2019-01-25

Bi-level Variable Selection in Semiparametric Transformation Models for Right Censored Data and Cure Rate Data

Zhong, Wenyan

UNIVERSITY OF CALGARY

Bi-Level Variable Selection in Semiparametric Transformation Models for Right Censored Data and Cure Rate Data

by

Wenyan Zhong

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN STATISTICS

CALGARY, ALBERTA

JANUARY, 2019

Abstract

In this dissertation, I investigated the bi-level variable selection in the semiparametric transformation models with right-censored data and the semiparametric mixture cure models with right-censored and cure rate data, respectively. The transformation models under the consideration include the proportional hazards model and the proportional odds model as special cases. In the framework of regularized regression, we proposed a computationally efficient estimation method that selects significant groups and variables simultaneously. Three penalty functions, i.e., Group bridge, adaptive group bridge and composite group bridge penalties which can integrate grouping structure of covariates, were adopted for bi-level variable selection purpose. In Chapter 2, the objective function, which consists of the negative weighted partial log-likelihood function plus one of the three penalties, has a parametric form and is convex with respect to the parameters. This leads to an easy implementation of the optimization algorithm for which convergence is guaranteed numerically. We showed that all the three proposed penalized estimators achieve the group selection consistency, and moreover, the adaptive group bridge estimator and the composite group bridge estimator enjoy the oracle properties, i.e., both estimators possess the group and individual selection consistency simultaneously and are asymptotically normal as if the true unimportant covariates were known. In Chapter 3, we further extended the bi-level variable selection procedure to the semiparametric mixture cure models. The semiparametric mixture cure models are formulated by a logistic regression for modelling the cure fraction and a class of semiparametric transformation models for modelling the survival function of remaining uncured individuals. Incorporating a cure fraction, the proposed model is more flexible than the standard survival models, and the proposed approach is capable to distinguish important covariates and groups from unimportant ones and estimate covariates' effects simultaneously in both the incidence and the latency parts. We proposed a new iterative E-M algorithm to handle two latent variables. We illustrated the finite sample performance of the proposed methods via simulations and two real data examples. Simula-

tion studies indicated that the proposed methods perform well even with relatively high dimension of covariates.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors, Dr. Jingjing Wu and Dr. Xuewen Lu for their constant guidance, inspirational encouragement and insightful comments on dissertation research and in my Ph.D. study. I really appreciate the opportunity that they led me to grow in this engaging statistical research area. The completion of this dissertation would not have been possible without their supervision and continuous support. Their enlightening guidance cultivated me with open-mindedness, the ability of critical thinking and attention to details that is essential for my professional development. The spirit of persistence, patience and creativity they conveyed will be beneficial to the rest of my life.

Besides my supervisors, I wish to express my profound appreciation to my supervisory committee members, Drs. Gemai Chen and Alexander de Leon, and my examination committee members, Drs. Hua Shen and Linglong Kong, for their precious time and helpful comments to improve this dissertation.

I am sincerely grateful to the professors in the Department of Mathematics and Statistics for all the support, encouragement and illuminative lectures I have received from them through these years. Further, I would like to extend my sincere gratitude to my fellow doctoral students and all my friends who have helped and supported me along the way.

Last but not the least, I am highly grateful to my parents for supporting me spiritually throughout my Ph.D. journey and my life in general.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
r.v.	random variable
c.d.f.	cumulative distribution function
p.d.f.	probability distribution function
PH	proportional hazards
PO	proportional odds
AFT	accelerated failure time
MLE	maximum likelihood estimation/estimator/estimate
NPMLE	nonparametric maximum likelihood estimation/estimator/estimate
LSE	least-squares estimator
EM	expectation-maximization
LASSO	least absolute shrinkage and selection operator
SCAD	smoothly clipped absolute deviation
MCP	minimax concave penalty
$N_p(\mu, \Sigma)$	p-dimensional multivariate normal distribution with mean μ and covariance matrix Σ
$\|\cdot\|_\gamma$	L_γ-norm
$\xrightarrow{a.s.}$	converge almost surely

Chapter 1

INTRODUCTION

Resulted from the speedy development in modern evolving technology, an increasing number of potential predictors could be easily gathered with low cost to avoid missing critical factors for statistical modelling. This promotes the fact that high dimensional data are becoming increasingly prevalent in various fields including finance, clinical trials, neuroimaging, genomics and many others. To improve model interpretability, stability and prediction accuracy, selecting a subset of significant variables plays a pivotal role in building statistical models. It is an even more demanding and challenging task for modelling censored data when we are provided a large set of covariates. The existing methods tend to place strict assumptions on either covariate effects or censoring mechanism, which end up with either failing to predict future outcomes accurately or making them highly implausible to implement and solve real world problems. In this dissertation, we will consider survival models which allow more complex covariate patterns than the commonly used Cox proportional hazards (PH) model and the proportional odds (PO) model. To incorporate the possibly known grouping information of covariates, we generalize three efficient bi-level variable selection approaches to a class of semiparametric transformation models. We further extend the idea to the more general semiparametric mixture cure models.

In this chapter, we conduct literature review and list the areas and methods to which this dissertation is related. More specifically, we review survival analysis in Section 1.1, semiparametric transformation models in Section 1.2, mixture cure models in Section 1.3 and variable selection methods in Section 1.4. Some technical details of the methods mentioned in Section 1.4 will be described in subsequent chapters along with our proposed methods.

1.1 Survival Analysis

Survival analysis corresponds to a set of statistical methods for studying data where the outcome variable is the time that takes for the occurrence of an event of interest. The event can be marriage, bankruptcy, death, the occurrence of a disease, a machine part fails, among others. In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. The response time to event or survival time is always positive, usually continuous, and can be measured in days, weeks, years, etc. The presence of censoring is another major feature in survival analysis, which represents a particular type of missing data that distinguishes the statistical analysis of survival data from analyzing other types of outcomes. Subjects are said to be censored when the information about their survival time is incompletely observed, i.e., the exact survival time is unknown. Censoring that is random and non-informative is usually required in order to avoid bias in a survival analysis. The non-informative assumption means the censoring is non-informative about the event of interest, that is, the censoring is caused by something other than the impending failure.

Different censoring mechanisms lead to different types of censored data. In this dissertation, we restrict ourselves to right-censored data as it is the most common type of censoring in survival analysis. For right-censored data, let T be a non-negative r.v. representing the waiting time until the occurrence of an event and C be potential censoring time, then the observed data consists of $\{\tilde{T}, \delta\}$, where $\tilde{T} = \min\{T, C\}$ is the observed time and $\delta = I(T \leqslant C)$ is the censoring indicator and $I(\cdot)$ is the indicator function. Let $Z(t)$ represent the associated d-dimensional vector of covariates which may depend on the time t. One of the important objectives in survival analysis is to assess the relationship between risk factors $Z(t)$ and survival time T and make future prediction of subject's survival probabilities. The probability of surviving past a certain point in time may be of more interest than the expected time of event. The survival function is defined as

$$S(t) \;=\; Pr(T > t) \;=\; 1 - F(t),$$

where F is the cumulative distribution function (c.d.f.) of T. The survival function gives the

probability that a subject will survive past time t. The survival function $S(t)$ is non-increasing and takes value 1 when $t = 0$ and value 0 when $t \to \infty$. In theory, the survival function is smooth, however, we usually observe events on a discrete time scale in practice. Another useful function in the context of survival analysis is the hazard function $h(t)$ which describes the probability of an event or its hazard if the subject survived up to that particular time point t. It is the instantaneous rate at which events occur given no previous events, formulated as

$$h(t) \;=\; \lim_{\Delta t \to 0^+} \frac{Pr(t < T \leqslant t + \Delta t | T > t)}{\Delta t} \;=\; \frac{f(t)}{S(t)}.$$

The cumulative hazard describes the accumulated risk up to time t, $H(t) = \int_0^t h(u)du$, it is a non-decreasing function. If we know any one of the functions $S(\cdot)$, $H(\cdot)$, or $h(\cdot)$, we can derive the other two functions by

$$h(t) \;=\; \frac{f(t)}{S(t)} \;=\; -\frac{\partial \log S(t)}{\partial t},$$

$$H(t) \;=\; -\log S(t),$$

$$S(t) \;=\; \exp[-H(t)].$$

The literature on survival analysis provides extensive approaches to modelling and estimating S or H, ranging from fully nonparametric to completely parametric. Kaplan-Meier estimator is a frequently used nonparametric statistic that allows us to estimate the survival function without a parametric assumption on underlying probability distribution. This estimator was proposed by Kaplan and Meier (1958) as

$$\hat{S}(t) \;=\; \prod_{t_i \leqslant t} \left[1 - \frac{E_i}{Y_i}\right],$$

where Y_i is the number of subjects who are at risk at time t_i and E_i is the number of subjects who experience the event of interest at time t_i. Thus, the conditional probability that a subject who survives just prior to time t_i experiences the event of interest at time t_i is equal to E_i/Y_i. Note that $\hat{S}(t) = 1$ when t is less than the smallest observed survival time $\min\{t_i, i = 1, \ldots, n\}$, and $\hat{S}(t)$ is not well defined when t is greater than the largest observed survival time $\max\{t_i, i = 1, \ldots, n\}$.

By imposing parametric assumptions, we can also estimate survival curves via parametric survival functions such as, to name a few, Exponential, Weibull, Gamma, log-normal and log-logistic. Exponential model employs the exponential distribution

$$f(t) \;=\; \lambda \exp(-\lambda t), \quad \lambda > 0, t \geqslant 0$$

as the probability density function (p.d.f.) of survival time. Based on this exponential assumption, one obtains the corresponding survival function and hazard function being respectively

$$S(t) \;=\; \exp(-\lambda t),$$

$$h(t) \;=\; \lambda.$$

We can see that its hazard function is a constant which makes this model inapplicable for many real datasets.

Weibull distribution has being used widely in survival analysis, which includes exponential distribution as a special case. For Weibull survival model, the related functions are given by

$$f(t) \;=\; \alpha\lambda t^{\alpha-1} \exp(-\lambda t^{\alpha}), \quad \alpha, \lambda > 0, t \geqslant 0,$$

$$S(t) \;=\; \exp(-\lambda t^{\alpha}),$$

$$h(t) \;=\; \alpha\lambda t^{\alpha-1},$$

where α is the shape parameter and λ is the scale parameter. This model has more flexibility of fitting different types of real survival data, as its hazard function can take any form of increasing, decreasing or constant with different values of α and λ.

