $\boldsymbol{s}_{-j}^* = \boldsymbol{s}_{-j}$. Then, it follows that $q(\boldsymbol{s}^* \mid \boldsymbol{s}) = p_\tau(s_j^* \mid \boldsymbol{s}_{-j}) = p_\tau(s_j^* \mid \boldsymbol{s}_{-j}^*)$. Similarly, it can be shown that $q(\boldsymbol{s} \mid \boldsymbol{s}^*) = p_\tau(s_j \mid \boldsymbol{s}_{-j}^*) = p_\tau(s_j \mid \boldsymbol{s}_{-j})$. This implies that

$$\begin{aligned}
\frac{p_\tau(\boldsymbol{s}^*)q(\boldsymbol{s} \mid \boldsymbol{s}^*)}{p_\tau(\boldsymbol{s})q(\boldsymbol{s}^* \mid \boldsymbol{s})} &= \frac{p_\tau(\boldsymbol{s}^*)p_\tau(s_j \mid \boldsymbol{s}_{-j}^*)}{p_\tau(\boldsymbol{s})p_\tau(s_j^* \mid \boldsymbol{s}_{-j})} \\
&= \frac{p_\tau(\boldsymbol{s}^*)p_\tau(s_j \mid \boldsymbol{s}_{-j})}{p_\tau(\boldsymbol{s})p_\tau(s_j^* \mid \boldsymbol{s}_{-j}^*)} \\
&= \frac{p_\tau(\boldsymbol{s}_{-j}^*)}{p_\tau(\boldsymbol{s}_{-j})} \\
&= 1,
\end{aligned}$$

where the last equality holds from the fact that $\boldsymbol{s}_{-j} = \boldsymbol{s}_{-j}^*$. This completes our proof. □

Now, we define the proposal distribution by

$$q(\boldsymbol{s}^* \mid \boldsymbol{s}) = \frac{\exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(\boldsymbol{s}^*)\right\} I\left\{\boldsymbol{s}^*_{-j} = \boldsymbol{s}_{-j}\right\}}{\exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(s^*_j = 1, \boldsymbol{s}^*_{-j})\right\} + \exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(s^*_j = 0, \boldsymbol{s}^*_{-j})\right\}}.$$

In SA, one of the key features is the annealing process, that is, the temperature τ decreases as the iteration number increases. Unlike SA, in our proposed method, we consider increasing the temperature for the following reason. Under the given temperature, the selected best model can be either a global optimum or a local optimum. If the best model is obtained at a local optimum, increasing the temperature will improve the chance to get out from the local trap and move forward to the global optimum. If the current best model is attained at the global optimum, then the best model continuously remains the same as τ increases. Hence, in this case, we can conclude the convergence to the global optimum.

Given the GIC we want to optimize, our proposed method works in the following way:

- Step 1: Start from an initial state of $\boldsymbol{s} = (s_1, s_2, ..., s_p)$ with an initial temperature $\tau = \tau_0$, use $\hat{\boldsymbol{s}} = (\hat{s}_1, \cdots, \hat{s}_p)$ to store the best model, set $r = 0$, which counts the number of iterations, and set an initial value for k, which controls the maximum number of covariates selected in the model.

- Step 2: Implement Gibbs sampler to generate a Markov chain and update $\boldsymbol{s}$ by iterating the following procedure:

  - Generate $\boldsymbol{s}^* = (s^*_1, s^*_2, \cdots, s^*_p)$ by setting $s^*_1 = 1 - s_1$ and $s^*_l = s_l$ for $l \neq 1$.

    If $\sum_{j=1}^{p} s^*_j > k$, skip the following and move to update s_2, otherwise, calculate $\mathrm{GIC}(\boldsymbol{s}^*)$.

    If $\mathrm{GIC}(\boldsymbol{s}^*) < \mathrm{GIC}(\hat{\boldsymbol{s}})$, update $\hat{\boldsymbol{s}} = \boldsymbol{s}^*$ and set $r = 0$, $\tau = \tau_0$. Otherwise, set $r = r + 1$.

    Generate a Bernoulli trial with success probability w defined in 2.9. If we obtain 1, update $\boldsymbol{s} = \boldsymbol{s}^*$, otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.

  - Generate $\boldsymbol{s}^* = (s^*_1, s^*_2, \cdots, s^*_p)$ by setting $s^*_2 = 1 - s_2$ and $s^*_l = s_l$ for $l \neq 2$.

If $\sum_{j=1}^{p} s_j^* > k$, skip the following and move to update s_3, otherwise, calculate GIC($\boldsymbol{s}^*$).

If GIC($\boldsymbol{s}^*$) < GIC($\hat{\boldsymbol{s}}$), update $\hat{\boldsymbol{s}} = \boldsymbol{s}^*$ and set $r = 0$, $\tau = \tau_0$. Otherwise, set $r = r + 1$.

Generate a Bernoulli trial with success probability w defined in 2.9. If we obtain 1, update $\boldsymbol{s} = \boldsymbol{s}^*$, otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.

⋮

– Generate $\boldsymbol{s}^* = (s_1^*, s_2^*, \cdots, s_p^*)$ by setting $s_p^* = 1 - s_p$ and $s_l^* = s_l$ for $l \neq p$.

If $\sum_{j=1}^{p} s_j^* > k$, skip the following and move to update s_1, otherwise, calculate GIC($\boldsymbol{s}^*$).

If GIC($\boldsymbol{s}^*$) < GIC($\hat{\boldsymbol{s}}$), update $\hat{\boldsymbol{s}} = \boldsymbol{s}^*$ and set $r = 0$, $\tau = \tau_0$. Otherwise, set $r = r + 1$.

Generate a Bernoulli trial with success probability w defined in 2.9. If we obtain 1, update $\boldsymbol{s} = \boldsymbol{s}^*$, otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.

The success probability in the Bernoulli trial is defined as

$$w = \frac{\exp\{-\frac{1}{\kappa\tau}\text{GIC}(\boldsymbol{s}^*)\}}{\exp\{-\frac{1}{\kappa\tau}\text{GIC}(\boldsymbol{s}^*)\} + \exp\{-\frac{1}{\kappa\tau}\text{GIC}(\boldsymbol{s})\}}. \tag{2.9}$$

- Step 3: Repeat Step 2 until $r > pm$ (m has a prespecified value).
- Step 4: Repeat Step 2 and 3 with $\boldsymbol{s} = \hat{\boldsymbol{s}}$, $r = 0$ for a sequences of values of $\tau = \{\tau_2, \tau_3, \cdots, \tau_{\max}\}$, where $\tau_{t+1} > \tau_t$, until $\tau = \tau_{\max}$. $\tau_{\max}$ is the maximum temperature with a prespecified value.

The final model $\hat{\boldsymbol{s}}$ from the above procedure will be the estimated global optimal model, with the corresponding estimated value of model selection criterion GIC($\hat{\boldsymbol{s}}$). The proposed algorithm is summarized in Algorithm 1.

Algorithm 1 Global optimal model selection

Start from an initial state of $\boldsymbol{s} = (s_1, s_2, ..., s_p)$ with $\tau = \tau_0$, use $\hat{\mathbf{s}} = (\hat{s}_1, \hat{s}_2..., \hat{s}_p)$ to store the best model, set $r = 0$, and define k to control the maximum model size. The algorithm proceeds as follows:

- Step 1: For $j = 1, \cdots, p$, update s_j by repeating the following steps until $r > pm$, where m is set to control the number of iterations.

  (a) Define $\boldsymbol{s}^*$ by $s_j^* = 1 - s_j$ and $s_\ell^* = s_\ell$ for $\ell \neq j$.

  (b) If $\sum_{i=1}^p s_i^* > k$, skip Steps (c)–(d) below and jump to the next update for $j + 1$. Otherwise, calculate $\text{GIC}(\boldsymbol{s}^*)$.

  (c) If $\text{GIC}(\boldsymbol{s}^*) < \text{GIC}(\hat{\boldsymbol{s}})$, update $\hat{\mathbf{s}} = \boldsymbol{s}^*$ and reset $r = 0$ and $\tau = \tau_0$. Otherwise, set $r = r + 1$.

  (d) We update $\boldsymbol{s} = \boldsymbol{s}^*$ if we obtain 1 from a Bernoulli trial with the success probability

$$\omega = \frac{\exp\left\{-\frac{1}{\kappa\tau}\text{GIC}(\boldsymbol{s}^*)\right\}}{\exp\left\{-\frac{1}{\kappa\tau}\text{GIC}(\boldsymbol{s}^*)\right\} + \exp\left\{-\frac{1}{\kappa\tau}\text{GIC}(\boldsymbol{s})\right\}}.$$

  Otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.

- Step 2: Repeat Step 1 with $\boldsymbol{s} = \hat{\boldsymbol{s}}$, $r = 0$, and $\tau = \tau_{t+1}(> \tau_t)$ until $\tau = T_{\max}$, where $T_{\max}$ is a prespecified maximum temperature.

In Appendix A, we provide a demo `R` code for implementing Step 1 of the proposed algorithm when GIC is assumed to be BIC.

2.5 Simulation Study

In this section, we conduct a simulation study to investigate the performance of our proposed method under the Cox proportional hazards model. For the choice of GIC, we select the

extended BIC (EBIC), which is the most popular choice for high-dimensional model selection (Chen and Chen (2012), Luo et al. (2015), Foygel and Drton (2010)):

$$\text{EBIC}(\boldsymbol{s}) = -2l\left(\hat{\boldsymbol{\beta}}(\boldsymbol{s})\right) + \log(n_0)|\boldsymbol{s}| + 2\gamma\binom{n}{|\boldsymbol{s}|}, \tag{2.10}$$

where $\gamma \in [0,1]$ is a prespecified tuning parameter and we set $\gamma = 1$ so that EBIC always satisfies model selection consistency even when $p > n$ (Chen and Chen, 2008).

When we implement the proposed method using Algorithm 1, we set $\boldsymbol{s} = (0, 0, ..., 0)$, $k = 15$, $\tau \in \{1, 4/3\}$ and $m = 10$ for the initial setting, called 'proposed method (null)'. To perform sensitivity analysis in the setting of the initial estimate of $\boldsymbol{s}$, we also consider a random estimate of $\boldsymbol{s}$ by randomly choosing six elements of $\boldsymbol{s}$ to be one and setting the remaining elements to be zero, called 'proposed method (random)'. For the purpose of comparison, we employ the following four methods that are commonly used in high-dimensional variable selection: (1) lasso, (2) SCAD (smoothly clipped absolute deviation, Fan and Li (2001)), (3) MCP (minimax concave penalty, Zhang (2010)), and (4) elastic net. The simulation study is conducted by using R, where lasso and elastic net are implemented by the `glmnet` package, and MCP and SCAD are implemented by the `ncvreg` package. For lasso, SCAD, MCP, and elastic net, we use the convex surrogate approach given in Section 2.3 with a grid of tuning parameters that are generated by the `R` packages.

We generate survival time T_i^* and censoring time C_i for $i = 1, \ldots, n$ independently as follows:

- $T_i^* = -\frac{\log(U_i)}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}$, where $U_i \overset{iid}{\sim} \text{Unif}(0,1)$, $\mathbf{x}_i \overset{iid}{\sim} \text{N}_p(0, \Sigma)$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$, and $\boldsymbol{\beta}$ is the p-dimensional vector with $\beta_1 = \beta_9 = 0.8$, $\beta_4 = \beta_{12} = -0.7$, $\beta_5 = \beta_{13} = 0.6$, and $\beta_j = 0$ for $j \neq 1, 4, 5, 9, 12, 13$.

- $C_i \overset{iid}{\sim} \text{Exp}(\eta)$, where $\eta = 0.22$ (about 25% censoring rate) or $\eta = 0.57$ (about 40% censoring rate).

To consider various high-dimensional data settings, we consider the following 12 scenarios in the data-generating process:

(1). $n = 200$, $p = 100$, and censor rate= 25% ($\eta = 0.22$).

(2). $n = 200$, $p = 1000$, and censor rate= 25% ($\eta = 0.22$).

(3). $n = 200$, $p = 2000$, and censor rate= 25% ($\eta = 0.22$).

(4). $n = 500$, $p = 100$, and censor rate= 25% ($\eta = 0.22$).

(5). $n = 500$, $p = 1000$, and censor rate= 25% ($\eta = 0.22$).

(6). $n = 500$, $p = 2000$, and censor rate= 25% ($\eta = 0.22$).

(7). $n = 200$, $p = 100$, and censor rate= 40% ($\eta = 0.57$).

(8). $n = 200$, $p = 1000$, and censor rate= 40% ($\eta = 0.57$).

(9). $n = 200$, $p = 2000$, and censor rate= 40% ($\eta = 0.57$).

(10). $n = 500$, $p = 100$, and censor rate= 40% ($\eta = 0.57$).

(11). $n = 500$, $p = 1000$, and censor rate= 40% ($\eta = 0.57$).

(12). $n = 500$, $p = 2000$, and censor rate= 40% ($\eta = 0.57$).

To evaluate the performance of finding the global optimum model, for each method, we count the number of cases in which the EBIC evaluated at the optimal model is smaller than the other methods over 100 Monte Carlo replications. We denote by $\mathrm{F}_{\min}$ the ratio of finding the smallest EBIC out of the 100 replications. In addition, to access the variable selection performance, we calculate the false-positive rate (FPR) and the false-negative rate (FNR),

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \quad \text{and} \quad \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}},$$

where TP, FP, TN and FN denote the number of true non-zeros, false non-zeros, true zeros, and false zeros, respectively.

The simulation result is summarized in Tables 2.1 and 2.2, where Time represents the average execution time in minutes over 30 replications when a Windows 10 computer with

an Inter Core i7-8650U processor and 16 GB of memory is used. The result clearly shows that our proposed method always has the highest frequency of finding the smallest EBIC compared with other methods for all 12 scenarios. In addition, the proposed method is less sensitive to the choice of the initial estimate of $\boldsymbol{s}$. This implies that our proposed method successfully identifies the global optimal model for the EBIC model selection procedure. It is also worth noting that the proposed method always achieves the lowest level of FPR and FNR for all 12 scenarios. This means that the proposed method provides the best performance in identifying the true model for high-dimensional variable selection. As mentioned earlier, EBIC possesses model selection consistency, that is, the global optimum model tends to be the true model with high probability when the sample size is large. Hence, the superiority of our proposed method in model selection can be regarded as another evidence to demonstrate that the proposed method successfully finds the global optimal model in terms of EBIC.

2.6 Real data analysis

In this section, we conduct real data analysis with the Diffuse Large B-Cell Lymphoma (DLBCL) data (Alizadeh et al., 2000; Rosenwald et al., 2002). DLBCL has been known to be the most common type of non-Hodgkin lymphoma in the United States and worldwide (https://lymphoma.org/aboutlymphoma/nhl/dlbcl/). The dataset used in this analysis is publicly available at the R package `ROC632`. The data contain information about 240 DLBCL patients who were monitored using a Lyphochip cDNA microarray with 7399 gene expressions. Since 5 observations have survival time equal to 0, we eliminate them and use the information of the remaining 235 patients for our analysis. In the dataset, the censoring rate is 0.434.

First, we perform a pre-screening procedure to screen out redundant covariates that are obviously unrelated to the survival time. For every single covariate, we obtain p-value by fitting the Cox model with the single covariate and then exclude it from the analysis if the obtained p-value is greater than 0.05. After the screening procedure, 1163 genes are finally selected for our analysis. Then, we apply the proposed method and the existing methods

(lasso, SCAD, MCP and elastic net) as in Section 2.5. Table 2.3 displays our analysis result. The result shows that our proposed method provides the smallest EBIC (=1307.769). This implies that the model selected by our proposed method receives the strongest support from the observed data.

2.7 Concluding remarks

We have proposed a global optimal model selection procedure with GIC using the notion of statistical mechanics. The superiority of the proposed method in high-dimensional variable selection has been shown by the simulation study and real data analysis.

While we have restricted our attention to the Cox model in this chapter, the proposed method can be easily adapted to different parametric and semi-parametric survival models by replacing the partial likelihood with the likelihood or pseudo-likelihood functions. In addition, various choices of GIC can be considered in the proposed framework. For recent developments in model selection criterion that belongs to GIC, see Kim et al. (2016) and references therein.

Table 2.1: *Simulation result with censoring rate=25%*

(n, p)	Method	F_{min}	FPR	FNR	Time (mins)
(200,100)	Proposed method (null)	0.97	0.0022 (0.0006)	0.0050 (0.0037)	0.6997 (0.0353)
	Proposed method (random)	0.98	0.0023 (0.0006)	0.0050 (0.0037)	0.7301 (0.0341)
	LASSO	0.17	0.0182 (0.0016)	0.0867 (0.0163)	0.0123 (0.0001)
	SCAD	0.25	0.0147 (0.0015)	0.0650 (0.0148)	0.0066 (0.0001)
	MCP	0.61	0.0067 (0.0009)	0.0333 (0.0113)	0.0071 (0.0001)
	Elastic Net	0.08	0.0197 (0.0018)	0.1517 (0.0212)	0.0130 (0.0001)
(200,1000)	Proposed method (null)	0.84	0.0025 (0.0007)	0.1167 (0.0235)	7.4433 (0.4144)
	Proposed method (random)	0.87	0.0030 (0.0007)	0.0900 (0.0193)	8.2580 (0.4954)
	LASSO	0.05	0.0039 (0.0008)	0.5300 (0.0193)	0.0488 (0.0011)
	SCAD	0.05	0.0048 (0.0010)	0.5183 (0.0204)	0.0662 (0.0016)
	MCP	0.13	0.0111 (0.0016)	0.3167 (0.0293)	0.0434 (0.0010)
	Elastic Net	0.05	0.0033 (0.0008)	0.5767 (0.0160)	0.0587 (0.0016)
(200,2000)	Proposed method (null)	0.85	0.0011 (0.0003)	0.2133 (0.0288)	14.7546 (0.9356)
	Proposed method (random)	0.86	0.0014 (0.0004)	0.1650 (0.0261)	15.1516 (0.7879)
	LASSO	0.13	0.0016 (0.0005)	0.6217 (0.0148)	0.0226 (0.0001)
	SCAD	0.14	0.0021 (0.0006)	0.6150 (0.0162)	0.0296 (0.0002)
	MCP	0.15	0.0108 (0.0017)	0.4400 (0.0279)	0.0172 (0.0002)
	Elastic Net	0.12	0.0013 (0.0004)	0.6350 (0.0120)	0.0266 (0.0001)
(500,100)	Proposed method (null)	1.00	0.0013 (0.0005)	0.0000 (0.0000)	1.3108 (0.0821)
	Proposed method (random)	1.00	0.0013 (0.0005)	0.0000 (0.0000)	1.3460 (0.0904)
	LASSO	0.87	0.0015 (0.0005)	0.0000 (0.0000)	0.0448 (0.0004)
	SCAD	0.96	0.0011 (0.0004)	0.0000 (0.0000)	0.0299 (0.0004)
	MCP	0.99	0.0010 (0.0003)	0.0000 (0.0000)	0.0305 (0.0004)
	Elastic Net	0.63	0.0049 (0.0008)	0.0000 (0.0000)	0.0456 (0.0004)
(500,1000)	Proposed method (null)	1.00	0.0011 (0.0003)	0.0000 (0.0000)	12.1821 (0.5317)
	Proposed method (random)	1.00	0.0011 (0.0003)	0.0000 (0.0000)	12.4426 (0.5394)
	LASSO	0.68	0.0040 (0.0008)	0.0017 (0.0017)	0.1033 (0.0026)
	SCAD	0.87	0.0018 (0.0005)	0.0000 (0.0000)	0.1783 (0.0040)
	MCP	0.99	0.0012 (0.0004)	0.0000 (0.0000)	0.1291 (0.0033)
	Elastic Net	0.51	0.0094 (0.0012)	0.0033 (0.0023)	0.1222 (0.0034)
(500,2000)	Proposed method (null)	1.00	0.0014 (0.0004)	0.0000 (0.0000)	28.1613 (1.6829)
	Proposed method (random)	1.00	0.0014 (0.0004)	0.0000 (0.0000)	29.5035 (1.2440)
	LASSO	0.57	0.0064 (0.0011)	0.0100 (0.0052)	0.1151 (0.0004)
	SCAD	0.70	0.0050 (0.0010)	0.0017 (0.0017)	0.1149 (0.0007)
	MCP	0.96	0.0019 (0.0004)	0.0000 (0.0000)	0.0726 (0.0002)
	Elastic Net	0.32	0.0129 (0.0015)	0.0017 (0.0065)	0.1320 (0.0004)

Table 2.2: *Simulation result with censoring rate=40%*

(n, p)	Method	F_{min}	FPR	FNR	Time (mins)
(200,100)	Proposed method (null)	0.95	0.0018 (0.0005)	0.0383 (0.0120)	0.8466 (0.0416)
	Proposed method (random)	0.98	0.0022 (0.0006)	0.0350 (0.0104)	0.8786 (0.0455)
	LASSO	0.08	0.0139 (0.0015)	0.2183 (0.0232)	0.0210 (0.0005)
	SCAD	0.16	0.0171 (0.0018)	0.1367 (0.0212)	0.0121 (0.0003)
	MCP	0.40	0.0104 (0.0013)	0.0683 (0.0148)	0.0128 (0.0003)
	Elastic Net	0.06	0.0143 (0.0017)	0.3000 (0.0248)	0.0224 (0.0005)
(200,1000)	Proposed method (null)	0.79	0.0035 (0.0008)	0.2417 (0.0292)	7.2024 (0.4769)
	Proposed method (random)	0.81	0.0040 (0.0010)	0.2333 (0.0284)	8.1090 (0.5236)
	LASSO	0.18	0.0020 (0.0005)	0.6317 (0.0145)	0.0461 (0.0013)
	SCAD	0.18	0.0027 (0.0007)	0.6200 (0.0158)	0.0538 (0.0008)
	MCP	0.19	0.0078 (0.0013)	0.4633 (0.0275)	0.0354 (0.0006)
	Elastic Net	0.18	0.0015 (0.0004)	0.6483 (0.0123)	0.0519 (0.0007)
(200,2000)	Proposed method (null)	0.76	0.0025 (0.0006)	0.3800 (0.0304)	13.0769 (1.0701)
	Proposed method (random)	0.89	0.0029 (0.0007)	0.2717 (0.0289)	14.4596 (0.9596)
	LASSO	0.23	0.0017 (0.0004)	0.6633 (0.0103)	0.0232 (0.0001)
	SCAD	0.24	0.0018 (0.0004)	0.6600 (0.0108)	0.0275 (0.0002)
	MCP	0.24	0.0049 (0.0011)	0.5917 (0.0196)	0.0161 (0.0002)
	Elastic Net	0.23	0.0011 (0.0003)	0.6683 (0.0102)	0.0275 (0.0001)
(500,100)	Proposed method (null)	1.00	0.0013 (0.0004)	0.0000 (0.0000)	1.0255 (0.0462)
	Proposed method (random)	1.00	0.0013 (0.0004)	0.0000 (0.0000)	1.1192 (0.0519)
	LASSO	0.78	0.0033 (0.0006)	0.0000 (0.0000)	0.0448 (0.0007)
	SCAD	0.96	0.0015 (0.0004)	0.0000 (0.0000)	0.0237 (0.0004)
	MCP	1.00	0.0013 (0.0004)	0.0000 (0.0000)	0.0247 (0.0004)
	Elastic Net	0.58	0.0069 (0.0011)	0.0000 (0.0000)	0.0457 (0.0007)
(500,1000)	Proposed method (null)	1.00	0.0017 (0.0004)	0.0000 (0.0000)	12.6398 (0.5068)
	Proposed method (random)	1.00	0.0017 (0.0004)	0.0000 (0.0000)	13.6700 (0.6852)
	LASSO	0.46	0.0079 (0.0010)	0.0015 (0.0075)	0.0151 (0.0023)
	SCAD	0.62	0.0061 (0.0010)	0.0017 (0.0017)	0.1572 (0.0033)
	MCP	0.97	0.0020 (0.0005)	0.0000 (0.0000)	0.1128 (0.0027)
	Elastic Net	0.36	0.0121 (0.0013)	0.0267 (0.0088)	0.1264 (0.0032)
(500,2000)	Proposed method (null)	0.99	0.0013 (0.0006)	0.0000 (0.0000)	24.3697 (0.8678)
	Proposed method (random)	0.98	0.0015 (0.0007)	0.0000 (0.0000)	26.6497 (1.4551)
	LASSO	0.30	0.0135 (0.0014)	0.0250 (0.0080)	0.1195 (0.0005)
	SCAD	0.43	0.0115 (0.0015)	0.0167 (0.0060)	0.1054 (0.0006)
	MCP	0.82	0.0040 (0.0008)	0.0017 (0.0017)	0.0663 (0.0005)
	Elastic Net	0.13	0.0163 (0.0017)	0.0583 (0.0120)	0.1388 (0.0007)

Table 2.3: *Real data analysis result with DLBCL data.*

	Index set of selected genes	EBIC
Our method	{260,663,1127,1162}	1307.769
Lasso	{ 705 }	1322.035
SCAD	{ 705 }	1322.035
MCP	{ 260,705,867 }	1318.424
Elastic net	{ 705 }	1322.035

Chapter 3

Best predictive model selection for high-dimensional survival data

3.1 Introduction

In survival analysis, prediction is critical in many fields, including life health science research challenges and disease investigations. Predicting life expectancy can be a complicated but essential topic. With the popularity of high-dimensional data, how to find the optimal model that delivers the best prediction performance has attracted increasing attention, as only a small proportion of the covariates are truly related to the response.

For the purpose of survival model analysis, numerous survival regression models with a wide variety of distributions have been thoroughly investigated. One of the most commonly used parametric models is called the accelerated failure time (AFT) model (Wei, 1992). It provides a widely utilized approach for estimating the effects of the covariates on the response, where the effect of a covariate is to accelerate or decelerate the survival time by some constant. For the goal of making predictions, a variety of measurements can be employed, such as the mean time to failure (MTTF), median survival time, and minimum prediction error survival time (MPET). They provide different ways of generating predictions from various perspectives, each with its own set of characteristics. With the AFT model,

once the effects of the covariates have been estimated, the values can be plugged into the prediction measurement for generating predictions.

When implementing model selection with high-dimensional survival data, a frequent approach is to define an appropriate model selection criterion that quantifies the best model. Then the best model can be found with various methodologies. A variety of model selection criteria can be employed, for example, the Bayesian information criterion (BIC) (Schwarz et al., 1978), Akaike information criterion(AIC) (Akaike, 1974), and extended BIC (Chen and Chen, 2008). All of these criteria can be referred to as the generalized information criterion (GIC) (Atkinson, 1980; Kim et al., 2012; Zhang et al., 2010). The model that optimizes the model selection criterion is the best fitting model. However, finding the best fitting model is not the same as finding the best predictive model, because the best fitting model cannot guarantee the best prediction performance. The penalized likelihood estimation method (e.g. lasso Tibshirani (1996), adaptive lasso Zou (2006), and SCAD Fan and Li (2001)) also serves as a way of implementing model selection for the high-dimensional survival data. These approaches work by cross-validation to find the best tuning parameters in making predictions, which can be considered as a convex surrogate. However, due to the limitation of the coverage for the solution path of the tuning parameters, these methods still can not guarantee the selection of the best predictive model.