Log-normal is another model that has been widely employed in survival analysis due to its link to normal distribution. A non-negative r.v. T is said to follow the log-normal distribution if its logarithm $Y = \log T$ follows a normal distribution. For log-normal survival model, the related functions are given by

$$f(t) \;=\; \frac{\exp\left[-1/2\left\{\frac{\log(t)-\mu}{\sigma}\right\}^2\right]}{t(2\pi)^{1/2}\sigma}, \quad \sigma > 0, t \geqslant 0,$$

$$S(t) \;=\; 1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\},$$
$$h(t) \;=\; f(t)/S(t),$$

where μ and σ are the location parameter and scale parameter of Y respectively and $\Phi(\cdot)$ denotes the c.d.f. of a standard normal r.v..

Accelerated failure time (AFT) model describes a linear relationship between a vector of explanatory covariates Z and the natural logarithm of the survival time T. The logarithm transformation is used in order to convert positive survival times to observations on the entire real line. The AFT model takes the form

$$\log(T) \;=\; -\beta^{\top}Z+\varepsilon,$$

where β is the vector of regression coefficients and ε is the random error term with a completely unspecified distribution function. Each error distribution yields a specific model for survival time T. For example, if the error distribution is the standard normal, then the survival time has log-normal regression model. If the error has the logistic distribution, then the survival time has the log-logistic model. If the error has the extreme value distribution, then the survival time follows the Weibull regression model. With an unspecified error term in the model, the AFT model could be categorized into the class of semiparametric survival models. Semiparametric models combine a parametric model for some components and a nonparametric model for the remaining. It bypasses some restrictions of using fully parametric model and promotes model flexibility. Hence, semiparametric survival models is the domain of this dissertation.

1.2 Semiparametric Transformation Models

Since the development of the Cox PH model by Cox (1972), a wide variety of semiparametric survival models have been proposed to model censored data in survival analysis. The Cox PH model does not impose assumptions on the underlying p.d.f. but it assumes that the hazards of subgroups are constant over time, hence the name "proportional hazards". The conditional hazard

function for a subject with d-dimensional covariate vector X (not include constant term) is assumed to be

$$h(t|X) \;=\; h_0(t)\exp\left(\beta^\top X\right),$$

where β is the unknown vector of regression coefficients, $h_0(t)$ is the unknown baseline hazard function that describes the risk of subjects with $X = 0$ and serves as a reference level or pivot, and $\exp(\beta^\top X)$ is the relative risk, a proportionate increase or decrease in risk associated with the set of characteristics X. Estimates of regression coefficients could be obtained without a parametric assumption on the baseline hazard function, however this model assumes that the covariates have multiplicative effects on the hazard function. This PH assumption is often violated in real applications. For example, sometimes it is reasonable to assume that the subject-specific hazard function becomes more similar to the baseline hazard when time lasts. Thus the PO regression model (Bennett, 1983) may be used in this case because it assumes that the odds of survival are proportional. Given the vector of covariates X, the survival function $S(\cdot)$ is modelled as

$$\left[\frac{S(t)}{1-S(t)}\right] \;=\; \left[\frac{S_0(t)}{1-S_0(t)}\right]\exp\left(\beta^\top X\right),$$

where β is the unknown vector of regression coefficients and $S_0(t)$ is the unknown baseline survival function. Similar to the Cox PH model, estimation of the regression coefficients could be obtained without a parametric assumption on the baseline log-odds function. The PO model assumes that the covariates have multiplicative effect on the odds of survival beyond time t, implying that the ratio of hazards corresponding to different covariates converges to unity as time lasts.

The Cox PH and the PO models are special cases of transformation survival models, a class of semiparametric linear transformation models that relates an unknown transformation of survival time linearly to covariates (Kalbfleisch and Prentice, 2011). Under linear transformation models, estimations of regression coefficients may still be obtained without a parametric assumption on baseline hazard function, as in the Cox PH and the PO models. However, since the PH and the PO assumptions are relaxed in the transformation models, the coefficients no longer have intuitive

interpretations except for these two special cases. The flexibility of transformation models makes the shape of subject-specific hazard function vary according to the transformation chosen. Let T be the failure time and Z the associated d-dimensional vector of fixed and time-invariant covariates. Then the semiparametric linear transformation model can be written as

$$H(T) \;=\; -\beta^\top Z + \varepsilon, \tag{1.1}$$

where H is the unknown monotone transformation function, ε is a r.v. with known distribution independent of Z, and β is the unknown d-dimensional vector of regression coefficients. Model (1.1) reduces to the Cox PH model when ε follows the extreme value distribution, and reduces to the PO model when ε follows the standard logistic distribution. With use of counting process, Zeng and Lin (2006, 2007b) proposed a class of more general transformation models to accommodate time-varying covariates and recurrent events. Let $N(t)$ denote the counting process recording the number of events that have occurred by time t and $Z(t)$ the possibly time-varying covariates. The cumulative hazards function of $N(t)$, conditional on $Z(t)$, takes the form

$$\Lambda(t|Z(\cdot)) \;=\; G\left[\int_0^t \exp\left\{\beta^\top Z(s)\right\} \mathrm{d}\Lambda(s)\right], \tag{1.2}$$

where G is a thrice continuously differentiable and strictly increasing function with $G(0) = 0$, $G'(0) > 0$ and $G(\infty) = \infty$. When $N(\cdot)$ has a single jump at each survival time t and the covariate Z is time-invariant, model (1.2) is reduced to

$$\Lambda(t|Z) \;=\; G\left[\exp\left(\beta^\top Z\right)\Lambda_0(t)\right],$$

which is equivalent to

$$\log\Lambda_0(t) = -\beta^\top Z + \log G^{-1}(-\log\varepsilon_0),$$

where ε_0 has a uniform distribution. Therefore, model (1.2) reduces to the linear transformation model (1.1) with $H(t) = \log\Lambda_0(t)$ and the error term $\varepsilon = \log G^{-1}(-\log\varepsilon_0)$. Specifying the function G while leaving the function Λ_0 unspecified is equivalent to specifying the distribution of ε while leaving the function H unspecified.

Zeng and Lin (2007b) considered two classes of functions for G in (1.2). One is Box-Cox transformation

$$G(x) = \begin{cases} \dfrac{(1+x)^{\rho}-1}{\rho}, & \rho > 0, \\ \log(1+x), & \rho = 0. \end{cases} \tag{1.3}$$

The other is logarithm transformation

$$G(x) = \begin{cases} \dfrac{\log(1+rx)}{r}, & r > 0, \\ x, & r = 0. \end{cases} \tag{1.4}$$

When $G(x) = x$, i.e., $\rho = 1$ in (1.3) or $r = 0$ in (1.4), it leads to the Cox PH model. When $G(x) = \log(1+x)$, i.e., $\rho = 0$ in (1.3) or $r = 1$ in (1.4), it leads to the PO model.

Estimations of regression coefficients can be obtained using nonparametric maximum likelihood approach, in which the unknown survival distribution is treated as an infinite-dimensional parameter. The likelihood is maximized over β and cumulative hazard function's jump sizes Λ_k at each observed failure time. Zeng and Lin (2007b) showed that the nonparametric maximum likelihood estimators (NPMLEs) of β and Λ_k's are consistent, asymptotically normally distributed and asymptotically efficient. The maximization of likelihood is computationally intensive, especially when the number of observed evens is large. A number of algorithms are available for computing the NPMLEs of β and Λ_k's, and one of them is the expectation-maximization (EM) algorithm.

1.3 Mixture Cure Models

In survival analysis, there is a commonly unstated assumption that all subjects will eventually experience the event of interest when the follow-up time is long enough. However, the development and advancement of medical technologies and pharmacology have made it possible for some patients to be cured of certain type of diseases through treatment. In such situations, a portion of the subjects are risk-free of the event of interest, even after prolonged periods of observation, and those subjects are referred as long-term survivors. Classical survival models, such as the Cox PH, PO and AFT models, do not directly account for the cured portion of the sample, because the suscep-

tibility of the event of interest is assumed for all subjects. Since the non-susceptible subjects will never experience the event, their survival times are always right-censored. The cure models are a class of survival models for modelling such situations. A numerous applications of cure models can be found in medical and epidemiological studies in the last two decades. As pointed out by Farewell (1986), cure models should not be used indiscriminately. To validate the implementation of cure models, there must be obvious empirical and biological evidence of a non-susceptible subpopulation. Cure models can be categorized into three classes, namely, mixture cure models, promotion time cure models and unifying models.

The mixture cure models assume that the underlying population is a mixture of susceptible and non-susceptible subgroups. The mixture cure models have two separately modelled parts, one for the probability of cure, referred as incidence part, and the other the depiction of the survival distribution for uncured group, referred as latency part. In this formulation, the association between outcome and covariates is captured by two groups of parameters: (1) the incidence rate of the event among all subjects, and (2) the survival distribution for those that are susceptible to the event. With right-censored data, various mixture cure models have been investigated by many authors, including Farewell (1977, 1982), Kuk and Chen (1992), Ross and Xian (1996), Sy and Taylor (2000), Peng and Dear (2000), Lu and Ying (2004), Zhang and Peng (2009), among others. Farewell (1982) proposed a logistic regression for the cured proportion and a Weibull regression model for the survival function of the uncured group. Under this mixture formulation, researchers have extended the cure model for various data settings. As examples of notable extensions, Kuk and Chen (1992) proposed a semiparametric generalization of Farewell's model by using a Cox PH model for the uncured group with a Monte Carlo simulation based estimation method; Sy and Taylor (2000) and Peng and Dear (2000) developed maximum likelihood methods for the joint estimation of parameters in both parts by employing the nonparametric form of the likelihood and an EM algorithm; Chen and Wang (2000) and Zhang and Peng (2009) used accelerated hazards model for the uncured group.

The promotion time cure model, an alternative method for modelling survival data with a cure fraction, was introduced and investigated by Yakovlev and Tsodikov (1996) and Tsodikov (1998); Tsodikov et al. (2003). Chen et al. (1999), Ibrahim et al. (2001), and Yin and Ibrahim (2005) extended this model to a Bayesian framework. This model initially stems from kinetic studies of carcinogenic cells, where the number of carcinogenic cells was considered as the underlying cause of cancer recurrence. The model used in these studies was based on the underlying biological mechanism and the assumption that the number of metastasis-competent tumour cells after a certain treatment follows a Poisson distribution. One appeal of this model is that the survival distribution for either cured subjects or noncured subjects can be integrated into one single formulation with the cured subjects being assumed to have survival time of infinity. For the i-th individual with covariate X_i, the survival function of this subject is given by $S(t|X_i) = \exp[-\theta(X_i)F(t)]$, where F is a proper baseline distribution function and θ is a known link function that delineates the effect of the covariate X on the survival function S. Under the promotion time cure model, the cure rate is $\lim_{t\to\infty} S(t|X_i) = \exp(-\theta(X_i))$. Moreover, when $\theta(X_i) = \exp(\beta^\top X_i)$ and β includes an intercept term β_0, the model becomes the usual Cox PH model subject to a bounded cumulative baseline hazard function $H(t) = F(t)\exp(\beta_0)$. This bounded cumulative hazard function leads to an improper survival function, i.e., $\lim_{t\to\infty} S(t) > 0$. Tsodikov (1998) developed an estimation method based on the profile likelihood approach. Most of the current studies on the promotion time cure models are based on Bayesian inference methods.