In the context of the Bayesian framework, the idea of the median probability model is proposed by Barbieri et al. (2004) for optimal predictive Bayesian model selection. Motivated by the idea of the median probability model, we propose a new way for finding the best predictive model with high-dimensional survival data. The key idea is to incorporate the median probability model into the frequentist framework via the concept of Boltzmann distribution. The resulted algorithm brings the idea from the Bayesian framework to the frequentist framework. The proposed method enables us to generate a sequence of candidate models using the Gibbs sampler, from which the best predictive model can be defined. In the parametric survival regression model, with the estimated values of the covariates, prediction can be calculated respectively. For model specification, the accelerate failure time (AFT) model with Weibull distribution is employed for the explanation, which is discussed

in Chapter 3.2. In Chapter 3.3, the idea of the median probability model is introduced. The details of our proposed method are explained in Chapter 3.4. To evaluate the performance of the proposed method, simulation study and real data analysis are demonstrated in Chapter 3.5 and Chapter 3.6.

3.2 Model setup

3.2.1 Accelerated failure time (AFT) model

For time-to-event data with n independent observations, assume that the ith individual's actual death time is T_i^*, the corresponding observed failure time can be expressed as $T_i = \min(T_i^*, C_i)$, where C_i is the censoring time for the ith individual, with the censoring indicator denoted by $\delta_i = I\{T_i^* \leq C_i\}$, and $I\{\cdot\}$ is the indicator function.

In survival analysis, revealing the relationship between the response and various covariates is of considerable interest to many disease studies. Making predictions is also the primary aim for a large number of research questions in this field. Among various parametric models, the accelerated failure time (AFT) model (Wei, 1992) is one of the most commonly utilized models. This model makes the assumption that the effects of the variables will cause the lifetime to accelerate or decelerate by a constant amount. It demonstrates the connection between the covariates $\mathbf{x}$ and the log of survival time $Y = \log T$ as follows:

$$Y_i = \log T_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \sigma \epsilon_i, \tag{3.1}$$

where $i = 1, \cdots, n$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^{\mathrm{T}}$ is the $p-$dimensional coefficient vector, which is of our primal interest. $\sigma\epsilon_i$ is the error term, in which σ is called the scale parameter and ϵ_i is the random disturbance term, usually assumed to be independent identically distributed with some density function $f(\epsilon)$. The model (3.1) defines a broad class of models, and depending on the distribution we specify for ϵ_i, we will obtain different models with different properties.

For the purpose of explanation, the AFT model in (3.1) can be alternatively stated in a

different way:

$$\begin{aligned} T_i &= \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\exp(\sigma\epsilon_i) \\ &= e^{\eta_i}T_0, \end{aligned}$$

where $T_0 = \exp(\sigma\epsilon_i)$ and $\eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ for $i = 1, \cdots, n$. This type of expression clearly shows that the effects of the covariates act multiplicatively on time (e.g., when $e^{\eta_i} = 0.5$, the subject effectively ages at twice the normal speed).

In the context of right-censored scenario with independent observations $\{(T_i, \mathbf{x}_i, \delta_i), i = 1, ..., n\}$, the regression parameter $\boldsymbol{\beta}$ in model (3.1) can be estimated by maximum likelihood estimation, in which the likelihood function is defined as

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i:\delta_i=1}[f(T_i, \mathbf{x}_i|\boldsymbol{\beta})] \times \prod_{i:\delta_i=0}[1 - F(T_i, \mathbf{x}_i|\boldsymbol{\beta})] \\ &= \prod_{i=1}^{n}[f(T_i, \mathbf{x}_i|\boldsymbol{\beta})]^{\delta_i} \times [S(T_i, \mathbf{x}_i|\boldsymbol{\beta})]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[\frac{f(T_i, \mathbf{x}_i|\boldsymbol{\beta})}{S(T_i, \mathbf{x}_i|\boldsymbol{\beta})}\right]^{\delta_i} \times S(T_i, \mathbf{x}_i|\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n}[h(T_i, \mathbf{x}_i|\boldsymbol{\beta})]^{\delta_i} \times [S(T_i, \mathbf{x}_i|\boldsymbol{\beta})], \end{aligned} \tag{3.2}$$

where $f(\cdot \mid \boldsymbol{\beta})$ is the probability density function, $F(\cdot \mid \boldsymbol{\beta})$ is the cumulative distribution function, $S(\cdot \mid \boldsymbol{\beta})$ gives the survival function, and $h(\cdot \mid \boldsymbol{\beta})$ denotes the hazard function. The hazard function, also known as the instantaneous failure rate, is always of importance in survival data analysis because it shows the risk of an event occurring at any given point in time. A flexible hazard function is one of the characteristics of some of the regularly used distributions. One of the distributions, called Weibull distribution, will be examined in greater detail in the next section.

3.2.2 Weibull distribution under AFT model

In survival analysis, different types of data distributions provide various properties on the hazard function. Some of the widely used distributions, such as the exponential distribution, or the Gompertz distribution, lead to a constant or monotone increasing hazard function. In reality, however, populations with an unchanging or constantly growing hazard function are rare. Then the Weibull distribution (Fréchet, 1927; Rosin, 1933; Weibull et al., 1951) is proposed, which provides more flexibilities for the hazard function.

Weibull distribution, also referred to as the type III extreme value distribution, is a form of distribution that is frequently employed for survival data analysis. This distribution includes the exponential distribution as the special case, and can be uniquely determined by three parameters, the location parameter $\alpha \in \mathbb{R}$, the scale parameter $\rho > 0$, and the shape parameter $\gamma \in \mathbb{R}$. Depending on the value we specify for γ, the Weibull distribution can provide a wide range of flexibilities, including monotone growing, decreasing, and constant hazard functions. As a rule of thumb, the location parameter $\alpha \in \mathbb{R}$ is usually assumed to be zero, which simplifies the Weibull distribution to a two-parameter distribution.

If a random variable T follows the Weibull distribution, $T \sim W(\rho, \gamma)$, the probability density function of T is defined as

$$f(t) = \frac{\gamma}{\rho}\left(\frac{t}{\rho}\right)^{\gamma-1} \exp\left[-\left(\frac{t}{\rho}\right)^{\gamma}\right],$$

where $\rho > 0$, and $\gamma > 0$. The corresponding cumulative distribution function, survival function, and hazard function can be derived as

$$\begin{aligned} F(t) &= 1 - \exp\left[-\left(\frac{t}{\rho}\right)^{\gamma}\right], \\ S(t) &= 1 - F(t) = \exp\left[-\left(\frac{t}{\rho}\right)^{\gamma}\right], \\ h(t) &= \frac{f(t)}{S(t)} = \frac{\gamma}{\rho}\left(\frac{t}{\rho}\right)^{\gamma-1}. \end{aligned}$$

From the hazard function $h(t)$, it can be shown that $\gamma > 1$ provides a monotone increasing

hazard function, $\gamma = 1$ leads the hazard function to be a constant, and $\gamma < 1$ makes the hazard function consistently decreasing.

Another well-known distribution for survival analysis is called the Gumbel distribution (Gumbel, 1935), also known as the type I extreme value distribution, which is expressed as the log of the Weibull distribution. If we assume $w = \log T$, where $T \sim W(\rho, \gamma)$, then we get $w \sim G(\mu, \sigma)$. The log transformation of T transforms the support from $\{T \geq 0\}$ to $\{-\infty < w < \infty\}$, and the probability density function of w can be derived as

$$f(w) = \frac{1}{\sigma} \exp\left[\left(-\frac{w-\mu}{\sigma}\right) - \exp\left(-\frac{w-\mu}{\sigma}\right)\right], \tag{3.3}$$

where $\mu = \log \rho \in (-\infty, \infty)$, $\sigma = \frac{1}{\gamma} \in (0, \infty)$. The corresponding cumulative distribution function, survival function and hazard function can be derived as follows:

$$\begin{aligned} F(w) &= 1 - \exp\left[-\exp\left(\frac{w-\mu}{\sigma}\right)\right], \\ S(w) &= \exp\left[-\exp\left(\frac{w-\mu}{\sigma}\right)\right], \\ h(w) &= \frac{f(w)}{S(w)} = \frac{1}{\sigma}\exp\left(\frac{w-\mu}{\sigma}\right). \end{aligned}$$

For the AFT model we specify in (3.1), which can be alternatively presented as

$$Y_i = \log(T_i) = \beta_0 + \beta_1 x_{i1} + ...\beta_p x_{ip} + \sigma\epsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sigma\epsilon_i,$$

where $i = 1, 2, .., n$. If we assume ϵ_i follows the standard Gumbel distribution, $\epsilon_i \overset{iid}{\sim} G(0, 1)$, with the probability density function defined as

$$f(\epsilon) = \exp[-(\epsilon + \exp(-\epsilon))],$$

it leads the corresponding probability density function of T_i to be

$$f(t) = \frac{1/\sigma}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\left[\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right]^{\frac{1}{\sigma}-1}\exp\left[-\left(\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right)^{\frac{1}{\sigma}}\right].$$

Referring to the probability density function in (3.3), the above probability density function indicates that the survival time T_i follows a Weibull distribution with parameters $\rho = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\gamma = \frac{1}{\sigma}$, which can be denoted by $T_i \overset{iid}{\sim} W(\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \frac{1}{\sigma})$. And the corresponding cumulative distribution function, survival function, and hazard function for T_i can be derived as

$$\begin{aligned}
F(t) &= 1 - \exp\left[-\left(\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right)^{\frac{1}{\sigma}}\right], \\
S(t) &= 1 - F(t) = \exp\left[-\left(\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right)^{\frac{1}{\sigma}}\right], \\
h(t) &= \frac{f(t)}{S(t)} = \frac{1/\sigma}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\left[\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right]^{\frac{1}{\sigma}-1}.
\end{aligned}$$

Once the distribution of the data is specified, the values of the parameters $(\boldsymbol{\beta}, \sigma)$ under the Weibull AFT model can be estimated. A typical common strategy is the maximum likelihood estimation.

With n independent observations $t_1, t_2, ...t_n$ under the AFT model, we assume $\epsilon_i \overset{iid}{\sim} G(0,1)$, the likelihood function in (3.2) can be derived as

$$\begin{aligned}
L(\boldsymbol{\beta}, \sigma|\boldsymbol{T}, \boldsymbol{\delta}) &= \prod_{i=1}^{n}\{f(t_i)\}^{\delta_i}\{S(t_i)\}^{1-\delta_i} \\
&= \prod_{i=1}^{n}\{h(t_i)\}^{\delta_i}S(t_i) \\
&= \prod_{i=1}^{n}\left\{\frac{1/\sigma}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\left[\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right]^{\frac{1}{\sigma}-1}\right\}^{\delta_i}\exp\left[-\left(\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right)^{\frac{1}{\sigma}}\right].
\end{aligned}$$

To maximize the likelihood function, a straightforward approach is to take the log of the likelihood function and maximize the corresponding log-likelihood function, which can be

obtained as follows:

$$
\begin{aligned}
l(\boldsymbol{\beta}, \sigma|\boldsymbol{T}, \boldsymbol{\delta}) &= \log L(\boldsymbol{\beta}, \sigma|\boldsymbol{T}, \boldsymbol{\delta}) \\
&= \log \prod_{i=1}^{n} \left\{ \frac{1/\sigma}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \left[\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \right]^{\frac{1}{\sigma}-1} \right\}^{\delta_i} \exp\left[-\left(\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \right)^{\frac{1}{\sigma}} \right] \\
&= \sum_{i=1}^{n} \delta_i \log \left\{ \frac{1/\sigma}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \left[\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \right]^{\frac{1}{\sigma}-1} \right\} - \sum_{i=1}^{n} \left(\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \right)^{\frac{1}{\sigma}} \\
&= \log \frac{1}{\sigma} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i (\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \left(\frac{1}{\sigma} - 1 \right) \sum_{i=1}^{n} \delta_i (\log t_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \\
&\quad - \sum_{i=1}^{n} \left(\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \right)^{\frac{1}{\sigma}} \\
&= \log \frac{1}{\sigma} \sum_{i=1}^{n} \delta_i + \left(\frac{1}{\sigma} - 1 \right) \sum_{i=1}^{n} \delta_i (\log t_i) - \frac{1}{\sigma} \sum_{i=1}^{n} \delta_i (\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sum_{i=1}^{n} \left(\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \right)^{\frac{1}{\sigma}}.
\end{aligned}
$$

Different numerical approaches (e.g., the Newton-Raphson method Newton (1736)) can be applied to the optimization procedure with the detailed form of the log-likelihood function. Then the maximum likelihood estimation of the parameters $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ can be obtained.

3.2.3 Prediction under the Weibull AFT model

When it comes to generating predictions in the context of AFT models, statisticians have several options. The mean time to failure (MTTF) (Ho and Silva, 2006), median survival time, and minimum prediction error survival time (MPET) (Henderson et al., 2001) are widely used ways of making predictions. In this chapter, we choose to use the mean time to failure (MTTF) as the objective of the prediction, and the logic can be extended to any other measurements.