Some earlier research underlined the connection and differences between the mixture cure model and the promotion time cure model, which encouraged the formulation of the third type of cure models, the unifying model. As an alternative to the aforementioned two types of cure models, the unifying model embeds both the mixture cure model and the promotion time cure model. It is not as commonly used as the other two kinds of cure models. Works for the unifying model includes Zeng et al. (2006), Taylor and Liu (2007) and Gu et al. (2011), among others.

Compared to traditional survival models, the cure models are less interpretable due to the com-

plex model structure, especially when the number of explanatory variables is large. To alleviate this issue, variable selection technique could be applied to reduce model complexity. In this dissertation, we propose a model selection procedure under the mixture cure models which identifies significant groups and variables in order to construct a parsimonious mixture cure model for prediction.

1.4 Variable Selection Methods

In survival analysis, a key problem is to assess the covariate effect on survival outcome. With the advanced modern technology, an increasing number of variables can be easily collected with low cost and accessible for researchers to predict survival outcomes. Under high-dimensional data setting, many traditional variable selection approaches, such as stepwise procedures and best subset selection, are no longer feasible and computationally expensive. In the last few decades, processes of selecting explanatory variables have been extensively studied under linear models with non-censored data. There is abundant literature exploring techniques and computational algorithms for variable selection that emerged in the last few decades. A large number of variable selection methods have been developed under the framework of penalized likelihood including the criterion-based procedures. In an attempt to select significant variables and estimate regression coefficients automatically and simultaneously, a family of variable selection approaches through regularization were proposed. Let X denote the vector of covariates and y be the response. Generally, the objective function of regularized estimation is given by

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \mathrm{P}_{\lambda}(\beta),$$

where $\mathrm{L}(\cdot)$ is a given empirical loss function (e.g., least-squares, negative log-likelihood-based function); $\mathrm{P}_{\lambda}(\cdot)$ is a specified non-negative penalty function that penalizes large coefficients of parameters with the intercept term being left out; λ is a regularization/tuning parameter that controls the trade-off between the fit and the parsimony of a model. An estimator of β can be obtained by minimizing the above function for a given value of λ, yielding $\hat{\beta}_{\lambda} = \arg\min_{\beta} \mathrm{Q}(\beta|\mathrm{X},\mathrm{y})$.

Various choices of the function $P_\lambda(\cdot)$ can be found in the literature, such as ridge penalty (Hoerl and Kennard, 1970), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), bridge penalty (Frank and Friedman, 1993; Fu, 1998), smoothly clipped absolute deviation (SCAD) method (Fan and Li, 2001), elastic net method (Zou and Hastie, 2005) and minimax concave penalty (MCP) method (Zhang et al., 2011) for individual variable selection; group LASSO (Bakin et al., 1999; Yuan and Lin, 2006), spare group LASSO (Friedman et al., 2010) and group MCP (Breheny and Huang, 2009) for group selection; group bridge (Huang et al., 2009), adaptive group bridge (Seetharaman, 2013), composite group bridge (Seetharaman, 2013) and composite MCP (Breheny and Huang, 2009) for bi-level variable selection.

1.4.1 Individual variable selection methods

Ridge regression estimator was proposed by Hoerl and Kennard (1970) for linear models to deal with multicollinearity in predictors, while Li and Luan (2002) applied the ridge penalty to the Cox PH model for high-dimensional censored data. The objective function of ridge regression is defined as

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda\|\beta\|_2^2.$$

Throughout this thesis, we use $\|\cdot\|_\gamma$ to denote the L_γ-norm of a vector, i.e., $\|x\|_\gamma = (|x_1|^\gamma + |x_2|^\gamma + \cdots + |x_p|^\gamma)^{1/\gamma}$, where γ is a positive real number. Minimizing the objective function gives a closed-form solution $\hat{\beta}^{ridge} = \left(n^{-1}\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I}\right)^{-1} n^{-1}\mathrm{X}^T\mathrm{y}$, which is similar to the least-squares estimator (LSE) but with an additional "ridge" down the diagonal of the matrix to be inverted. The L_2-norm penalty on the regression coefficients yields a unique and more stable estimator of β than un-regularized estimators. However, ridge estimator does not give a way of selecting important variables and cannot provide accurate predictions in the presence of scale difference among regression coefficients.

Tibshirani (1996) proposed the prominent LASSO approach for estimation and variable selection simultaneously under the linear model. The objective function of LASSO regression is given

by

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda\|\beta\|_1.$$

Like ridge penalty, LASSO penalizes large regression coefficients and shrinks estimates towards zero. Nevertheless, unlike the ridge penalty, due to the singularity of $|\beta_i|$ at $\beta_i = 0$, LASSO produces solution with some coefficient estimates being exactly zeros which effectively removes those predictors from the model to achieve model sparsity. Tibshirani et al. (1997), Gui and Li (2005), and Park et al. (2006) extended the LASSO approach to the Cox PH model using different optimization algorithms.

Frank and Friedman (1993) considered a class of regression estimators using L_γ penalty with $0 \leqslant \gamma \leqslant 2$, a natural generalization of L_0 penalty. The objective function is given by

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) = \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda \sum_{k=1}^{d} |\beta_k|^\gamma,$$

where $\lambda > 0$ and $\gamma > 0$. It bridges the best subset penalty (L_0 penalty) and the ridge penalty (L_2 penalty), and hence the name "bridge" regression. When $\gamma = 0$, the penalty term corresponds to L_0 penalty that equals the number of nonzero parameters in a model and has been used in *AIC* and *BIC* model selection criteria; when $0 < \gamma < 1$, it corresponds to a class of concave penalties; when $1 \leqslant \gamma \leqslant 2$ it corresponds to a class of convex penalties, and especially the lasso penalty when $\gamma = 1$ and the ridge penalty when $\gamma = 2$. Under appropriate regularity conditions with $0 < \gamma < 1$, bridge estimators enjoy selection consistency and are capable to distinguish in sparse high-dimensional settings between covariates with coefficients exactly zero and those with nonzero coefficients.

Fan and Li (2001) developed the SCAD penalty defined as

$$\mathrm{P}_{\lambda,\gamma}(\beta) = \begin{cases} \lambda|\beta|, & \text{if } |\beta| \leqslant \lambda, \\ -\frac{|\beta|^2 - 2\gamma\lambda|\beta| + \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < |\beta| \leqslant \gamma\lambda, \\ \frac{(\gamma+1)\lambda^2}{2}, & \text{if } |\beta| > \gamma\lambda, \end{cases}$$

where $\gamma > 2$ are tuning parameters. Fan and Li (2001) suggested using $\gamma \approx 3.7$ in SCAD. The SCAD penalty function is singular at the origin and has continuous first-order derivative. The

SCAD penalty is a non-concave function with a constant penalty on large coefficients. It is a smooth transition from L_1 penalty to L_0 penalty.

To assess whether a penalty is "good" or not, Fan and Li (2001) advocated the following three criteria for the correspondingly penalized estimator, known as *oracle property*:

(1) Sparsity: the resulting estimator sets small estimated coefficients to exact zeros so that variable selection and model complexity reduction are achieved;

(2) Unbiasedness: the resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, in order to reduce model bias;

(3) Continuity: the resulting estimator is continuous in data in order to avoid instability in model prediction and be robust against small model perturbation.

An oracle estimator is defined as the one that enjoys these three properties. The name "oracle" stands for the feature that an estimator correctly identifies the true model and efficiently estimates the nonzero components as if the true model were known. Ideally, small penalty should be applied to large coefficients for small bias, while large penalty applied to small coefficients for better sparsity. Since LASSO imposes the same level of penalization on each coefficient, LASSO estimators do not enjoy the oracle property. Most of the penalties producing oracle estimators are concave, suffering from multiple local minima numerically, such as adaptive LASSO and bridge. Non-concave penalties SCAD and MCP, allow the estimated coefficients to converge numerically to large values more quickly than LASSO. Although these methods all shrink small coefficients towards zeros, MCP and SCAD have comparatively less shrinkage effect on nonzero coefficients.

1.4.2 Group selection methods

The aforementioned methods focus on individual variable selection. However, grouping structures arise naturally in many statistical modelling problems. For example, in gene expression analysis, genes belonging to the same biological pathway can be considered as a group. In genetic association studies, genetic markers from the same gene can be considered as a group. Therefore, it is

more plausible to take the grouping structure into account in the analysis of such data in order to increase interpretability. Besides, grouping structure describes the inherent interconnections and the ignorance of it may lead to an inefficient model.

Consider the general regression problem with n observations and d single variables consisting of J groups. Then the linear regression model could be written as

$$\mathrm{Y} \;=\; \sum_{j=1}^{J} \mathrm{X}_{A_j} \beta_{A_j} + \varepsilon,$$

where for $j = 1, \ldots, J$, A_j is a subset of $\{1, \ldots, d\}$ with size d_j, Y is a $n \times 1$ column vector of response variable, X_{A_j} is a $n \times d_j$ design matrix corresponding to the j-th group, β_{A_j} is a $d_j \times 1$ coefficient vector corresponding to the j-th group, and $\varepsilon \sim N_n(0, \sigma^2 I)$ is a $n \times 1$ random column vector of error term. Yuan and Lin (2006) proposed group LASSO penalty under this linear regression model assuming the J design matrices are orthonormal. The objective function of group LASSO approach is given by

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_{R_j},$$

where λ_n is a non-negative penalty parameter, c_j is a constant used to adjust for group size d_j, R_j is a $d_j \times d_j$ positive definite weight matrix, and $\|\beta_{A_j}\|_{R_j} = (\beta_{A_j}^{\top} R_j \beta_{A_j})^{1/2}$. Group LASSO uses weighted L_2-norm of the grouped coefficients in the penalty function. Note that when $d_j = 1$, $c_j = 1$ and R_j is the identity matrix, the group LASSO penalty is reduced to the regular LASSO penalty. When $R_j = I_j$, the group LASSO penalty can be represented as

$$\sum_{j=1}^{J} c_j \|\beta_{A_j}\|_{R_j} \;=\; \sum_{j=1}^{J} c_j \left(\beta_{A_j}^{\top} R_j \beta_{A_j}\right)^{1/2} \;=\; \sum_{j=1}^{J} \left(\sum_{k \in A_j} c_j^2 \beta_k^2\right)^{1/2},$$

which involves two penalties, a specified bridge penalty for group selection and a weighted ridge penalty to select variables in an all-in-all-out fashion.