Under the assumption of the AFT regression model with Weibull distribution, predicting

the MTTF can be derived as follows (Liu, 2018):

$$
\begin{aligned}
E(Y) &= E(\log(T)) \\
&= \int_{-\infty}^{\infty} y f(y) dy \\
&= \int_{-\infty}^{\infty} y \frac{1}{\sigma} \exp\left(\frac{y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right) \exp\left[-\exp\left(\frac{y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right)\right] dy \\
&= \frac{1}{\sigma} \int_{0}^{\infty} (\sigma \log z + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) z \exp(-z) d(\sigma \log z + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \\
&= \int_{0}^{\infty} (\sigma \log z + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \exp(-z) dz \\
&= \int_{0}^{\infty} \sigma \frac{\partial}{\partial \alpha} [z^{\alpha} \exp(-z)]_{\alpha=0} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \\
&= \sigma \frac{\partial}{\partial \alpha} \left[\int_{0}^{\infty} z^{\alpha} \exp(-z)\right]_{\alpha=0} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \\
&= \sigma \Gamma(\alpha)'_{\alpha=1} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta},
\end{aligned}
\tag{3.4}
$$

where $z = \exp(\frac{y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma})$, and $\Gamma(\alpha)'_{\alpha=1}$ is the negative of the Euler-Mascheroni Constant (≈ -0.57721 Euler (1740)). It provides the log of the anticipated survival time for each individual.

To obtain the mean time to failure (MTTF) of the ith individual, the first step is to get the estimation of the parameters $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$. Then, by (3.4), the prediction for the ith individual can be calculated as

$$
E(T_i) = \exp\left[\hat{\sigma}\Gamma(\alpha)'_{\alpha=1} + \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right].
$$

In addition to making point prediction, we can also construct the confidence interval for the mean time between failures (MTTF) with the Delta method. With the estimated values of the parameters $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$, the standard error of MTTF can be calculated as

$$
SE = \left\{ \begin{pmatrix} \frac{\partial E(T_i)}{\partial \hat{\boldsymbol{\beta}}} \\ \frac{\partial E(T_i)}{\partial \hat{\sigma}} \end{pmatrix}^{\mathrm{T}} \Sigma_{(\hat{\boldsymbol{\beta}}, \hat{\sigma})} \begin{pmatrix} \frac{\partial E(T_i)}{\partial \hat{\boldsymbol{\beta}}} \\ \frac{\partial E(T_i)}{\partial \hat{\sigma}} \end{pmatrix} \right\}^{\frac{1}{2}}.
$$

where $\Sigma_{(\hat{\boldsymbol{\beta}}, \hat{\sigma})}$ is the variance-covariance matrix of $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$. And the $(1-\alpha)\%$ confidence interval

is defined as

$$E(T_i) - z_{1-\frac{\alpha}{2}} SE < T_i < E(T_i) + z_{1-\frac{\alpha}{2}} SE,$$

where α controls the type I error, and z represents the quantile of the standard normal distribution.

3.3 Best predictive model selection

For survival data analysis, the problem of best predictive model selection has been extensively studied based on parametric models with a relatively small number of covariates. However, for high-dimensional survival data, identifying the best predictive model is still theoretically and computationally challenging. In Chapter 2, we propose a method for finding the global optimal model that optimizes the model selection criterion. The selected model is known to be the best fitting model. However, finding the best fitting model is not the same as finding the best predictive model, as the best fitting model cannot guarantee the best predictive performance. A recent literature proposes the idea of optimal Bayesian predictive model selection, which gives the definition of the median probability model. Motivated by the idea of the median probability model, we propose a new algorithm for finding the best predictive model under the frequentist framework.

3.3.1 Median probability model

The idea of the median probability model, originally proposed by Barbieri et al. (2004), is a popular and efficient method for finding the optimal predictive model among normal linear models from the Bayesian perspective. Under the Bayesian framework, it is generally assumed that the model with the highest posterior likelihood is the best predictive model. However, the conclusion can be held only under very strict conditions: only if two models are being entertained (Berger, 1997) or in the variable selection problem for linear models having orthogonal design matrices (Clyde, 1999; Clyde and George, 2000, 1999). However, these conditions can be easily violated. Then the idea of the median probability model is

proposed, which is defined as the model consisting of those variables with overall posterior probability greater than or equal to 1/2. This idea provides a way of finding the model that delivers the best predictive performance under a generalized condition.

The posterior model probabilities $P(\boldsymbol{s}|T)$ is determined by some Markov chain Monte Carlo (MCMC) schemes under the Bayesian framework, and the posterior inclusion probability for each variable j is defined as

$$p_j = \sum_{\boldsymbol{s}\in\mathcal{S}:s_j=1} P(\boldsymbol{s}|T), \tag{3.5}$$

where $\mathcal{S}$ is a set of candidate models, $\boldsymbol{s}$ is any sub-model with model index $\boldsymbol{s} = (s_1, s_2, ..., s_p)$, s_i being either 1 or 0 as covariate $\mathbf{x}_i$ is in or out of the model. This gives the overall posterior probability that variable j is in the model.

If it exists, the median probability model $\boldsymbol{s}^*$ is defined to be the model consisting of those variables whose posterior inclusion probability is greater than or equal to 1/2. The corresponding $\boldsymbol{s}^*$ can be formally defined by

$$s_j^* = \begin{cases} 1, & \text{if } p_j \geq \dfrac{1}{2}, \\ 0, & \text{othewise.} \end{cases} \tag{3.6}$$

This model selection procedure can be summarized as follows:

- Step 1: Set an initial candidate model $\boldsymbol{s} = \boldsymbol{s}_0$.
- Step 2: Draw a new candidate model $\boldsymbol{s}^*$ from a proposal distribution $q(\boldsymbol{s}^* \mid \boldsymbol{s})$, which represents the conditional probability of $\boldsymbol{s}^*$ given the current status of $\boldsymbol{s}$.
- Step 3: Repeat step 2 for certain times until we get a big enough set of candidate models $\mathcal{S}$, then calculate the posterior inclusion probability for each variable from (3.5).
- Step 4: The median probability model is given by (3.6), which provide the index of predictors included in the model.

3.3.2 Boltzmann distribution connects Bayesian and frequentist approach

The idea of the median probability model is proposed under the Bayesian framework, and how to specify the proposal distribution can be tricky without any prior information. By applying the idea of statistical mechanics, we propose to treat the Boltzmann distribution as the posterior distribution. Then we bring the idea from the Bayesian framework to the frequentist framework.

The idea of the Boltzmann distribution (Boltzmann, 1868; Gibbs, 1902) comes from system energy optimization. It gives the probability that a system being in a certain state as a function of that state's energy and the temperature of the system. The distribution can be expressed as

$$p(\boldsymbol{s}) \propto \exp\left\{-\frac{E(\boldsymbol{s})}{\kappa\tau}\right\},$$

where $p(\boldsymbol{s})$ is the probability of the system being in state $\boldsymbol{s}$, $\boldsymbol{s} = (s_1, s_2, ..., s_p)$ indicates a candidate model, it gives the indices of the covariates included in the model. $s_j \in \{0, 1\}$ for $j = 1, 2, ..., p$, with $s_j = 1$ representing the jth covariate included in the model, otherwise excluded. $E(\boldsymbol{s})$ is the energy of that state, κ is the Boltzmann's constant and τ is the temperature.

In order to incorporate the idea of Boltzmann distribution with model selection, we can replace the energy function $E(\boldsymbol{s})$ by a model selection criterion, which is also referred to as the generalized information criterion GIC($\boldsymbol{s}$). Optimizing GIC($\boldsymbol{s}$) is one of the principles that guide the choice of optimum model under the context high-dimensional survival data. By treating the Boltzmann distribution as the posterior distribution corresponding to the candidate model $\boldsymbol{s}$, the Markov chain Monte Carlo (MCMC) method can be employed for generating candidate models. Given the GIC($\boldsymbol{s}$) we specify, the corresponding Boltzmann distribution can be expressed as

$$b_\tau(\boldsymbol{s}) \propto \exp\left\{-\frac{\text{GIC}(\boldsymbol{s})}{\kappa\tau}\right\}.$$

A general form the the GIC($\boldsymbol{s}$) can be expressed as

$$\text{GIC}(\boldsymbol{s}) = -2l(\boldsymbol{\beta}(\boldsymbol{s})) + \text{pen}_\lambda(|\boldsymbol{s}|),$$

where $|\boldsymbol{s}| = \sum_{i=1}^{p} s_j$ gives the total number of covariates in the model, $l(\boldsymbol{\beta}(\boldsymbol{s}))$ is the log-likelihood of the model $\boldsymbol{s}$, and $\text{pen}_\lambda(|\boldsymbol{s}|)$ is the penalty function, which increases when the number of covariates selected in the model increases. λ is the tuning parameter, which controls the degrees of penalization.

Various forms of GIC($\boldsymbol{s}$) can be utilized as the model selection criterion. Among all possible choices, the extended Bayesian information criteria (EBIC) proposed by Chen and Chen (2008) provides the benefit of model selection consistency with high-dimensional survival data. It is thus utilized as the model selection criterion for the remainder of this chapter. The EBIC is defined as

$$\text{EBIC}_\gamma(\boldsymbol{s}) = -2l\left(\hat{\boldsymbol{\beta}}(\boldsymbol{s})\right) + \log(n_0)\ |\boldsymbol{s}| + 2\gamma \log \binom{p}{|\boldsymbol{s}|}, \tag{3.7}$$

where $\boldsymbol{\beta}(\boldsymbol{s})$ denotes the sub-vector of $\boldsymbol{\beta}$ with indices contained in $\boldsymbol{s}$, $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ is the maximum likelihood estimator of $\boldsymbol{\beta}(\boldsymbol{s})$, $|\boldsymbol{s}| = \sum_{i=1}^{p} s_j$ gives the number of covariates included in the model, $\binom{p}{|\boldsymbol{s}|}$ provides the number of all possible models with $|\boldsymbol{s}|$ indices, and γ is a tuning parameter between 0 and 1.

In addition to employing the EBIC($\boldsymbol{s}$) as the model selection guidance, another intriguing and beneficial component relies on the tuning parameter γ. Instead of treating γ as a fixed number, we treat it as a random variable and provide a way of generating and updating it based on its posterior distribution. As it is shown in the simulation study, under our proposed framework, treating γ as a random variable improves the prediction performance. The details are discussed in section 3.3.3.

3.3.3 Proposed algorithm

Combining the strength of the median probability model and the Boltzmann distribution, we propose a new best predictive model selection method under the frequentist framework. The proposed method employs the idea of $\text{EBIC}_\gamma(\boldsymbol{s})$ (3.7) as the guidance of model selection. As it has been shown in Chen and Chen (2008), assume that $p = O(n^K)$ for some constant K, if $\gamma > 1 - \frac{1}{2K}$, then, under the asymptotic identifiability condition, the probability that any model other than the true model will be selected tends to zero, which means the consistency property can be held. So we set the range for the tuning parameter as $\gamma > 1 - \frac{1}{2K}$, where $K = \frac{\log p}{\log n}$.

Also, motivated by the idea of Gibbs sampling, we propose a way of generating a candidate model based on the current model by simply adding or removing one new predictor. Let $\boldsymbol{s} = (s_1, ..., s_p)^{\mathrm{T}} \in \mathbb{R}^p$ be the current model. For a given j, we define a candidate model by $\boldsymbol{s}^* = (s_1^*, \ldots, s_p^*)$ such that $s_k^* = s_k$ if $k \neq j$, and $s_j^* = 1 - s_j$.

Utilizing the Boltzmann distribution, given a value of the tuning parameter γ , the posterior distribution of a model $\boldsymbol{s}$ can be derived as

$$
\begin{aligned}
p(\boldsymbol{s}|\gamma) &\propto \exp\left\{-\frac{\text{EBIC}(\boldsymbol{s})}{2\tau}\right\} \mathbb{I}(|\boldsymbol{s}| \leq k_n) \\
&\propto \exp\left\{-\frac{-2l(\beta(\boldsymbol{s})) + \log(n_0)||\beta(\boldsymbol{s})||_0 + 2\gamma \log\binom{p}{|\boldsymbol{s}|}}{2\tau}\right\} \mathbb{I}(|\boldsymbol{s}| \leq k_n) \\
&\propto \exp\left\{-\frac{\gamma \log\binom{p}{|\boldsymbol{s}|}}{\tau}\right\} \mathbb{I}(|\boldsymbol{s}| \leq k_n) \\
&\propto \binom{p}{|\boldsymbol{s}|}^{-\gamma} \mathbb{I}(|\boldsymbol{s}| \leq k_n) \\
&= m(\boldsymbol{s}),
\end{aligned}
$$

where k_n represents the maximum number of covariates selected in the model, which controls the model complexity.

Also, note that $\sum_{|\boldsymbol{k}|=1}^{k_n} m(\boldsymbol{k}) = \infty$, it gives the fact that the posterior distribution of $\boldsymbol{s}$ is improper. In order to make it a proper posterior distribution, we make the transformation

as follows:

$$\begin{aligned} p(\boldsymbol{s}|\gamma) &\propto \frac{m(\boldsymbol{s})}{\sum_{|\boldsymbol{k}|=1}^{k_n} m(\boldsymbol{k})} \\ &\propto \frac{m(\boldsymbol{s})}{\sum_{|\boldsymbol{k}|=1}^{k_n} \binom{p}{|\boldsymbol{k}|}\binom{p}{|\boldsymbol{k}|}^{-\gamma}} \\ &\propto \frac{\binom{p}{|\boldsymbol{s}|}^{-\gamma}}{\sum_{|\boldsymbol{k}|=1}^{k_n} \binom{p}{|\boldsymbol{k}|}^{1-\gamma}} \mathbb{I}(|\boldsymbol{s}| \leq k_n). \end{aligned}$$

For the posterior distribution of γ, in order to make it aligned with the principle of consistency, we set it as $\gamma > 1 - \frac{1}{2K}$ with a uniform distribution, where $K = \frac{\log p}{\log n}$. Then the posterior distribution of γ can be derived via the Bayes' theorem as follows:

$$\begin{aligned} p(\gamma|\boldsymbol{s}) \propto & p(\boldsymbol{s}|\gamma)p(\gamma) \\ \propto & \frac{1}{\sum_{|\boldsymbol{k}|=1}^{k_n} \binom{p}{|\boldsymbol{k}|}^{1-\gamma} / \binom{p}{|\boldsymbol{s}|}^{-\gamma}} \mathbb{I}(|\boldsymbol{s}| \leq k_n) \mathbb{I}_{(1-\frac{\log n}{2\log p}, \infty)}(\gamma). \end{aligned}$$

With all the information above, our proposed method can be summarized with the following procedure:

- Step 1: Start from an initial state of a candidate model $\boldsymbol{s}^{(0)} = (s_1^{(0)}, s_2^{(0)}, ..., s_p^{(0)})$ with a initial value of $\gamma = \gamma_0$, set an initial value for k_n, which controls the maximum number of covariates selected in the model.