Later Wang et al. (2007) considered group SCAD regression analysis for microarray gene expression data with objective function defined as

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \sum_{j=1}^{J} \mathrm{P}_{\lambda,\gamma}\left(\|\beta_{A_j}\|_q\right),$$

where $\mathrm{P}_{\lambda,\gamma}$ is the SCAD penalty and q takes value 1 or 2. When $q = 1$, the penalty is called L_1 group SCAD penalty, which is a combination of the unbiased SCAD for group level and the LASSO penalty for individual variable level (thus bi-level selection). When $q = 2$, the penalty is called L_2 group SCAD penalty, a combination of the unbiased SCAD for group level and the ridge penalty for individual level (thus group selection but not individual selection).

Above gives two examples of penalties selecting important groups of variables in an all-in-all-out fashion.

1.4.3 Bi-level variable selection methods

In model selection, when one is interested in identifying not only important groups of variables but also important individual variables, bi-level variable selection techniques should be considered. Bi-level variable selection identifies zero vectors, and selects and estimates nonzero components in a nonzero vector. Penalties for bi-level variable selection can be divided into two classes according to the way they are formulated, namely, additive penalties and hierarchical penalties.

Friedman et al. (2010) proposed the sparse group LASSO penalty as a simple way to achieve bi-level selection, which is defined as

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) = \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda_1 \sum_{k=1}^{d} |\beta_k| + \lambda_2 \sum_{j=1}^{J} \|\beta_{A_j}\|_2.$$

This is an additive penalty using a LASSO penalty and a group LASSO penalty. Note that it is common to reparameterize it using λ and α, with $\lambda_1 = \alpha\lambda$ and $\lambda_2 = (1-\alpha)\lambda$. It is equivalent to LASSO when $\alpha = 1$, group LASSO when $\alpha = 0$, and a $50-50$ mix when $\alpha = 0.5$. The sparse group LASSO adds the LASSO and group LASSO penalties which are both convex, and as a result the sparse group LASSO is also convex. Following this additive fashion, one can mix other penalties to obtain different bi-level variable selection penalties.

Hierarchical penalties consist of two components, an outer penalty p_O applied to the group level and an inner penalty p_I applied to the individual variable level. The hierarchical bi-level

variable selection has an objective function given by

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) = \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \sum_{j=1}^{J} \mathrm{p}_O \left\{ \sum_{k\in A_j} \mathrm{p}_I(|\beta_k|) \right\},$$

where p_O and p_I usually depend on various tuning/regularization parameters. The group LASSO could also be thought of as in this form, with $\mathrm{p}_O(\theta) = c_j|\theta|^{1/2}$ and $\mathrm{p}_I(\beta) = \beta^2$. In this hierarchical framework, group and individual penalties interact in a multiplicative manner, as opposed to an additive manner of penalties such as sparse group LASSO. The outer penalty p_O must be nonconvex, i.e., its rate of penalization must be decreasing as the size of the group increases.

Huang et al. (2009) proposed the group bridge penalty which encourages variable selection at both individual and group levels and the resulting estimator enjoys group selection consistency. The objective function of group bridge penalized regression model is

$$\mathrm{Q}(\beta|\mathrm{X},\mathrm{y}) \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_1^{\gamma} \;=\; \mathrm{L}(\beta|\mathrm{X},\mathrm{y}) + \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k\in A_j} |\beta_k| \right)^{\gamma},$$

where λ_n is a non-negative penalty parameter, c_j is a constant used to adjust for group size d_j, i.e., $c_j \propto d_j^{1-\gamma}$ with $0 < \gamma < 1$ being a bridge index. Here we require γ strictly between 0 and 1 to have the concavity of bridge function. The group bridge uses the L_1-norm of grouped coefficients in the penalty function. Essentially, it is a combination of the bridge penalty for group selection and the LASSO penalty for within-group selection.

The implementation of LASSO penalty at individual level leads to the loss of individual selection consistency of resulting group bridge estimator. To remedy this issue, Seetharaman (2013) considered adaptive group bridge and composite group bridge penalties for linear model. The adaptive group bridge penalty uses a weighted L_1-norm of the grouped coefficients in the penalty function, given by

$$\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta^{(j)}|) \;=\; \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_{1,w}^{\gamma} \;=\; \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} w_k|\beta_k| \right)^{\gamma},$$

where $\|\cdot\|_{1,w}$ is a w-weighted L_1-norm of a vector (i.e., $\|\beta_{A_j}\|_{1,w} = \sum_{k\in A_j} w_k|\beta_k|$) and w_k is a within-group weight for β_k. Note that the performance of this adaptive group bridge estimator

heavily relies on the weights which are usually chosen adaptively as a function of some initial estimate of β.

The composite group bridge penalty is given by

$$\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta^{(j)}|) \;=\; \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_k|^{\mu} \right)^{\gamma},$$

where $\mu, \gamma \in (0,1]$ and the rest of the terms are similarly defined as that in the group bridge penalty. The composite group bridge method subsumes and extends both the bridge and group bridge approaches. When $\mu = 1$, it reduces to the group bridge method; when $\gamma = 1$, it reduces to the bridge penalty. It is also evident that the composite group bridge criterion is closely related to adaptive group LASSO method. When $\mu \in (0,1)$ and $\gamma \in (0,1)$, the composite group bridge produces an oracle estimator. Compared with group bridge penalty, the adaptive group bridge and composite group bridge estimators possess better selection consistency at within-group level. In general, a bi-level model selection procedure simultaneously performs two selection tasks: (1) discovery of the relationship between groups and the survival outcome, (2) identification of individual variables that are related to the outcome. In this thesis, we mainly consider bi-level variable selection approaches for survival models with right-censored data and data with long-term survivors, under the general semiparametric transformation models.

The rest of this thesis is organized as follows. In Chapter 2, we proposed three bi-level variable selection methods for semiparametric transformation models with right-censored data and develop effective algorithms to implement the procedures. We extended group bridge approach to semiparametric transformation models and established the $\sqrt{n}$-consistency and oracle properties of the adaptive group bridge and composite group bridge estimators for right-censored data. Simulation studies and real data analysis were conducted to illustrate the bi-level variable selection approaches. In Chapter 3, we studied the bi-level variable selection in mixture cure models for right-censored data with long-term survivors. A logistic regression model is used for modelling the cure proportion and a semiparametric transformation model is employed for fitting the survival function of uncured subjects. We proposed an effective and computationally efficient algorithm

for bi-level variable selection in mixture cure models. We used simulation studies and real data analysis to demonstrate the performance of the proposed algorithms for finite samples with right-censored data and data with long-term survivors. At last, we present a summary of the thesis and discuss possible future works in Chapters 4 and 5 respectively. Please kindly note that both Chapters 2 and 3 are written in the form of manuscripts for immediate publication to journals.

Chapter 2

BI-LEVEL SELECTION IN SEMIPARAMETRIC TRANSFORMATION MODELS WITH RIGHT-CENSORED DATA

In this chapter, we consider a large class of transformation models including the Cox PH model and proportional odds model as special cases. There are only a few literatures on variable selection for this class of models and all are at individual variable level. To fill the gap of variable selection at both group and individual levels, we consider bi-level variable selection approaches. We propose a penalized nonparametric maximum likelihood estimation (NPMLE) with three different penalties, i.e., group bridge (GB), adaptive group bridge (AGB) and composite group bridge (CGB), and develop their respective computational algorithms. We prove that the resulted estimations from AGB and CGB have desirable properties such as efficiency and oracle property. Our simulation studies demonstrate that for the semiparametric transformation models all the three penalties work well in bi-level variable selection, while AGB and CGB outperform GB when the sparsity level of significant groups is high. To demonstrate the implementation of the proposed procedures, simulation studies and two real data examples are analyzed and presented.

2.1 Introduction

In survival analysis, there is a big concentration on assessing covariate effect on censored survival outcome. Owed to the availability of advanced modern technologies, big data can be collected in a much easier way nowadays. In big data, it is common that the number of observed covariates is very large while only relatively few have effect on a response variable. Thus, to make prediction model simpler but work equivalently well as full model, it is desirable to identify those important

covariates among all collected ones. In the past few decades, there emerged abundant literature exploring techniques and computational algorithms for variable selection under linear models with non-censored data. A large number of variable selection techniques have been proposed under the framework of penalized likelihood, including the criterion-based procedures, the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the bridge penalty (Fu, 1998), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006) and the minimax concave penalty (MCP) (Zhang et al., 2011), among others. Without taking consideration of the inherent interconnections among predictors, these methods perform variable selection at individual level only. However, grouping structure arise naturally in many real cases. For example, groups of medical measurements may be taken in clinical studies; in gene expression analysis, genes belonging to the same biological pathway can be consisdered as a group; in genetic association studies, genetic markers from the same gene or haplotype can be grouped. To incorporate grouping structure, a series of group selection approaches have been developed. For instance, Yuan and Lin (2006) proposed group LASSO penalty; like the group LASSO but imposing either a SCAD penalty, bridge penalty or MCP on the norm of each group, Wang et al. (2007) used group SCAD penalty analyzing microarray gene expression data, Huang et al. (2009) introduced group bridge penalty in the general linear model setting, and Breheny and Huang (2009) investigated group MCP method. For the purpose of bi-level variable selection, Breheny and Huang (2009) studied composite MCP approach and summarized a general hierarchical framework for constructing a bi-level variable selection penalties, while Breheny (2015) proposed recently the group exponential LASSO penalty.

Some of the aforementioned variable selection criteria and procedures for linear regression analysis have been extended to censored data in survival analysis. For instance, Tibshirani et al. (1997) and Zhang and Lu (2007) extended LASSO and adaptive LASSO respectively to Cox proportional hazards (PH) model. Wang et al. (2009) proposed a hierarchically penalized Cox regression procedure with grouped variables. Fan and Li (2002) extended SCAD-penalized likelihood

approach to Cox PH model and frailty model. Lu and Zhang (2007) studied the LASSO and the adaptive LASSO procedures under proportional odds model. Wang (2016) studied adaptive elastic net for variable selection in proportional odds model. Both Cox PH model and proportional odds model belong to a class of semiparametric transformation models. Li and Gu (2012) considered adaptive LASSO for general transformation models with right censored data and Liu and Zeng (2013) developed an adaptive LASSO procedure to perform variable selection for semiparametric transformation models with right-censored data. To the best of our knowledge, there is no literature that investigates group selection or bi-level variable selection problems for semiparametric transformation models. To fill this gap, we study the bi-level variable selection problem for this class of models with right-censored data.

In this chapter, we extend the work of Liu and Zeng (2013) by incorporating grouping information in order to enjoy the advantage of identifying both important groups and important individual variables simultaneously. A group selection procedure based on a penalized weighted log-likelihood function is developed for semiparametric transformation models with right-censored data. This work can also be viewed as an extension of (Huang et al., 2014) which studied the Cox PH model, a special case in the class of semiparametric transformation models. Moreover, our proposed approach enjoys the oracle property in the sense that our estimator performs as well as the oracle one assuming the true model is known.

This chapter is organized as follows. In Section 2.2, we review the NPMLE for the semiparametric transformation models. In Section 2.3, we propose a penalized NPMLE for the purpose of bi-level variable selection with three different penalties, i.e., group bridge, adaptive group bridge and composite group bridge. The asymptotic properties of the resulted estimators are presented in Section 2.4. Their finite-sample performance are examined in Section 2.5 through simulation studies and in Section 2.6 through analysis of two real benchmark datasets. Section 2.7 gives some final discussions, while finally Section 2.8 is devoted to the proofs of theorems.

2.2 Nonparametric MLE

A general semiparametric transformation model assumes that given covariates $Z(\cdot)$, the cumulative hazard function of survival time T is

$$\Lambda(t|Z(\cdot)) \;=\; G\left[\int_0^t \exp\left\{\beta^\top Z(s)\right\} \mathrm{d}\Lambda(s)\right], \tag{2.1}$$

where $\Lambda(\cdot)$ is an unspecified increasing function and G is a known thrice continuously differentiable and strictly increasing function with $G(0)=0$, $G'(0)>0$ and $G(\infty)=\infty$. The transformation function G is assumed to have the form $G(x) = -\log\int_0^\infty \exp(xt)\phi(t)\mathrm{d}t$, where $\phi(t)$ is a known p.d.f. on $[0,\infty)$. Following Zeng and Lin (2007b), we consider the widely used logarithmic transformation class of functions

$$G(x) \;=\; \begin{cases} \dfrac{\log(1+rx)}{r}, & r>0, \\ x, & r=0. \end{cases} \tag{2.2}$$

When $G(x)=x$, i.e. $r=0$ in (2.2), it leads to Cox PH model. When $G(x)=\log(1+x)$, i.e. $r=1$ in (2.2), it reduces to the proportional odds model. This class of transformations corresponds to setting $\phi(t)$ being gamma density with unit mean and variance r.

Suppose there is a random sample of n subjects chosen from a large study cohort. Let $\tilde{T}_i$ and C_i denote the survival/failure time and censoring time respectively of the i^{th} subject. We observe the time $T_i = \min\{\tilde{T}_i, C_i\}$ and the censoring indicator $\delta_i = I(\tilde{T}_i \leqslant C_i)$, where $I(\cdot)$ is the indicator function. Let $Z_i(\cdot) = (Z_{i1}(\cdot),\ldots,Z_{id}(\cdot))^\top$ be the corresponding $d\times 1$ vector of time-varying covariates for the i^{th} subject. Thus with right-censoring, the observed data are triplets $(T_i,\delta_i,Z_i(\cdot))$, $i=1,\ldots,n$. Assume that $\tilde{T}_i$ and C_i are conditionally independent given $Z_i(\cdot)$, and that the censoring mechanism is non-informative. Under the semiparametric transformation model (2.1), the likelihood function of the observed data is

$$\begin{aligned} \prod_{i=1}^n \left(\Lambda'(T_i)\exp\left\{\beta^\top Z_i(T_i)\right\} G'\left[\int_0^{T_i}\exp\left\{\beta^\top Z_i(s)\right\}\mathrm{d}\Lambda(s)\right]\right\}^{\delta_i} \\ \times \exp\left(-G\left[\int_0^{T_i}\exp\left\{\beta^\top Z_i(s)\right\}\mathrm{d}\Lambda(s)\right]\right), \end{aligned} \tag{2.3}$$

where $\Lambda'(T_i)$ is the derivative of the cumulative hazard function Λ at time T_i.

The above likelihood function includes both the finite-dimensional parameter β and the infinite-dimensional parameter Λ, and the likelihood may not be a concave function of these parameters. Also, due to the transformation G, the partial likelihood for β does not exist. Thus, it is no longer feasible to directly apply the penalized methods in Fan and Li (2001) or Zhang and Lu (2007) for variable selection. To tackle this difficulty, we adopt the method of Zeng and Lin (2007a) by introducing a latent variable ζ which can be interpreted as the frailty in survival analysis. Then given $Z(\cdot)$ and ζ, the cumulative hazard function of survival time T in model (2.1) is equivalent to

$$\Lambda(t|Z(\cdot),\zeta) \;=\; \zeta \int_0^t \exp\left\{\beta^\top Z(s)\right\} \mathrm{d}\Lambda(s). \tag{2.4}$$

because

$$\begin{aligned} S(T|Z(\cdot)) \;&=\; \Pr(T>t|Z(\cdot)) \\ &=\; E\left[\Pr\left(T>t|Z(\cdot),\zeta\right)|Z(\cdot)\right] \\ &=\; E\left(\exp\left[-\zeta\int_0^t \exp\beta^\top Z(s)\mathrm{d}\Lambda(s)\right]|Z(\cdot)\right) \\ &=\; \int_0^\infty \exp\left[-\zeta\int_0^t \exp\beta^\top Z(s)\mathrm{d}\Lambda(s)\right]\phi(\zeta)\mathrm{d}\zeta \\ &=\; \exp\left(-G\left[\int_0^t \exp\beta^\top Z(s)\mathrm{d}\Lambda(s)\right]\right). \end{aligned}$$

Since in model (2.4) the cumulative baseline hazard function Λ is an unknown function, NPMLE could be implemented. With the latent variable ζ, the likelihood function (2.3) becomes

$$\begin{aligned} L_n(\beta,\Lambda(\cdot)|T,\delta,Z(\cdot),\zeta) \;&=\; \prod_{i=1}^n \left[\zeta_i \Delta\Lambda(T_i)\exp\left\{\beta^\top Z_i(T_i)\right\}\right]^{\delta_i} \\ &\quad\times \exp\left[-\zeta_i\int_0^{T_i}\exp\left\{\beta^\top Z_i(s)\right\}\mathrm{d}\Lambda(s)\right]\times\phi(\zeta_i), \end{aligned}$$

and the corresponding log-likelihood function is (after removing constant terms)

$$\ell_n(\beta,\Lambda(\cdot)|T,\delta,Z(\cdot),\zeta) \;=\; \sum_{i=1}^n\left[\delta_i\left\{\log\Delta\Lambda(T_i)+\beta^\top Z_i(T_i)\right\} - \zeta_i\sum_{T_k\leqslant T_i}\exp\left\{\beta^\top Z_i(T_k)\right\}\Delta\Lambda(T_k)\right]. \tag{2.5}$$

Here for NPMLE, we replace $\Lambda'(T_i)$ in the likelihood and log-likelihood functions by the discretized version $\Delta\Lambda(T_i)$, the jump size of Λ at observed failure time T_i, and we need to maximize the log-likelihood function over both β and $\Delta\Lambda(T_i)$.

Zeng and Lin (2007a) showed that the NPMLEs of β and $\Delta\Lambda(T_i)$ for recurrent events are consistent, asymptotically normal and asymptotically efficient. The maximization process is computationally intensive, especially for a large number of failures. A number of algorithms are available for the computation of NPMLEs of β and $\Delta\Lambda(T_i)$, such as the expectation-maximization (EM) algorithm employed by Zeng and Lin (2007a). With the complete data $(T_i, \delta_i, Z_i(\cdot), \zeta_i)$, $i = 1, \ldots, n$, the EM algorithm applied to the log-likelihood function (2.5) under model (2.4) includes the following two steps.

E-Step: Compute the expected log-likelihood based on current estimates $\tilde{\beta}^{(m-1)}$ and $\widetilde{\Delta\Lambda}(T_i)^{(m-1)}$, conditional on the observed data. Specifically, first compute the posterior expectation of the latent variables, i.e. $\bar{\zeta}_i^{(m)} \equiv E\left[\zeta_i | T, \delta, Z(\cdot), \tilde{\beta}^{(m-1)}, \widetilde{\Delta\Lambda}(T_i)^{(m-1)}\right]$, $i = 1, \ldots, n$, at the m^{th} iteration, based on the posterior density of ζ_i. Then the conditional expectation of the log-likelihood on $\tilde{\beta}^{(m-1)}$, $\widetilde{\Delta\Lambda}(T)^{(m-1)}$ and data is

$$\begin{aligned} & Q\left(\beta, \Delta\Lambda(T) | T, \delta, Z(\cdot), \tilde{\beta}^{(m-1)}, \widetilde{\Delta\Lambda}(T)^{(m-1)}\right) \\ = \; & E_{\zeta | T, \delta, Z(\cdot), \tilde{\beta}^{(m-1)}, \widetilde{\Delta\Lambda}(T)^{(m-1)}}\left[\ell_n(\beta, \Lambda(\cdot) | T, \delta, Z(\cdot), \zeta)\right] \\ = \; & \sum_{i=1}^{n}\left[\delta_i\left\{\log\Delta\Lambda(T_i) + \beta^\top Z_i(T_i)\right\} - \bar{\zeta}_i \sum_{T_k \leqslant T_i} \exp\left\{\beta^\top Z_i(T_k)\right\}\Delta\Lambda(T_k)\right\}. \end{aligned} \tag{2.6}$$

Here $T = (T_1, \ldots, T_n)^\top$, $\delta = (\delta_1, \ldots, \delta_n)^\top$, $Z = (Z_1, \ldots, Z_n)^\top$, $\Delta\Lambda(T) = (\Delta\Lambda(T_1), \ldots, \Delta\Lambda(T_n))^\top$ and $\widetilde{\Delta\Lambda}(T) = (\widetilde{\Delta\Lambda}(T_1), \ldots, \widetilde{\Delta\Lambda}(T_n))^\top$

M-Step: Compute the estimates $\beta^{(m)}$ and $\Delta\Lambda(T)^{(m)}$ which maximize the expected log-likelihood $Q\left(\beta^{(m)}, \Delta\Lambda(T)^{(m)} | T, \delta, Z(\cdot), \tilde{\beta}^{(m-1)}, \widetilde{\Delta\Lambda}(T)^{(m-1)}\right)$ obtained in the E-step.