- Step 2: Given γ, implement the Gibbs sampler to generate a new state $\boldsymbol{s}^{(i)}$ based on the current state $\boldsymbol{s}^{(i-1)}$ with the following rules:

  - Generate $\boldsymbol{s}^{(i)} = \left(s_1^{(i)}, s_2^{(i)}, \cdots, s_p^{(i)}\right)$ by setting $s_1^{(i)} = 1 - s_1^{(i-1)}$ and $s_l^{(i)} = s_l^{(i-1)}$ for $l \neq 1$.

    If $\sum_{j=1}^{p} s_j^{(i)} > k_n$, skip the following and move to generate $s_2^{(i)}$. Otherwise, calculate EBIC($\boldsymbol{s}^{(i)}$) and EBIC($\boldsymbol{s}^{(i-1)}$).

    Generate a Bernoulli trial with success probability w defined in (3.8). If we obtain 1, keep $s_1^{(i)} = s_1^{(i)}$, otherwise, update $s_1^{(i)} = s_1^{(i-1)}$.

– Generate $\boldsymbol{s}^{(i)} = \left(s_1^{(i)}, s_2^{(i)}, \cdots, s_p^{(i)}\right)$ by setting $s_2^{(i)} = 1 - s_2^{(i-1)}$ and $s_l^{(i)} = s_l^{(i-1)}$ for $l \neq 2$.

If $\sum_{j=1}^{p} s_j^{(i)} > k_n$, skip the following and move to generate $s_3^{(i)}$. Otherwise, calculate EBIC($\boldsymbol{s}^{(i)}$) and EBIC($\boldsymbol{s}^{(i-1)}$).

Generate a Bernoulli trial with success probability w defined in (3.8). If we obtain 1, keep $s_2^{(i)} = s_2^{(i)}$, otherwise, update $s_2^{(i)} = s_2^{(i-1)}$.

$\vdots$

– Generate $\boldsymbol{s}^{(i)} = \left(s_1^{(i)}, s_2^{(i)}, \cdots, s_p^{(i)}\right)$ by setting $s_p^{(i)} = 1 - s_p^{(i-1)}$ and $s_l^{(i)} = s_l^{(i-1)}$ for $l \neq p$.

If $\sum_{j=1}^{p} s_j^{(i)} > k_n$, skip the following and move to generate $s_1^{(i+1)}$. Otherwise, calculate EBIC($\boldsymbol{s}^{(i)}$) and EBIC($\boldsymbol{s}^{(i-1)}$).

Generate a Bernoulli trial with success probability w defined in (3.8). If we obtain 1, keep $s_p^{(i)} = s_p^{(i)}$, otherwise, update $s_p^{(i)} = s_p^{(i-1)}$.

The success probability in the Bernoulli trial is defined as

$$w = \frac{\exp\{-\frac{1}{\kappa\tau}\text{EBIC}(\boldsymbol{s}^{(i)})\}}{\exp\{-\frac{1}{\kappa\tau}\text{EBIC}(\boldsymbol{s}^{(i)})\} + \exp\{-\frac{1}{\kappa\tau}\text{EBIC}(\boldsymbol{s}^{(i-1)})\}}. \tag{3.8}$$

- Step 3: Given $\boldsymbol{s}^{(i)}$, generate a new sample of γ via the grid method as follows:

  – Define a sequence of values for $\gamma = (\gamma_1, \gamma_2, ...\gamma_g)$.

  – Calculate the probability for each γ_j via the posterior probability as follows:

$$p(\gamma_j|\boldsymbol{s}^{(i)}) \propto \frac{1}{\sum_{|\boldsymbol{k}|=1}^{k_n} \binom{p}{|\boldsymbol{k}|}^{1-\gamma_j} / \binom{p}{|\boldsymbol{s}^{(i)}|}^{-\gamma_j}} \mathbb{I}(|\boldsymbol{s}^{(i)}| \leq k_n) \mathbb{I}_{(1-\frac{\log n}{2\log p}, \infty)}(\gamma_j).$$

  – Randomly pick up a value $\gamma^* \in (\gamma_1, \gamma_2, ...\gamma_g)$ with the success probability of w_j for each γ_j

$$w_j = \frac{p(\gamma_j|\boldsymbol{s}^{(i)})}{\sum_{j=1}^{g} p(\gamma_j|\boldsymbol{s}^{(i)})},$$

update $\gamma = \gamma^*$.

- Step 4: Repeat Step 2 and 3 for R times after some burning period (throw away a certain number of the initially generated samples), and obtain the R samples $(\boldsymbol{s}^{(1)}, ...\boldsymbol{s}^{(R)})$. Calculate the fraction of times each variable is selected in the R samples:

$$p_j = \frac{\sum_{i=1}^{R} \mathbb{I}(s_j^{(i)} = 1)}{R}.$$

- Step 5: The best predictive model $\boldsymbol{s}^*$ is defined as the one containing the covariates with $p_i \geq 1/2$.

$$s_j^* = \begin{cases} 1, & \text{if } p_j \geq \dfrac{1}{2}, \\ 0, & \text{othewise.} \end{cases}$$

Denote $\boldsymbol{s} = (s_1..., s_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $s_j \in \{0, 1\}$ for $j = 1, 2, ..., p$. The details of the proposed algorithm works as follows:

Algorithm 2 Best predictive model selection

Start from an initial state of $\boldsymbol{s}^{(0)} = (s_1^{(0)}, s_2^{(0)}, ..., s_p^{(0)})$, and an initial value of $\gamma = \gamma_0$, set the upper bound for the number of covariates as k_n, the algorithm works as the following steps:

- Step 1: For $j = 1, 2, ..., p$, generate a new candidate model $\boldsymbol{s}^{(i)}$ based on the current candidate model $\boldsymbol{s}^{(i-1)}$ as follows:

  (a) Define $\boldsymbol{s}^{(i)}$ by $s_j^{(i)} = 1 - s_j^{(i-1)}$ and $s_l^{(i)} = s_l^{(i-1)}$ for $l \neq j$.

  (b) If $\sum_{j=1}^{p} s_j^{(i)} > k_n$, skip Step (c) and move to $s_{j+1}^{(i)}$.
  Otherwise, calculate EBIC($\boldsymbol{s}^{(i)}$) and EBIC($\boldsymbol{s}^{(i-1)}$)

  (c) We keep $s_j^{(i)} = s_j^{(i)}$ if we obtain 1 from a Bernoulli trial with the success probability:

$$w = \frac{\exp\left\{-\frac{1}{\kappa\tau}\text{EBIC}(\boldsymbol{s}^{(i)})\right\}}{\exp\left\{-\frac{1}{\kappa\tau}\text{EBIC}(\boldsymbol{s}^{(i)})\right\} + \exp\left\{-\frac{1}{\kappa\tau}\text{EBIC}(\boldsymbol{s}^{(i-1)})\right\}}.$$

  Otherwise , update $s_j^{(i)} = s_j^{(i-1)}$

- Step 2: Given $\boldsymbol{s}^{(i)}$, define a sequence of values for $\gamma = (\gamma_1, \gamma_2, ...\gamma_g)$ and pick up a sample $\gamma* \in (\gamma_1, \gamma_2, ...\gamma_g)$ with the success probability

$$w_{\gamma_j} = \frac{p(\gamma_j|\boldsymbol{s}^{(i)})}{\sum_{j=1}^{g} p(\gamma_j|\boldsymbol{s}^{(i)})},$$
$$p(\gamma_j|\boldsymbol{s}^{(i)}) \propto \frac{1}{\sum_{|\boldsymbol{k}|=1}^{k_n} \binom{p}{|\boldsymbol{k}|}^{1-\gamma_j} / \binom{p}{|\boldsymbol{s}^{(i)}|}^{-\gamma_j}} \mathbb{I}(|\boldsymbol{s}^{(i)}| \leq k_n) \mathbb{I}_{(1-\frac{\log n}{2\log p}, \infty)}(\gamma_j),$$

  update $\gamma = \gamma^*$

- Step 3: Repeat Step 1 and 2 for R times after some burning period and obtain R samples $(\boldsymbol{s}^{(1)}, ...\boldsymbol{s}^{(R)})$. Calculate the proportion for each of the element equals 1:

$$p_j = \frac{\sum_{i=1}^{R} \mathbb{I}(s_j^{(i)} = 1)}{R}.$$

- Step 3: The best model ($\boldsymbol{s}^*$) is the one containing the covariates with $p_i \geq 1/2$.

$$s_j^* = \begin{cases} 1, & \text{if } p_j \geq \dfrac{1}{2}, \\ 0, & \text{othewise.} \end{cases}$$

3.4 Simulation study

In this section, we conduct the simulation study to investigate the predictive performance of our proposed method under the accelerate failure time (AFT) model.

When we implement our proposed method in Algorithm 2, we set $\boldsymbol{s}^0 = (0, 0, ..., 0)$, $R = 200$, $\tau = 1$, $\kappa = 2$ and $k_n = n_0^{3/4}$ (n_0 gives the number of uncensored observations in the survival data) for our initial setting, called "Proposed method". The first 20% generated samples are thrown away as the burning period. To investigate the benefit of treating γ as a random variable, we also consider fixing γ as one, called "Proposed method ($\gamma = 1$)".

To compare the performance, we consider the following three popular methods which are commonly employed in high-dimensional variable selection: (1) lasso (Tibshirani, 1996), (2) SCAD (Fan and Li, 2001), (3) MCP (Zhang, 2010). The simulation study is conducted by R, where lasso, SCAD and MCP are conducted by the `cv.nvcsurv` function in the `ncvreg` package, under which the 10-fold cross-validation is employed for finding the best tuning parameter to determine the selected model.

The simulation data are separately generated for two parts, one for the training data and another for the testing data. For training data, the survival time T_i^* and censoring time C_i for $i = 1, \cdots, n$ independent individuals are generated as follows:

- $\mathbf{x}_i \overset{iid}{\sim} \mathrm{N}_p(0, \Sigma)$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$.
- $\boldsymbol{\beta}$: the p-dimensional coefficient vector with values: $\beta_1 = \beta_9 = 1.4$, $\beta_4 = \beta_{12} = -1.3$, $\beta_5 = \beta_{13} = 1.2$ and $\beta_j = 0$ for $j \neq 1, 4, 5, 9, 12, 13$.
- $T_i^* = -\log(U_i) \times \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, where $U_i \overset{iid}{\sim} \mathrm{Unif}(0, 1)$.
- $C_i \overset{iid}{\sim} \mathrm{Exp}(\eta)$, where η=0.22 (for about 25% censoring rate), or η=(0.47,0.46,0.45) (for about 40% censoring rate).

For the testing data, the survival time T_i^{**} for $i = 1, \cdots, n$ independent individuals are generated as follows:

- $\mathbf{x}_i^{**} \overset{iid}{\sim} \mathrm{N}_p(0, \Sigma)$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$.

- $\boldsymbol{\beta}$: $\beta_1 = \beta_9 = 1.4$, $\beta_4 = \beta_{12} = -1.3$, $\beta_5 = \beta_{13} = 1.2$ and $\beta_j = 0$ for $j \neq 1, 4, 5, 9, 12, 13$.
- $T_i^{**} = -\log(U_i) \times \exp(\mathbf{x}_i^{**\mathrm{T}}\boldsymbol{\beta})$, where $U_i \overset{iid}{\sim} \mathrm{Unif}(0, 1)$.

To consider various high-dimensional data settings, we implement the following six scenarios in the data generating process:

Training data:

(1) $n = 200$, $p = 100$, censor rate= 25% censoring time $\boldsymbol{C} \sim \exp(0.12)$.

(2) $n = 300$, $p = 100$, censor rate= 40% censoring time $\boldsymbol{C} \sim \exp(0.47)$.

(3) $n = 200$, $p = 500$, censor rate= 25% censoring time $\boldsymbol{C} \sim \exp(0.12)$.

(4) $n = 300$, $p = 500$, censor rate= 40% censoring time $\boldsymbol{C} \sim \exp(0.46)$.

(5) $n = 200$, $p = 1000$, censor rate= 25% censoring time $\boldsymbol{C} \sim \exp(0.12)$.

(6) $n = 300$, $p = 1000$, censor rate= 40% censoring time $\boldsymbol{C} \sim \exp(0.45)$.

Testing data:

(1) $n^* = 100$, $p = 100$.

(2) $n^* = 100$, $p = 100$.

(3) $n^* = 100$, $p = 500$.

(4) $n^* = 100$, $p = 500$.

(5) $n^* = 100$, $p = 1000$.

(6) $n^* = 100$, $p = 1000$.

To evaluate the performance of prediction, we calculate the mean squared prediction error (MSPE) for each method over 100 Monte Carlo replications, where the MSPE is calculated as:

$$\mathrm{MSPE} = \frac{\sum_{i=1}^{100}\sum_{j=1}^{n^*}(\hat{y}_{ij} - y_{ij})^2}{100 \times n^*},$$

where $\hat{y}_{ij} = \log(\hat{T}_{ij})$ is the log of the predicted survival time for the jth individual from the ith testing dataset, and $y_{ij} = \log(T_{ij})$ is the log of the observed survival time.