After convergence, we obtain the NPMLEs $\tilde{\beta}$ and $\widetilde{\Delta\Lambda}(T_i)$'s. Based on the NPMLEs, we compute the posterior expectation $E[\zeta_i | T, \delta, Z(\cdot), \tilde{\beta}, \widetilde{\Delta\Lambda}(T)]$ which is denoted as weight $\tilde{e}_i$, $i = 1, \ldots, n$. As

shown in the Appendix of (Zeng and Lin, 2007a),

$$\tilde{e}_i \;=\; \begin{cases} G'\left[\sum_{k=1}^{n} I\left(T_k \leqslant T_i\right) \exp\left\{\beta^\top Z_i(T_k)\right\} \Delta\Lambda(T_k)\right], & \delta_i = 0, \\ -\dfrac{G''\left[\sum_{k=1}^{n} I\left(T_k \leqslant T_i\right) \exp\left\{\beta^\top Z_i(T_k)\right\} \Delta\Lambda(T_k)\right]}{G'\left[\sum_{k=1}^{n} I\left(T_k \leqslant T_i\right) \exp\left\{\beta^\top Z_i(T_k)\right\} \Delta\Lambda(T_k)\right]} & \\ +G'\left[\sum_{k=1}^{n} I\left(T_k \leqslant T_i\right) \exp\left\{\beta^\top Z_i(T_k)\right\} \Delta\Lambda(T_k)\right], & \delta_i = 1. \end{cases}$$

Given $\tilde{w}_i$, we differentiate the function (2.6) with respect to $\Delta\Lambda(T_i)$ and set it to zero, obtaining

$$\Delta\Lambda(T_i) \;=\; \frac{\delta_i}{\sum_{k=1}^{n} I\left(T_k \geqslant T_i\right) \tilde{e}_k \exp\left\{\beta^\top Z_k(T_i)\right\}}.$$

Substituting $\Delta\Lambda(T_i)$ back into function (2.5) gives

$$\ell_n(\beta) \;=\; \sum_{i=1}^{n} \delta_i \left(\beta^\top Z_i(T_i) - \log\left[\sum_{k=1}^{n} I(T_k \geqslant T_i) \tilde{e}_k \exp\left\{\beta^\top Z_k(T_i)\right\}\right]\right), \tag{2.7}$$

a function of β very similar to the Cox partial log-likelihood. This function (2.7) is called weighted version of the partial log-likelihood function which is used to incorporate penalties for variable selection. As pointed out by Zeng and Lin (2007a,b), (2.7) inherits several good properties. For instance, the resulted estimator $\tilde{\beta}$ is an efficient MLE of β and it is a strictly concave function of β. These advantages of function (2.7) ensure nice theoretical properties of the estimator after variable selection.

2.3 Bi-level Variable Selection

To include grouping structure in the transformation model (2.4), we assume that the d variables in the possibly time-varying covariate vector $Z(\cdot)$ can be categorized into J factors (groups of variables). Let A_j be a size d_j subset of $\{1,\ldots,d\}$ representing known grouping structure of the design vectors, $j = 1,\ldots,J$. Here we consider the scenario that the d variables are partitioned into J groups, and thus $d = \sum_{j=1}^{J} d_j$. Let $\beta_{A_j} \in \mathbb{R}^{d_j}$ be the coefficient vector corresponding to the j^{th} factor consisting of d_j input variables.

For the purpose of variable selection in transformation models, l_q-penalized regression is considered where the regression coefficients are partly or all penalized via a l_q-norm penalty to achieve a parsimonious statistical model and high prediction accuracy. In general, let β denote the parameter of interest, then the penalized estimator of β is obtained by minimizing $Loss(\beta) + p_\lambda(\beta)$, where the loss function $Loss(\beta)$ is usually chosen to be either the residual sum of squared errors (RSS) in regression or negative log-likelihood function otherwise, the penalty function $p_\lambda(\cdot)$ measures the complexity of the enclosed coefficient vector and λ is a non-negative tuning parameter controlling the degree of penalization. Now under the semiparametric transformation model (2.4) with grouping information, variable selection can be realized by minimizing the penalized negative partial log-likelihood (2.7) in a weighted version

$$Q_n(\beta) \;=\; -\frac{1}{n}\ell_n(\beta) + \sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta_{A_j}|). \tag{2.8}$$

For choosing the tuning parameter λ_n, Akaike information criterion (AIC), Bayesian information criterion (BIC) and generalized cross validation score (GCV) are reasonable and commonly used criteria since the unknown $\beta = \left(\beta_{A_1}^\top, \ldots, \beta_{A_J}^\top\right)^\top \in \mathbb{R}^d$ is the parameter of interest, we aim at bi-level selection, i.e., identifying both non-zero vectors and non-zero components in non-zero vectors in β.

2.3.1 Penalty functions

To accommodate grouping information in semiparametric transformation models, we consider three different penalty functions: group bridge, adaptive group bridge, and composite group bridge.

(a) Group bridge penalty

Group bridge penalty was first proposed in Huang et al. (2009) under multivariate linear regression model which simultaneously selects variables at both group level and individual variable level. As shown in (Huang et al., 2009), it enjoys the oracle property in group selection in the sense that it can correctly select important groups with probability converging to one. Afterwards, it has

been extended to the Cox PH model by Huang et al. (2014). The group bridge penalty function is defined as

$$\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta_{A_j}|) \;=\; \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_1^{\gamma} \;=\; \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k \in A_j} |\beta_k| \right)^{\gamma}, \tag{2.9}$$

where the constant $\gamma \in (0,1)$ and c_j's are some between-group level weights accounting for group sizes. According to Huang et al. (2009), a simple choice of c_j is $c_j \propto d_j^{1-\gamma}$ adjusting for the size of each group of coefficients. The value β that minimizes $Q_n(\beta)$ in (2.8) with the group bridge penalty (2.9) is referred as a group bridge estimator. Note that when $d_j = 1$, the group bridge penalty in (2.9) simplifies to the standard bridge penalty. If further $c_j = 1$ and $\gamma = 1$, it is reduced to the standard LASSO penalty. By the last representation in (2.9), the group bridge penalty could be viewed as imposing a LASSO penalty at within-group level while employing a bridge penalty at between-group level. Due to the essence of LASSO, it fails to achieve the selection consistency at within-group level. This sparked more recent research on developing estimators which possess the oracle property at both levels. Since the adaptive LASSO penalty is a natural extension of LASSO penalty, a simple remedy named adaptive group bridge penalty, was proposed by Wang et al. (2009), in which the adaptive LASSO penalty is used at within-group level.

(b) Adaptive group bridge penalty

Adaptive group bridge penalty is defined as

$$\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta_{A_j}|) \;=\; \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_{1,w}^{\gamma} \;=\; \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k \in A_j} w_k |\beta_k| \right)^{\gamma}, \tag{2.10}$$

where w_k's are some within-group level weights, $\|\cdot\|_{1,w}$ denotes the w-weighted l_1-norm of a vector (i.e., $\|\beta_{A_j}\|_{1,w} = \sum_{k \in A_j} w_k |\beta_k|$), c_j's are some between-group level weights, and the rest of the terms are similarly defined as in (2.9). The value β that minimizes $Q_n(\beta)$ in (2.8) with the adaptive group bridge penalty (2.10) is referred as an adaptive group bridge estimator. The w_k's are usually chosen in the following way. Suppose some reliable initial estimator $\tilde{\beta}$ is available, e.g. a least-squares estimator when sample size is large relatively to model dimension, then we set

$w_k = |\tilde{\beta}_k|^{-\mu}$, where μ is a power parameter usually set to be 2. The adaptive group bridge estimator has bi-level selection consistency, however it has a drawback that its performance heavily relies on the quality of the chosen initial estimator through the adaptive weights.

(c) Composite group bridge penalty

Combining bridge penalty and group bridge penalty, Seetharaman (2013) proposed a composite group bridge penalty to conduct bi-level variable selection under multivariate linear penalized regression. To induce within-group selection consistency, another bridge penalty is adopted to replace the l_1-penalty in group bridge. The composite group bridge penalty has the form

$$\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta_{A_j}|) \;=\; \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_{\mu}^{\mu\gamma} \;=\; \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k \in A_j} |\beta_k|^{\mu} \right)^{\gamma}, \tag{2.11}$$

where $\mu, \gamma \in (0,1]$ and the rest of the terms are similarly defined as in (2.9). When $\mu = 1$ and $\gamma \neq 1$, it reduces to the group bridge; when $\gamma = 1$, it reduces to the form of bridge regression. The composite group bridge method subsumes and extends both the bridge and group bridge approaches. It is also evident that the composite group bridge criterion is closely related to adaptive group LASSO method. We shall mainly consider $\mu, \gamma \in (0,1)$, as we will show later that this region ensures the oracle property.

2.3.2 Computational algorithms

Due to the non-convexity of the penalty functions, it is challenging to directly maximize the objection function $Q_n(\beta)$ in (2.8) with the three penalties listed in Section 2.3.1. Following Huang et al. (2009, 2014), we formulate an equivalent minimization problem that is easier to be solved computationally. Let $\nabla \ell_n(\beta) = -\partial \ell_n(\beta)/\partial \beta$ and $\nabla^2 \ell_n(\beta) = -\partial^2 \ell_n(\beta)/\partial \beta^2$ denote the gradient vector and the Hessian matrix of ℓ_n respectively. Consider the Cholesky decomposition of $\nabla^2 \ell_n(\beta)$, i.e., $\nabla^2 \ell_n(\beta) = X^\top X$, and set the pseudo-response vector to be $Y = (X^\top)^{-1}[\nabla^2 \ell_n(\beta)\beta - \nabla \ell_n(\beta)]$. Then $-\ell_n(\beta)$ can be approximated by the quadratic form $\frac{1}{2}(Y - X\beta)^\top (Y - X\beta)$.