In addition, to access the model identification performance, we calculate the false-positive rate (FPR) and false-negative rate (FNR),

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}}, \quad \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}},$$

where TP, FP, TN and FN denote the number of true non-zeros, false non-zeros, true zeros, and false zeros respectively.

Table 3.1: *Simulation result with censoring rate=25%*

(n, p)	Method	MSPE	FPR	FNR
(200,100)	Proposed method	1.6727 (0.0326)	0.0023 (0.0005)	0.0000 (0.0000)
	Proposed method($\gamma = 1$)	1.6784 (0.0325)	0.0029 (0.0006)	0.0000 (0.0000)
	LASSO	2.1605 (0.0414)	0.2566 (0.0064)	0.0000 (0.0000)
	SCAD	1.7283 (0.0340)	0.0178 (0.0022)	0.0000 (0.0000)
	MCP	1.6969 (0.0339)	0.0090 (0.0014)	0.0000 (0.0000)
(200,500)	Proposed method	1.6724 (0.0318)	0.0022 (0.0005)	0.0000 (0.0000)
	Proposed method($\gamma = 1$)	1.6776 (0.0323)	0.0029 (0.0006)	0.0000 (0.0000)
	LASSO	2.7039 (0.0556)	0.3397 (0.0057)	0.0000 (0.0000)
	SCAD	1.8293 (0.0362)	0.0416 (0.0039)	0.0000 (0.0000)
	MCP	1.7259 (0.0344)	0.0122 (0.0021)	0.0000 (0.0000)
(200,1000)	Proposed method	1.6730 (0.0313)	0.0026 (0.0005)	0.0000 (0.0000)
	Proposed method($\gamma = 1$)	1.6775 (0.0312)	0.0030 (0.0006)	0.0000 (0.0000)
	LASSO	2.9797 (0.0531)	0.3757 (0.0053)	0.0000 (0.0000)
	SCAD	1.8768 (0.0365)	0.0617 (0.0043)	0.0000 (0.0000)
	MCP	1.7302 (0.0341)	0.0140 (0.0022)	0.0000 (0.0000)

Table 3.2: *Simulation result with censoring rate=40%*

(n, p)	Method	MSPE	FPR	FNR
(300,100)	Proposed method	1.6564 (0.0304)	0.0007 (0.0003)	0.0000 (0.0000)
	Proposed method($\gamma = 1$)	1.6602 (0.0305)	0.0012 (0.0003)	0.0000 (0.0000)
	LASSO	2.0389 (0.0378)	0.2793 (0.0076)	0.0000 (0.0000)
	SCAD	1.6890 (0.0319)	0.0140 (0.0028)	0.0000 (0.0000)
	MCP	1.6719 (0.0316)	0.0049 (0.0013)	0.0000 (0.0000)
(300,500)	Proposed method	1.6553 (0.0309)	0.0015 (0.0004)	0.0000 (0.0000)
	Proposed method($\gamma = 1$)	1.6587 (0.0310)	0.0018 (0.0047)	0.0000 (0.0000)
	LASSO	2.5785 (0.0464)	0.3581 (0.0070)	0.0000 (0.0000)
	SCAD	1.7502 (0.0330)	0.0276 (0.0048)	0.0000 (0.0000)
	MCP	1.6873 (0.0313)	0.0086 (0.0023)	0.0000 (0.0000)
(300,1000)	Proposed method	1.6533 (0.0302)	0.0018 (0.0005)	0.0000 (0.0000)
	Proposed method($\gamma = 1$)	1.6638 (0.0304)	0.0023 (0.0006)	0.0000 (0.0000)
	LASSO	2.7310 (0.0482)	0.4072 (0.0062)	0.0000 (0.0000)
	SCAD	1.7849 (0.0347)	0.0395 (0.0058)	0.0000 (0.0000)
	MCP	1.7002 (0.0314)	0.0124 (0.0026)	0.0000 (0.0000)

The simulation result is summarized in Tables 3.1 and 3.2. From the result above, we can find that our proposed method always gives the smallest value of MSPE among all the six scenarios, which means the proposed method performs best in prediction among all methods. Also, treating γ in the EBIC as a random variable gives a smaller value of MSPE than fixing γ as 1, which implies the posterior distribution we assigned to γ provides efficiency in best predictive model selection.

For model identification, all of the methods get 0 for FPR, which means they all have the ability to identify the true-positive (non-zero) predictors. However, for FNR, our proposed method always provides the smallest value, which means the proposed method is able to select

the model with the smallest number of redundant predictors. Also, treating γ in EBIC as a random variable gives a smaller value of FPR than fixing it as 1, which implies the posterior distribution we assigned to γ provides efficiency in improving model identification. Among other methods, MCP gives the smallest value of FPR, while lasso provides the largest value, which implies that it tends to select too many redundant variables.

The distribution of γ in EBIC is shown in 3.1. As can be seen in the histogram, the majority of the values of γ are generated inside a certain range of (1.0, 1.2). Despite the fact that the generated values are quite close to one, considering γ as a random variable still provides the advantages of improving the prediction precision and model identification accuracy.

In summary, compared with lasso, SCAD, and MCP, our proposed method performs best both in the performance of prediction and model identification. Also, treating γ in EBIC as a random variable gives better results than fixing it as 1, which implies the posterior distribution we assigned to γ provides benefits for high-dimensional survival model selection.

3.5 Real data application

In this section, we conduct the real data analysis with the Veteran's Administration lung cancer trial data (Kalbfleisch and Prentice, 2011). The dataset is publicly available and can be accessed from the R package `ncvreg`.

The original data contains information about 137 patients with eight covariates: stime (survival or follow-up time in days), status (dead or censored), treat (treatment: standard or test), age (patient's age in years), Karn (Karnofsky score of patient's performance on a scale of 0 to 100), diag.time (times since diagnosis in months at entry to trial), cell (one of four cell types), and prior (prior therapy or not).

In order to meet the high-dimensional setting, we generate 800 fake variables, each of the fake variables is generated based on bootstrapping from one of the eight variables in the original dataset. In this way, we have a total number of 808 covariates. Among the 137 patients, 128 of them have observed survival times, while the remaining nine patients'

survival times are censored, and the censoring rate is 6.57%.

When implementing the algorithm, we randomly split the data into two parts. The training data contains 80% of the randomly selected observed individuals and 80% of the randomly selected censored individuals, while the testing data contains the remaining 20% observed individuals. We repeat the random splitting 100 times, each time we apply our method and compare it with lasso, SCAD, and MCP in terms of MSPE. For each of the replications in our proposed method, we run the Gibbs sampling 300 times and throw away the first 100 generated samples as the burning period.

Table 3.3: *Real data analysis with Veteran's Administration lung cancer trial data*

	MSPE	Average # of covariates selected
Proposed	1.6132 (0.0498)	1.07
LASSO	1.8439 (0.0609)	11.17
SCAD	1.8331 (0.0595)	8.51
MCP	1.7464 (0.0611)	3.72

The comparison result is shown in Table 3.3. Compared with other methods, it is worth noticing that, on average, our proposed method provides the smallest value of the MSPE, which means it finds the model with the smallest prediction error. Hence, we can conclude that our proposed method performs best for the best predictive model selection.

3.6 Concluding remarks

We propose the best predictive model selection method by incorporating the idea of the median probability model with the Boltzmann distribution. The algorithm translated the initial notion from a Bayesian framework to a new idea of the frequentist framework. It provides a way of selecting the model with the best prediction performance for high-dimensional survival data. The superiority of the proposed method has been demonstrated by the simulation study and real data analysis.

While we focus our attention on the AFT model with Weibull distribution in this chapter,

the proposed method can be easily adapted to different parametric survival models with different types of data distribution. In addition, various choices of GIC can be taken into account in our proposed framework.

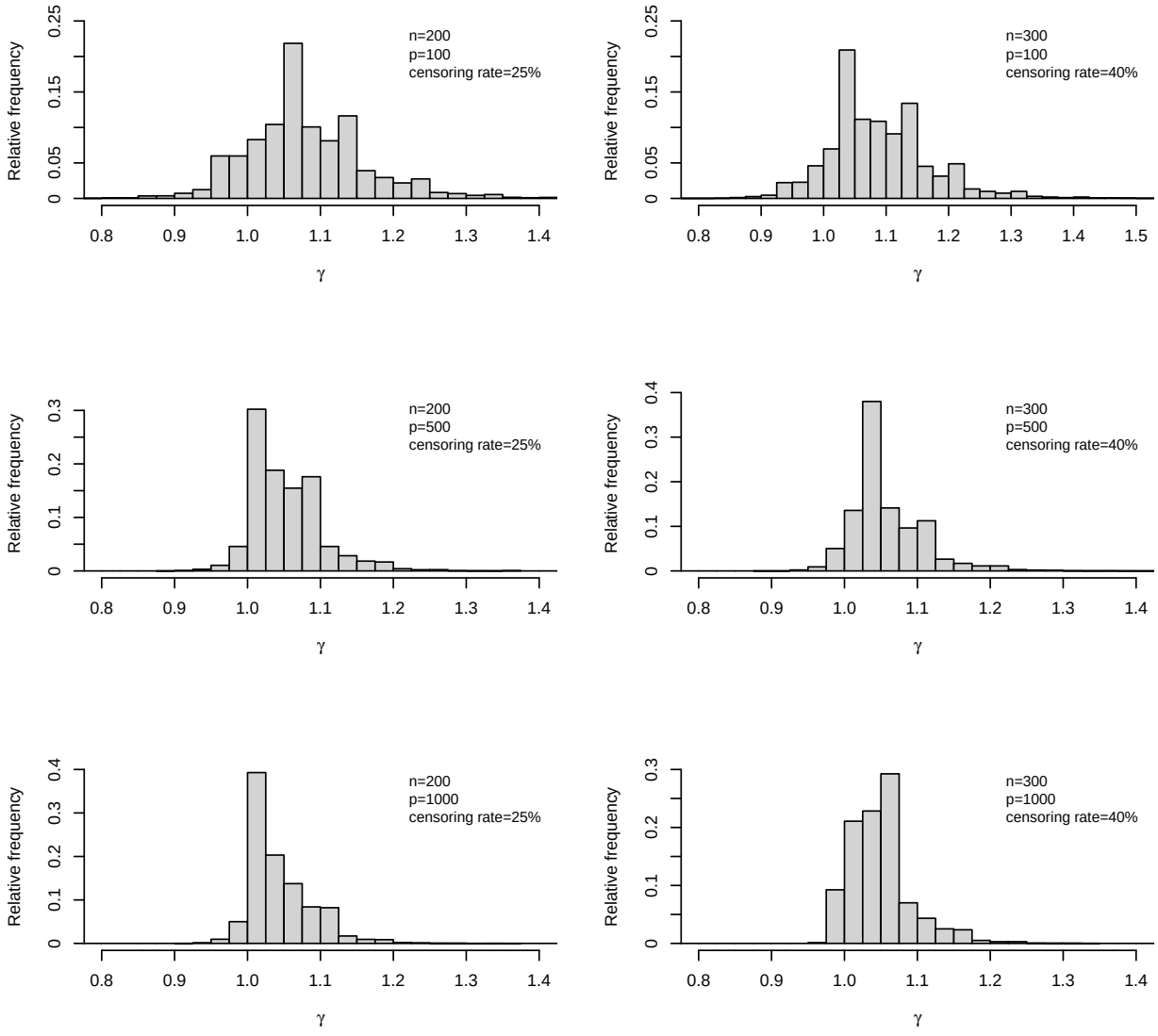

Figure 3.1: *The distribution of γ when treating it as a random variable in EBIC under different sample sizes (n), number of covariates (p), and censoring rates.*

Chapter 4

Robust variable selection approach to survival regression models

4.1 Introduction

In survival analysis, model selection has attracted growing attention from data scientists. In particular, with the advent of high-dimensional data, variable selection has become a critical step in overcoming the curse of dimensionality. Some of the semi-parametric and parametric models, like the Cox proportional hazards model (Cox, 1972) and the accelerated failure time (AFT) model (Wei, 1992), are widely utilized for survival data analysis. Despite the fact that these models are popular in survival analysis, they have some strict assumptions to reveal the relationship between the response and the covariates to guarantee some important properties, like the identifiability and consistency of the model selection results. However, in the real-life world, these assumptions can be violated (Gail et al., 1984; Lagakos et al., 1984; Lagakos, 1988; Morgan et al., 1986; O'neill, 1986; Solomon, 1984; Struthers and Kalbfleisch, 1986). In this chapter, we will develop a robust model selection method in the presence of misspecification problems.

In a parametric survival modeling context, the robust covariance matrix estimator, called the "sandwich" variance estimator, has been extensively studied (Gail et al., 1988; Huber,

1967; Kent, 1982; Royall, 1986; White, 1982). The "sandwich" estimator is known to be a proper variance estimator even when the model is misspecified. Borrowing the idea of the sandwich estimator, we develop a pseudo log-likelihood estimator, which will be used for developing a robust model selection criterion. However, finding the best model for the proposed model selection criterion is challenging since it can be used only for pairwise model comparison. Using the notion of simulated annealing (Kirkpatrick et al., 1983), we develop a probabilistic search algorithm for finding the global optimum model with respect to the proposed model selection criterion.

The rest of this chapter is organized as follows. Model misspecification and robust inference are discussed in Chapter 4.2. In Chapter 4.3, we propose a robust model selection criterion for pairwise comparison. In Chapter 4.4, we extend the proposed idea to high-dimensional survival model selection problems. In Chapter 4.5, future works are discussed.