(a) Group bridge algorithm

Recall for group bridge estimation, the objective function is $Q_n(\beta) = -\frac{1}{n}\ell_n(\beta) + \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_1^{\gamma}$. Following Huang et al. (2014), it is computational easier and more efficient to consider an equivalent minimization of function

$$W_n(\beta,\theta) \;=\; \frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\beta_{A_j}\|_1 + \tau_n \sum_{j=1}^{J} \theta_j$$

over (β,θ), where τ_n is a penalty tuning parameter. Huang et al. (2009) or Proposition 1 of Huang et al. (2014) proved that minimizing $Q_n(\beta)$ in (2.8) is equivalent to minimizing $W_n(\beta,\theta)$ if

$$\lambda_n = \tau_n^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}. \tag{2.12}$$

The minimization of W_n could be treated as a LASSO-type optimization problem, and thus the shooting algorithm and coordinate descent algorithm could be applied to solve it numerically. The iterative algorithm is given below.

Step 1. Use the EM algorithm presented in Section 2.3 to compute $\tilde{\beta}$ and $\widetilde{\Delta\Lambda}(T_i)$'s, then compute the weights $\tilde{e}_i$'s and formulate the weighted version of the partial log-likelihood function $\ell_n(\beta)$ given in (2.7).

Step 2. Initialize by setting $\beta^{(0)} = \tilde{\beta}$.

Step 3. At the m^{th} step, compute X and Y through the cholesky decomposition based on $\beta^{(m-1)}$ from previous step.

Step 4. Compute

$$\theta_j^{(m)} \;=\; c_j \left(\frac{1-\gamma}{\tau_n\gamma}\right)^{\gamma} \|\beta_{A_j}^{(m-1)}\|_1^{\gamma}, \;\; j = 1,\ldots,J.$$

Step 5. Use coordinate descent algorithm to calculate the updated estimate

$$\beta^{(m)} \;=\; \arg\min_{\beta} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{J} (\theta_j^{(m)})^{1-1/\gamma} c_j^{1/\gamma} \|\beta_{A_j}\|_1 \right\}.$$

Step 6. Repeat Steps 3-5 until $\|\beta^{(m)} - \beta^{(m-1)}\|_2 / \|\beta^{(m-1)}\|_2$ is small.

Since the group bridge penalty function is not convex, for a fixed τ_n value the above algorithm may only produce a local minimizer of $W_n(\beta,\theta)$. To solve this issue, based on the relationship between τ_n and λ_n in (2.12), we consider the grid values of tuning parameter λ over the range $[0,40]$ with increment size 0.2 in our simulation studies, and choose the one that gives the smallest value of $W_n(\beta,\theta)$ for the final group bridge estimate calculation.

(b) Adaptive group bridge algorithm

Like the algorithm for group bridge in part (a), the algorithm for adaptive group bridge approach needs one more initialization in Step 1, for the within-group weights w_k's, but has the rest steps the same. Similar to group bridge method, under (2.12) minimizing the objective function $Q_n(\beta)$ with adaptive group bridge penalty (2.10) is equivalent to minimizing

$$W_n(\beta,\theta) \;=\; \frac{1}{2}\|Y-X\beta)\|_2^2+\sum_{j=1}^{J}\theta_j^{1-1/\gamma}c_j^{1/\gamma}\|\beta_{A_j}\|_{1,w}+\tau_n\sum_{j=1}^{J}\theta_j.$$

The iterative algorithm is given as following.

Step 1. Use the EM algorithm presented in Section 2.3 to compute $\tilde{\beta}$ and $\widetilde{\Delta\Lambda}(T_i)$'s, then compute the weights $\tilde{e}_i$'s and formulate the weighted version of the partial log-likelihood function $\ell_n(\beta)$ given in (2.7).

Step 2. Initialize by setting $\beta^{(0)}=\tilde{\beta}$ and $w_k=|\tilde{\beta}_k|^{-\mu}$.

Step 3. At the m^{th} step, compute X and Y through the cholesky decomposition based on $\beta^{(m-1)}$ from previous step.

Step 4. Compute

$$\theta_j^{(m)} \;=\; c_j\left(\frac{1-\gamma}{\tau_n\gamma}\right)^{\gamma}\|\beta_{A_j}^{(m-1)}\|_{1,w}^{\gamma},\;\; j=1,\ldots,J.$$

Step 5. Use coordinate descent algorithm to calculate the updated estimate

$$\beta^{(m)} \;=\; \arg\min_{\beta}\left\{\frac{1}{2}\|Y-X\beta\|_2^2+\sum_{j=1}^{J}(\theta_j^{(m)})^{1-1/\gamma}c_j^{1/\gamma}\|\beta_{A_j}\|_{1,w}\right\}.$$

Step 6. Repeat Steps 3-5 until $\|\beta^{(m)} - \beta^{(m-1)}\|_2 / \|\beta^{(m-1)}\|_2$ is small.

(c) Composite group bridge algorithm

For composite group bridge approach, Seetharaman (2013) proved that minimizing the objective function $Q_n(\beta)$ with composite group bridge penalty (2.11) is equivalent to minimizing

$$W_n(\beta, \theta, \nu) \;=\; \frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \left(\sum_{k \in A_j} \nu_k^{1-1/\mu} |\beta_k| + \psi \sum_{k \in A_j} \nu_k \right) + \tau_n \sum_{j=1}^{J} \theta_j$$

over (β, θ, ν) if (2.12) holds and

$$\psi \;=\; \mu^{\mu/(1-\mu)}(1-\mu), \tag{2.13}$$

where $(\tau_n, \psi)^\top$ are some penalty parameters. Minimizing W_n over β with (θ, ν) held fixed, the optimization problem turns into an adaptive LASSO problem, which could be efficiently solved by implementing the iteratively re-weighted LASSO regression algorithm given as following.

Step 1. Use the EM algorithm presented in Section 2.3 to compute $\tilde{\beta}$ and $\widetilde{\Delta\Lambda}(T_i)$'s, and then compute the weights $\tilde{e}_i$'s and formulate the weighted version of the partial log-likelihood function $\ell_n(\beta)$ given in (2.7).

Step 2. Initialize by setting $\beta^{(0)} = \tilde{\beta}$.

Step 3. At the m^{th} step, compute X and Y through the cholesky decomposition based on $\beta^{(m-1)}$ from previous step.

Step 4. Compute

$$\begin{aligned} \theta_j^{(m)} \;&=\; c_j \left(\frac{1-\gamma}{\tau_n \gamma} \right)^{\gamma} \|\beta_{A_j}^{(m-1)}\|_\mu^{\mu\gamma}, \;\; j = 1, \ldots, J, \\ \nu_k^{(m)} \;&=\; \mu^{\mu/(\mu-1)} |\beta_k^{(m-1)}|^\mu, \;\; k = 1, \ldots, d. \end{aligned}$$

Step 5. Use coordinate descent algorithm to calculate the updated estimate

$$\begin{aligned} \beta^{(m)} \;&=\; \arg\min_{\beta} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{J} \left(\theta_j^{(m)}\right)^{1-1/\gamma} c_j^{1/\gamma} \sum_{k \in A_j} \left(\nu_k^{(m)}\right)^{1-1/\mu} |\beta_k| \right\}, \\ &=\; \arg\min_{\beta} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_n \sum_{k \in A_j} w_k^{(m)} |\beta_k| \right\}, \end{aligned}$$

where $w_k^{(m)} = \gamma\mu \sum_{j:k\in A_j} c_j \|\beta_{A_j}^{(m-1)}\|_\mu^{\mu\gamma-\mu} |\beta_k^{(m-1)}|^{\mu-1}$.

Step 6. Repeat Steps 3-5 until $\|\beta^{(m)} - \beta^{(m-1)}\|_2 / \|\beta^{(m-1)}\|_2$ is small.

The algorithm alternatively updates (θ, ν) and β through a blockwise coordinate descent algorithm. It always converges since at each step the non-negative objective function $W_n(\beta, \theta, \nu)$ decreases. Here θ and ν could be interpreted as group level and individual level weights adjusting the respective scales at different levels.

2.3.3 Tuning parameter selection

A tuning parameter in variable selection balances goodness-of-fit and sparsity, and thus the choice of a tuning parameter is essential to the performance of the penalized log-likelihood procedures discussed in this chapter. In practice, it is often to set a range for a tuning parameter and find the corresponding estimator for each fixed tuning parameter value in the range. Then a tuning parameter value is selected which optimizes a certain model selection criterion. In this chapter, tuning parameters are selected via the minimization of criterion AIC, BIC or GCV. Specifically, let $\hat{d}(\lambda_n)$ denote the cardinality of the set of indices of nonzero coefficients for a fixed tuning parameter λ_n. Then the AIC, BIC and GCV are given respectively by

$$AIC(\lambda_n) = \log\left[\frac{\ell_n(\hat{\beta})}{n}\right] + \frac{2\hat{d}(\lambda_n)}{n},$$

$$BIC(\lambda_n) = \log\left[\frac{\ell_n(\hat{\beta})}{n}\right] + \frac{\log(n)\hat{d}(\lambda_n)}{n},$$

$$GCV(\lambda_n) = \frac{\ell_n(\hat{\beta})}{n\left[1 - \hat{d}(\lambda_n)/n\right]^2}.$$

2.4 Asymptotic Properties

In the penalized objective function $Q_n(\beta)$ given by (2.8), $p_{\lambda_n}^{(j)}(|\beta_{A_j}|)$ is generally a d_j-variate penalty function of the parameters in the j^{th} group and satisfies conditions

$$p_{\lambda_n}^{(j)}(|\beta_{A_j}|) \geqslant 0, \;\; p_{\lambda_n}^{(j)}(0) = 0,$$

$$p_{\lambda_n}^{(j)}(|\beta_{A_j}|) \geqslant p_{\lambda_n}^{(j)}(|\beta_{A_j}^*|) \text{ if } |\beta_k| \geq |\beta_k^*| \text{ for all } k \in A_j.$$

Denote the true values of β and Λ by β_0 and Λ_0 respectively. We rewrite $\beta_0 = \left(\beta_{0,\mathscr{A}}^\top, \beta_{0,\mathscr{B}}^\top, \beta_{0,\mathscr{C}}^\top\right)^\top$, where $\mathscr{A} = \{k : \beta_{0,k} \neq 0\}$, $\mathscr{B} = \{k \in A_j : \beta_{0,k} = 0 \text{ and } \beta_{0,A_j} \neq 0\}$, and $\mathscr{C} = \{A_j : \beta_{0,A_j} = 0\}$, i.e., $\mathscr{A}$ contains the indices of nonzero coefficients, $\mathscr{B}$ contains the indices of all zero coefficients in nonzero groups, and $\mathscr{C}$ contains the indices of zero coefficients in zero groups. Thus $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$ are disjoint and partition the set of all indices of the d coefficients. We write $\mathscr{D} = \mathscr{B} \cup \mathscr{C}$, which contains the indices of all zero coefficients.