4.2 Model misspecification investigation

The problem of model misspecification occurs in many scenarios with real data since some structural assumptions can be easily violated. For example, the Cox proportional hazards model assumes the following relationship between failure time t and covariates vector $\mathbf{x}$:

$$h(t) = h_0(t) \times \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \tag{4.1}$$

where $h_0(t)$ is the unspecified baseline hazard function, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the p-dimensional coefficient vector, which is of our primary interest. Under the Cox proportional hazards model (4.1) with the right-censored scenario, estimation of $\boldsymbol{\beta}$ can be obtained by maximizing the partial likelihood function (Cox, 1975),

$$L(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\sum_{l \in R(T_i)} \exp(\mathbf{x}_l^{\mathrm{T}}\boldsymbol{\beta})} \right]^{\delta_i},$$

where $R(T_i)$ is the risk set at time T_i, which represents the number of individuals who survived at least until time T_i. Then the corresponding partial log-likelihood function is given as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \log \left\{ \sum_{l \in R(T_i)} \exp(\mathbf{x}_l^{\mathrm{T}} \boldsymbol{\beta}) \right\} \right].$$

If the assumption (4.1) is true, then the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ can have the asymptotic normality as follows (Andersen and Gill, 1982) :

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, V), \tag{4.2}$$

where V can be consistently estimated by $\hat{A}(\hat{\boldsymbol{\beta}})^{-1}$ such that

$$\hat{A}(\hat{\boldsymbol{\beta}}) = -\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

However, when the assumption (4.1) is violated, $\hat{A}(\hat{\boldsymbol{\beta}})^{-1}$ will not be a consistent estimator of V (Gail et al., 1984; Lagakos et al., 1984; Lagakos, 1988; Morgan et al., 1986; O'neill, 1986; Solomon, 1984; Struthers and Kalbfleisch, 1986).

Another example comes from a widely used parametric model, the so-called accelerated failure time (AFT) model. The AFT model is defined by

$$Y_i = \log T_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \sigma \epsilon_i, \ i = 1, \ldots, n,$$

where σ is the scale parameter, and ϵ_i is the random disturbance term, usually assumed to be independent and identically distributed with some density function $f(\epsilon)$. For survival

data, the likelihood function is obtained as

$$\begin{aligned} L(\boldsymbol{\beta}) &= \left\{ \prod_{i:\delta_i=1} f(T_i \mid \boldsymbol{\beta}) \right\} \times \left\{ \prod_{i:\delta_i=0} [1 - F(T_i \mid \boldsymbol{\beta})] \right\} \\ &= \prod_{i=1}^{n} f(T_i \mid \boldsymbol{\beta})^{\delta_i} S(T_i \mid \boldsymbol{\beta})^{1-\delta_i}, \\ &= \prod_{i=1}^{n} \left[\frac{f(T_i \mid \boldsymbol{\beta})}{S(T_i \mid \boldsymbol{\beta})} \right]^{\delta_i} S(T_i, \mid \boldsymbol{\beta}), \end{aligned}$$

where $\delta_i = I\{T_i < C_i\}$ is the censoring indicator for individuals $i = 1, \cdots, n$, $f(\cdot \mid \boldsymbol{\beta})$ is the probability density function, $F(\cdot \mid \boldsymbol{\beta})$ represents the cumulative distribution function, $S(\cdot \mid \boldsymbol{\beta})$ is the survival function, and $h(\cdot \mid \boldsymbol{\beta})$ denotes the hazard function. By assuming the Weibull distribution for the AFT model, we have

$$L(\boldsymbol{\beta}, \sigma | \boldsymbol{T}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \left\{ \frac{1/\sigma}{\exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})} \left[\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})} \right]^{\frac{1}{\sigma}-1} \right\}^{\delta_i} \exp\left[- \left(\frac{t}{\exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})} \right)^{\frac{1}{\sigma}} \right].$$

Then, the log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma | \boldsymbol{T}, \boldsymbol{\delta}) = & \log \frac{1}{\sigma} \sum_{i=1}^{n} \delta_i + \left(\frac{1}{\sigma} - 1 \right) \sum_{i=1}^{n} \delta_i (\log t_i) \\ & - \frac{1}{\sigma} \sum_{i=1}^{n} \delta_i (\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) - \sum_{i=1}^{n} \left(\frac{t_i}{\exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})} \right)^{\frac{1}{\sigma}}. \end{aligned}$$

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ under the AFT model. Then, valid inference can be made by using the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. However, when the AFT model is misspecified, making a valid inference is difficult since the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ cannot be properly estimated (Hattori, 2012; Ishii et al., 2021).

As illustrated above, the validity of the likelihood function (or the partial likelihood function) hinges on the correct specification of the model. To make the robust inference, the use of the sandwich estimator has been extensively studied (Gail et al., 1988; Huber, 1967; Kent, 1982; Royall, 1986; White, 1982). Assume $\boldsymbol{\theta}$ is the parameter of interest. Let $\hat{\boldsymbol{\theta}}$ be the

maximum likelihood estimator of $\boldsymbol{\theta}$ with respect to the log-likelihood function $l(\boldsymbol{\theta})$. Then, under some mild regularity conditions, it has been shown that V in (4.2) can be consistently estimated by the sandwich estimator,

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \hat{A}^{-1}(\hat{\boldsymbol{\theta}})\hat{B}(\hat{\boldsymbol{\theta}})\hat{A}^{-1}(\hat{\boldsymbol{\theta}}),$$

where

$$\hat{B}(\boldsymbol{\theta}) \;=\; \frac{1}{n}\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

As a result, the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ can be properly estimated by $\hat{V}(\hat{\boldsymbol{\theta}})/n$ even with a misspecified log-likelihood function. This motivates us to develop a robust pseudo log-likelihood estimator based on the sandwich estimator. More details are given in the following section.

4.3 Robust model selection criterion

As discussed in Chapter 2.2, the following generalized information criterion (GIC) plays an important role in model selection with high-dimensional survival data:

$$\text{GIC}(\boldsymbol{s}) = -2l(\hat{\boldsymbol{\beta}}(\boldsymbol{s})) + \text{pen}(|\boldsymbol{s}|),$$

where $\boldsymbol{s} = (s_1, \ldots, s_p)$ represents a candidate model such that $\beta_j \neq 0$ if $s_j = 1$ and $\beta_j = 0$ if $s_j = 0$ for $j = 1, \ldots, p$, $|\boldsymbol{s}| = \sum_{j=1}^{p} s_j$, $\boldsymbol{\beta}(\boldsymbol{s})$ denotes the sub-vector of $\boldsymbol{\beta}$ corresponding to $s_j = 1$ $(j = 1, \ldots, p)$, and $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ is the maximum likelihood estimation of $\boldsymbol{\beta}(\boldsymbol{s})$ under model $\boldsymbol{s}$.

In GIC, the log-likelihood, $l(\hat{\boldsymbol{\beta}}(\boldsymbol{s}))$, depends on the model assumption. In other words, model selection using GIC can suffer from the problem of model misspecification. To address this issue, using the sandwich estimator, we propose a robust pseudo log-likelihood under

the full model as follows:

$$-2\tilde{l}(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under the full model and $\hat{\Sigma}$ is the sandwich estimator of $\mathrm{var}(\hat{\boldsymbol{\beta}})$ under the full model. Let $\tilde{\boldsymbol{\beta}}(\boldsymbol{s})$ be the $p \times 1$ vector which consists of $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ and zeros accordingly. Then, the robust pseudo log-likelihood under model $\boldsymbol{s}$ is obtained as

$$-2\tilde{l}(\tilde{\boldsymbol{\beta}}(\boldsymbol{s})) = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}(\boldsymbol{s}))^{\mathrm{T}}\hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}(\boldsymbol{s})).$$

Hence, we define the robust model selection criterion, called pseudo-GIC by

$$\text{pseudo-GIC}(\boldsymbol{s}) = -2\tilde{l}(\tilde{\boldsymbol{\beta}}(s)) + \mathrm{pen}(|\boldsymbol{s}|). \tag{4.3}$$

One of the limitations of the proposed idea is that the dimension of the full model must be much lower than the sample size. However, in the high-dimensional data setting, the full model size is large, even larger than the sample size. In this case, the proposed idea is infeasible. To address this issue, we modify our idea for pairwise model comparison.

4.3.1 Robust pairwise model comparison

Assume we have two candidate models, $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, such that $\boldsymbol{s}_1$ is nested in $\boldsymbol{s}_2$ (i.e., $\boldsymbol{s}_1 \subset \boldsymbol{s}_2$). Then, by treating $\boldsymbol{s}_2$ as the full model, a modified pseudo log-likelihood can be defined by

$$-2\tilde{l}(\boldsymbol{\beta}(\boldsymbol{s}_2)) = (\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \boldsymbol{\beta}(\boldsymbol{s}_2))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}_2}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \boldsymbol{\beta}(\boldsymbol{s}_2)),$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2)$ is the maximum likelihood estimator of $\boldsymbol{\beta}(\boldsymbol{s}_2)$ and $\hat{\Sigma}_{\boldsymbol{s}_2}$ is the sandwich estimator of $\mathrm{var}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2))$. Let $\tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)$ be $|\boldsymbol{s}_2| \times 1$ vector, which consists of $\hat{\boldsymbol{\beta}}(\boldsymbol{s}_1)$ and zeros accordingly. Hence, the pseudo log-likelihood evaluated at $\tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)$ is given as

$$-2\tilde{l}(\tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)) = (\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}_2}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)).$$

Hence, the difference of pseudo log-likelihood between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ can be defined by

$$\begin{aligned}\Delta_{\boldsymbol{s}_1\boldsymbol{s}_2} =& 2\tilde{l}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2)) - 2\tilde{l}(\tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)) \\ =& 0 + (\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}_2}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)) \\ =& (\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}_2}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)).\end{aligned}$$

As a result, pairwise model comparison can be performed by using the following model selection criterion:

$$\begin{aligned}\delta_{\boldsymbol{s}_1\boldsymbol{s}_2} &= \Delta_{\boldsymbol{s}_1\boldsymbol{s}_2} + \mathrm{pen}(|\boldsymbol{s}_1|) - \mathrm{pen}(|\boldsymbol{s}_2|) \\ &= (\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}_2}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}_2) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}_2}(\boldsymbol{s}_1)) + \mathrm{pen}(|\boldsymbol{s}_1|) - \mathrm{pen}(|\boldsymbol{s}_2|).\end{aligned}$$

If $\delta_{\boldsymbol{s}_1\boldsymbol{s}_2} < 0$, model $\boldsymbol{s}_1$ is better than $\boldsymbol{s}_2$. Otherwise, $\boldsymbol{s}_2$ should be preferred than $\boldsymbol{s}_1$.

Now, we consider a more general case. Assume that we have two candidate models, $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, such that $\boldsymbol{s}_1 \not\subset \boldsymbol{s}_2$ and $\boldsymbol{s}_1 \not\supset \boldsymbol{s}_2$. In this case, the full model can be defined by $\boldsymbol{s}^* = \boldsymbol{s}_1 \cup \boldsymbol{s}_2$ such that $s_j^* = 1$ if $s_{1j} = 1$ or $s_{2j} = 1$. Then, the difference of pseudo log-likelihood between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ can be expressed as

$$\begin{aligned}\tilde{\Delta}_{\boldsymbol{s}_1\boldsymbol{s}_2} &= \Delta_{\boldsymbol{s}_1\boldsymbol{s}^*} - \Delta_{\boldsymbol{s}_2\boldsymbol{s}^*} \\ &= (\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_1))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}^*}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_1)) \\ &\quad -(\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_2))^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{s}^*}^{-1}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_2)),\end{aligned}$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*)$ is the maximum likelihood estimator of $\boldsymbol{\beta}(\boldsymbol{s}^*)$ and $\hat{\Sigma}_{\boldsymbol{s}^*}$ is the sandwich estimator of $\mathrm{var}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*))$. Hence, pairwise model comparison can be performed by the following model selection criterion:

$$\tilde{\delta}_{\boldsymbol{s}_1\boldsymbol{s}_2} = \tilde{\Delta}_{\boldsymbol{s}_1\boldsymbol{s}_2} + \mathrm{pen}(|\boldsymbol{s}_1|) - \mathrm{pen}(|\boldsymbol{s}_2|).$$

If $\delta_{\boldsymbol{s}_1\boldsymbol{s}_2} < 0$, $\boldsymbol{s}_1$ is preferred than $\boldsymbol{s}_2$. Otherwise, $\boldsymbol{s}_2$ is preferred. The proposed model selection

criterion includes many existing model selection criteria as a special case. For example, suppose that we consider the extended Bayesian information criterion (EBIC) (Chen and Chen, 2008), which is one of the most popular model selection criteria for high-dimensional data:

$$\mathrm{EBIC}_\gamma(\boldsymbol{s}) = -2l(\hat{\boldsymbol{\beta}}(\boldsymbol{s})) + |\boldsymbol{s}| \log(n_0) + 2\gamma \log \binom{p}{|\boldsymbol{s}|},$$

where γ is a tuning parameter between 0 and 1. Then, the proposed model selection criterion for pairwise comparison reduces to

$$\begin{aligned}
\tilde{\delta}\mathrm{EBIC}_{\boldsymbol{s}_1\boldsymbol{s}_2} &= \tilde{\Delta}_{\boldsymbol{s}_1\boldsymbol{s}_2} + \mathrm{pen}(|\boldsymbol{s}_1|) - \mathrm{pen}(|\boldsymbol{s}_2|) \\
&= (\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_1))^{\mathrm{T}} \hat{\Sigma}_{\boldsymbol{s}^*}^{-1} (\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_1)) \\
&\quad -(\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_2))^{\mathrm{T}} \hat{\Sigma}_{\boldsymbol{s}^*}^{-1} (\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - \tilde{\boldsymbol{\beta}}_{\boldsymbol{s}^*}(\boldsymbol{s}_2)) \\
&\quad +(|\boldsymbol{s}_1| - |\boldsymbol{s}_2|) \log(n_0) + 2\gamma \log \left\{ \binom{p}{|\boldsymbol{s}_1|} / \binom{p}{|\boldsymbol{s}_2|} \right\}.
\end{aligned}$$

4.3.2 Limitation of pairwise comparison

The proposed model selection criterion provides a robust method for pairwise model comparison. Keep in mind that our main objective is to identify the global optimal model among multiple candidate models. Unfortunately, due to the non-convexity of the model selection criterion, this pairwise comparison approach can not be directly used for the purpose of multiple model comparisons.

As discussed in Chapter 2, the stochastic optimization approach can be employed to identify the global optimum model in a non-convex optimization problem. Specifically, we repeatably implement the pairwise comparison procedure within the proposed simulated annealing algorithm. More details are given in the following section.

4.4 Robust global optimal model selection

To extend our proposed idea to multiple model comparisons, we borrow the idea of simulated annealing (SA) from Chapter 2.4.1. Let $\hat{\boldsymbol{s}} = (\hat{s}_1, \hat{s}_2, ..., \hat{s}_p)$ indicate the global best model, that is, $\hat{\boldsymbol{s}}$ satisfies $\tilde{\delta}_{\hat{\boldsymbol{s}},\boldsymbol{s}} < 0$ for every candidate model $\boldsymbol{s}$. However, when there are many candidate models, it is computationally infeasible to consider all possible combinations of pairs. To address this issue, we propose the following stochastic search procedure:

- Step 1: Start from an initial state of $\boldsymbol{s} = (s_1, s_2, ..., s_p)$ with an initial temperature $\tau = \tau_0$, use $\hat{\boldsymbol{s}} = (\hat{s}_1, \cdots, \hat{s}_p)$ to store the best model, set $r = 0$, which counts the number of iterations, and set an initial value for k, which controls the maximum number of covariates selected in the model.

- Step 2: Iterate the following steps for $j = 1, \ldots, p$:

  - Generate $\boldsymbol{s}^* = (s_1^*, s_2^*, \cdots, s_p^*)$ by setting $s_j^* = 1 - s_j$ and $s_\ell^* = s_\ell$ for $\ell \neq j$.

  - If $\sum_{j=1}^{p} s_j^* > k$, move to the next iteration; otherwise, calculate $\delta_{\boldsymbol{s}\boldsymbol{s}^*}$ and proceed the following steps:

    (i) If $\delta_{\boldsymbol{s}\boldsymbol{s}^*} > 0$, calculate $\tilde{\delta}_{\boldsymbol{s}^*\hat{\boldsymbol{s}}}$:

    If $\tilde{\delta}_{\boldsymbol{s}^*\hat{\boldsymbol{s}}} < 0$, update $\hat{\boldsymbol{s}} = \boldsymbol{s}^*$ and set $r = 0$, $\tau = \tau_0$. Otherwise, set $r = r + 1$.

    (ii) If $\delta_{\boldsymbol{s}\boldsymbol{s}^*} \leq 0$, calculate $\tilde{\delta}_{\boldsymbol{s}\hat{\boldsymbol{s}}}$:

    If $\tilde{\delta}_{\boldsymbol{s}\hat{\boldsymbol{s}}} < 0$, update $\hat{\boldsymbol{s}} = \boldsymbol{s}$ and set $r = 0$, $\tau = \tau_0$. Otherwise, set $r = r + 1$.

  - Generate $z \sim \text{Ber}(\omega)$, where

$$\omega = \frac{1}{1 + \exp[-\frac{\delta_{\boldsymbol{s}\boldsymbol{s}^*}}{\kappa\tau}]}.$$

    If $z = 1$, update $\boldsymbol{s} = \boldsymbol{s}^*$. Otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.

- Step 3: Repeat Step 2 until $r > pm$ (m has a prespecified value).

- Step 4: Repeat Step 2 and 3 with $\boldsymbol{s} = \hat{\boldsymbol{s}}$, $r = 0$ for a sequences of values of $\tau = \{\tau_2, \tau_3, \cdots, \tau_{\max}\}$, where $\tau_{t+1} > \tau_t$, until $\tau = \tau_{\max}$, where $\tau_{\max}$ is the maximum temperature with a prespecified value.

Note that the final model $\hat{\boldsymbol{s}}$ from the proposed algorithm will converge to the global optimum model as $r \to \infty$. Also, note that the temperature τ is introduced to improve the chance of jumping out from the local trap.

4.5 Future work and discussion

In this chapter, we have developed a robust model selection criterion for addressing model misspecification problems. An important feature is that the proposed framework includes many model selection criteria as a special case. In addition, the proposed algorithm has wide applicability to high-dimensional model comparison. While traditional model selection procedures require the structural assumption for the data to construct the log-likelihood function, the proposed model comparison method only requires the regression coefficient estimate and its variance estimator. It is important to note that the performance of the proposed idea has not been fully investigated yet. Hence, simulation study and real data analysis are required for future work.

Chapter 5

Concluding remarks

In this dissertation, we have proposed distinct strategies for addressing high-dimensional survival model selection problems. The proposed methods provide various ways for determining the optimum model archiving one of the properties: (1) the global optimal model that optimizes the model selection criterion, (2) the best model that guarantees the best prediction performance, (3) the global optimal model when a model is misspecified. The contributions, extensions, as well as limitations of our proposed methods can be summarized in the following aspects.

5.1 Contributions

We have developed innovative strategies for high-dimensional survival model selection. The proposed methods are motivated by statistical mechanics, which comes from the field of physics. Our key idea is to incorporate the notion of Bayesian inference into a model selection framework via the Boltzmann distribution. The Boltzmann distribution is used to define the probability distribution associated with the model selection criterion in the simulated annealing scheme that performs global optimization via stochastic search algorithms. The proposed framework serves as a bridge between Bayesian model selection and frequentist model selection by leveraging the idea of statistical mechanics. When compared to other

popular methods, the proposed methods provide outstanding performance. The simulation study and real data analysis have shown that the improvement is significant.

5.2 Extensions

While we have restricted our attention to the Cox proportional hazards model and the accelerate failure time model, the proposed methods perform well under various survival data models. The proposed ideas can also be adapted to a wide range of models, including non-parametric models, semi-parametric models, and parametric models with various types of data. As a result, a broader range of statistical applications can be encompassed within our proposed framework. For example, our proposed idea can be implemented to solve the problem of high-dimensional data model selection for linear regression or generalized linear regression.

5.3 Limitations

As shown in the simulation studies, our proposed methods perform well for selecting the best model for model fitting and prediction. Although model selection has been our primary focus of this dissertation, in practice, computational efficiency is also an important issue to be taken into consideration. The simulation result shows that our proposed method requires more time for implementation than the existing methods. This high computation cost can be considered as the major drawback of our proposed method. As we are dealing with high-dimensional survival data, it is important to improve computational efficiency. Various ways of adjustments could help us to speed up the computation. For example, we can speed up the R implementation by modifying our R codes with high-performance computation packages (e.g. Rcpp). At this moment, we will leave the improvement of computation efficiency for future work.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature 403*(6769), 503–511.

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

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika 67*(2), 413–418.

Barbieri, M. M., J. O. Berger, et al. (2004). Optimal predictive model selection. *The annals of statistics 32*(3), 870–897.

Berger, J. (1997). Bayes factors. in the encyclopedia of statistical sciences. *Update 3*, 20–29.

Boltzmann, L. (1868). Studien uber das gleichgewicht der lebenden kraft. *Wissenschafiliche Abhandlungen 1*, 49–96.

Casella, G. and E. I. George (1992). Explaining the gibbs sampler. *The American Statistician 46*(3), 167–174.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chen, J. and Z. Chen (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 555–574.

Clyde, M. (1999). Bayesian model averaging and model search strategies. *Bayesian statistics 6*, 157–185.

Clyde, M. and E. I. George (2000). Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 681–698.

Clyde, M. A. and E. I. George (1999). Empirical bayes estimation in wavelet nonparametric regression. In *Bayesian inference in wavelet-based models*, pp. 309–322. Springer.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

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

Euler, L. (1740). De progressionibus harmonicis observationes. *Commentarii academiae scientiarum Petropolitanae*, 150–161.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Foygel, R. and M. Drton (2010). Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems 23*, 604–612.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon. 6*, 93–116.

Gail, M., W.-Y. Tan, and S. Piantadosi (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika 75*(1), 57–64.

Gail, M. H., S. Wieand, and S. Piantadosi (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika 71*(3), 431–444.

Gibbs, J. W. (1902). *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. C. Scribner's sons.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical journal 52*(1), 70–84.

Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *5*(2), 115–158.

Hattori, S. (2012). Testing the no-treatment effect based on a possibly misspecified accelerated failure time model. *Statistics & Probability Letters 82*(2), 371–377.

Henderson, R., M. Jones, and J. Stare (2001). Accuracy of point predictions in survival analysis. *Statistics in medicine 20*(20), 3083–3096.

Ho, L. and A. Silva (2006). Unbiased estimators for mean time to failure and percentiles in a weibull regression model. *International Journal of Quality & Reliability Management*.

Huber, P. J. (1967). Under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*, Volume 5, pp. 221. Univ of California Press.

Ishii, R., K. Maruo, H. Noma, and M. Gosho (2021). Statistical inference based on accelerated failure time models under model misspecification and small samples. *Statistics in Biopharmaceutical Research 13*(4), 384–394.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

Kalbfleisch, J. D. and R. L. Prentice (2011). *The statistical analysis of failure time data*, Volume 360. John Wiley & Sons.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika 69*(1), 19–27.

Kim, Y., J.-J. Jeon, et al. (2016). Consistent model selection criteria for quadratically supported risks. *The Annals of Statistics 44*(6), 2467–2496.

Kim, Y., S. Kwon, and H. Choi (2012). Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research 13*, 1037–1057.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by simulated annealing. *science 220*(4598), 671–680.

Lagakos, S., , and D. Schoenfeld (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, 1037–1048.

Lagakos, S. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika 75*(1), 156–160.

Lin, D. Y. and L.-J. Wei (1989). The robust inference for the cox proportional hazards model. *Journal of the American statistical Association 84*(408), 1074–1078.

Liu, E. (2018). Using weibull accelerated failure time regression model to predict survival time and life expectancy. *BioRxiv*, 362186.

Luo, S., J. Xu, and Z. Chen (2015). Extended bayesian information criterion in the cox model with a high-dimensional feature space. *Annals of the Institute of Statistical Mathematics 67*(2), 287–311.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Morgan, T. M., S. Lagakos, and D. Schoenfeld (1986). Omitting covariates from the proportional hazards model. *Biometrics*, 993–995.

Newton, I. (1736). *The Method of Fluxions and Infinite Series: With Its Application to the Geometry of Curve Lines*. Nourse.

O'neill, T. J. (1986). Inconsistency of the misspecified proportional hazards model. *Statistics & probability letters 4*(5), 219–222.

Rosenwald, A., G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, et al. (2002). The

use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine 346*(25), 1937–1947.

Rosin, P. (1933). Laws governing the fineness of powdered coal. *Journal of Institute of Fuel 7*, 29–36.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review/Revue Internationale de Statistique*, 221–226.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Solomon, P. J. (1984). Effect of misspecification of regression models in the analysis of survival data. *Biometrika 71*(2), 291–298.

Struthers, C. A. and J. D. Kalbfleisch (1986). Misspecified proportional hazard models. *Biometrika 73*(2), 363–369.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine 16*(4), 385–395.

Volinsky, C. T. and A. E. Raftery (2000). Bayesian information criterion for censored survival models. *Biometrics 56*(1), 256–262.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine 11*(14-15), 1871–1879.

Weibull, W. et al. (1951). A statistical distribution function of wide applicability. *Journal of applied mechanics 18*(3), 293–297.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1–25.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*(2), 894–942.

Zhang, H. H. and W. Lu (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika 94*(3), 691–703.

Zhang, Y., R. Li, and C.-L. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*(489), 312–323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Appendix A

R code

```
### y: survival time
### status: censoring indicator
### x: covariates with dimension n by p
### tau: temperature
### k: upper bound controls the # of covariates in the model
n0 <- sum(status)
BIC<-function(s)
{
  if(sum(s) == 0)){
    fit0    <- coxph(Surv(y, status!=0) ~  1);
    return(-2*fit0$loglik)
  }else{
    fit <- coxph(Surv(time, status) ~ x[,s]);
    return(-2*fit$loglik[2]+log(n0)* length(fit$coef))
  }
}
s <- rep(0,ncol(x))
```

```
best.s <- s
best.BIC <- BIC(s)
r <- 1
while(r < m*p)
  {
     s.prev <- s
     for (j in 1:ncol(x))
     {
        if( length(which(s==1)) < k ) {
           s.1    <- s
           s.1[j] <- 1
           s.0    <- s
           s.0[j] <- 0
           BIC1   <- BIC(which(s.1!=0))
           BIC0   <- BIC(which(s.0!=0))
           BIC.min<- min(BIC1,BIC0)
           s.min  <- ifelse(BIC1 > BIC0, s.0, s.1)
           if( best.BIC > BIC.min ){
               r <- 1
               best.s   <- s.min
               best.BIC <- BIC.min
           }else{
               r <- r+1
           }
           w      <- exp(-BIC1/tau+BIC0/tau)
           prob   <- w/(1+w)
           s[j]   <- rbinom(1, 1, prob=prob)
```

```
45
46          }else { next }
47      }
```