Let $I(\beta_0)$ denote the Fisher information matrix based on the partial log-likelihood function of $\beta = \beta_0$, and $I_{\mathscr{A}}(\beta_{0,\mathscr{A}})$ the Fisher information matrix of $\beta_{\mathscr{A}} = \beta_{0,\mathscr{A}}$ knowing that $\beta_{0,\mathscr{D}} = 0$. Define

$$a_n = \max_{(j,k)} \left\{ \frac{\partial p_{\lambda_n}^{(j)}(|\beta_{0,A_j}|)}{\partial |\beta_{0,k}|} : k \in A_j, \beta_{0,k} \neq 0 \right\}, \tag{2.14}$$

$$b_n = \max_{(j,k)} \left\{ \frac{\partial^2 p_{\lambda_n}^{(j)}(|\beta_{0,A_j}|)}{\partial |\beta_{0,k}|^2} : k \in A_j, \beta_{0,k} \neq 0 \right\}. \tag{2.15}$$

We write the time-varying covariates $Z(\cdot)$ as $(Z_{\mathscr{A}}^\top(\cdot), Z_{\mathscr{D}}^\top(\cdot))^\top$, where $Z_{\mathscr{A}}(\cdot)$ and $Z_{\mathscr{D}}(\cdot)$ denote the important and the unimportant covariates corresponding to the index sets $\mathscr{A}$ and $\mathscr{D}$ respectively. Assume the following regularity conditions.

Condition 1. The function $\Lambda_0(t)$ is strictly increasing and continuously differentiable, and $\beta_0 \in \text{int}(\Omega)$ with known $\Omega \subset \mathbb{R}^d$.

Condition 2. With probability one, $Z(\cdot)$ has bounded total variation over $[0, \tau]$, where τ is the follow-up time. In addition, if there exists a vector α and a deterministic function $\alpha_0(t)$ such that with probability one $\alpha_0(t) + \alpha^\top Z(t) = 0$ for all t, then $\alpha = 0$ and $\alpha_0(t) = 0$.

Condition 3. With probability one, there exists a positive constant a_0 such that $\mathbb{P}(C \geqslant \tau | Z) > a_0$ and $\mathbb{P}(T \geqslant \tau | Z) > a_0$.

Condition 4. For any positive constant m_0, $\limsup_{x\to\infty} G^{-1}(m_0 x)\log\left[x\sup_{y\leqslant x} G'(y)\right] = 0$.

Condition 5.

$$\frac{\lambda_n^2}{n}\sum_{j=1}^{J} c_j^2 \|\beta_{0A_j}\|_1^{2\gamma-2}|A_j| \leqslant dM_n, \quad M_n = O_p(1),$$

where the constants c_j's satisfy $\min_{1\leqslant j\leqslant J} c_j \geqslant 1$ and $n^{1/2}\lambda_n = O_p(1)$ and $O_p(1)$ denotes a sequence bounded in probability.

These conditions are also used in Zeng and Lin (2006). *Conditions 1* and *3* are standard in survival models. *Condition 2* is equivalent to that the design matrix $(1, Z^\top(t))^\top$ is of full rank with positive probability for all $t \in [0,\tau]$, and is used to ensure strict concavity of the objective function $\ell_n(\beta)$. *Condition 4* specifies the tail behaviour of the transformation function $G(x)$ and it is easy to check that the logarithmic transformation (2.2) satisfies this condition.

The following two theorems give the asymptotic properties of the MLE $\hat{\beta}$ that minimizes the penalized partial likelihood (2.8) with penalty function either the group bridge (2.9), the adaptive group bridge (2.10) or the composite group bridge (2.11).

Theorem 2.4.1. *Under regularity conditions (A)-(D) of Andersen and Gill (1982), if $a_n = O_p(n^{-1/2})$ and $b_n \to 0$ with a_n and b_n defined in (2.14) and (2.15) respectively, then there exists a local minimizer $\hat{\beta}$ of $Q_n(\beta)$ in (2.8) with penalty function (2.9), (2.10) or (2.11) such that $\|\hat{\beta} - \beta_0\|_2 = O_p(n^{-1/2})$.*

Theorem 2.4.2. *Assume the conditions in Theorem 2.4.1 hold and, let $\hat{\beta} = \left(\hat{\beta}_{0,\mathscr{A}}^\top, \hat{\beta}_{0,\mathscr{B}}^\top, \hat{\beta}_{0,\mathscr{C}}^\top\right)^\top$ be a $\sqrt{n}$-consistent local minimizer of $Q_n(\beta)$ with penalty function (2.9), (2.10) or (2.11). Assume that the regularity conditions (A)-(D) of Andersen and Gill (1982) hold. Then*

(a) For $k \in \mathscr{D}$, i.e. $\beta_{0,k} = 0$, if $k \in A_j$ and $n^{1/2}\partial p_{\lambda_n}^{(j)}(|\hat{\beta}_{0,A_j}|)/\partial|\beta_{0,k}| \to \infty$ as $n \to \infty$, then we have $\hat{\beta}_k = 0$ with probability approaching one.

(b) For any $k \in \mathscr{A}$, i.e. $\beta_{0,k} \neq 0$, if $k \in A_j$, $b_n \to 0$ and $n^{1/2}\partial p_{\lambda_n}^{(j)}(|\beta_{0,A_j}|)/\partial|\beta_{0,k}| \to 0$ as $n \to \infty$, then under (a) we have that $n^{1/2}(\hat{\beta}_{\mathscr{A}} - \beta_{0,\mathscr{A}})$ converges in distribution to a zero-mean normal random variable with covariance matrix $I_{\mathscr{A}}^{-1}(\beta_{0,\mathscr{A}})$.

Theorem 1 indicates that by choosing proper λ_n and penalty functions $p_{\lambda_n}^{(j)}$'s, there exists a $\sqrt{n}$-consistent penalized partial log-likelihood estimator. Theorem 2 implies that by choosing proper λ_n and $p_{\lambda_n}^{(j)}$'s, the corresponding penalized partial log-likelihood estimator possesses the sparsity property that $\hat{\beta}_{\mathscr{D}} = 0$ with probability tending to one. Furthermore, the estimators of the nonzero coefficients, $\hat{\beta}_{\mathscr{A}}$, have the same asymptotic distribution as what they would have if the zero coefficients were known in advance. Asymptotically, therefore, the penalized partial log-likelihood estimator $\hat{\beta}$ performs as well as if the true underlying model was provided in advance, i.e., it possesses the oracle property of Fan and Li (2001, 2002).

Corollary 2.4.1. *If $\lambda_n = O_p(n^{-1/2})$, then there exists a $\sqrt{n}$-consistent local maximizer $\hat{\beta}_n = (\hat{\beta}_{n,\mathscr{A}}^\top, \hat{\beta}_{n,\mathscr{B}}^\top, \hat{\beta}_{n,\mathscr{C}}^\top)^\top$ for the semiparametric transformation model (2.4) with adaptive group bridge penalty; if further $n^{3/4}\lambda_n \to \infty$ as $n \to \infty$, then $\hat{\beta}_{n,\mathscr{C}} = 0$ with probability tending to one.*

Corollary 2.4.2. *If $\lambda_n = O_p(n^{-1/2})$, then there exists a root-n consistent local maximizer $\hat{\beta}_n = (\hat{\beta}_{n,\mathscr{A}}^\top, \hat{\beta}_{n,\mathscr{B}}^\top, \hat{\beta}_{n,\mathscr{C}}^\top)^\top$ for the composite group bridge penalized semiparametric transformation model (2.4); if further $n^{3/4}\lambda_n \to \infty$ as $n \to \infty$, then $\hat{\beta}_{n,\mathscr{C}} = 0$ with probability tending to one.*

Proofs of these corollaries follow the spirit of Fan and Li (2001, 2002) but are nontrivial extensions due to the grouping structure of the penalty. The proofs are deferred to Section 2.8.

2.5 Simulation Studies

In this section, we compare through simulation studies the performance of the three estimators based on group bridge, adaptive group bridge and composite group bridge respectively for semiparametric transformation models (2.2).

Failure times are generated from transformation models with logarithmic transformation function G specified by (2.2). We consider five transformation models indexed by $r = 0, 0.3, 0.5, 0.7, 1$. Note that $r = 0$ and $r = 1$ correspond to the proportional hazards model and the proportional odds model respectively. A Weibull distribution $\Lambda(t) = at^b$ with some $a, b > 0$ is assumed as the baseline cumulative hazard function. Specifically, the failure times $T = \{[(1/U)^r - 1]\exp(-\beta^\top Z)/(ar)\}^{1/b}$, where U is a random variable generated from the uniform distribution over $(0,1)$. Censoring times are generated from a uniform distribution over $(c/2, c)$, where c is chosen to obtain a censoring rate around 20%. The grouping structure of covariates and the coefficient values are described below for five different cases with Cases 1 and 2 for smaller number of covariates and Cases 3, 4 and 5 for larger number of covariates. In Cases 1, 2 and 3 the group sizes are the same while Cases 4 and 5 consider different group sizes. In addition, in Cases 1 and 3 the coefficients within each group are either all zeros or all non-zeros, while Cases 2, 4 and 5 consider mixed groups with some zeros and some non-zeros. In Cases 1 and 3, the sparsity level in the nonzero groups are zero, while in Cases 4, 2 and 5, it increases to 50%, 66.7% and 77.5%, respectively.

Case 1. There are 5 groups with 3 covariates within each group, and thus totally 15 covariates in the regression model. The true values of β_i's are taken to be $(\beta_1, \beta_2, \beta_3)^\top = (0.5, 1, 1.5)^\top$, $(\beta_4, \beta_5, \beta_6)^\top = (1,1,1)^\top$, $(\beta_7, \ldots, \beta_9)^\top = (0, \ldots, 0)^\top$, $(\beta_{10}, \ldots, \beta_{12})^\top = (0, \ldots, 0)^\top$, $(\beta_{13}, \ldots, \beta_{15})^\top = (0, \ldots, 0)^\top$. The covariates $(z_1, \ldots, z_{15})^\top$ are generated in the following way. We first simulate $V_1, \ldots, V_{15}$ from the standard normal distribution, and independently $Z_1, \ldots, Z_5$ from an $AR(1)$ model with the initial standard normal distribution and $\mathrm{Cov}(Z_i, Z_j) = 0.4^{|i-j|}$, $i, j = 1, \cdots, 5$. Then we set the covariate $z_j = (Z_{g_j} + V_j)/4$, $j = 1, \cdots, 15$, where g_j is the smallest integer greater than $(j-1)/3$, and the z_j's with the same value of g_j belong to the same group.

Case 2. For comparison purpose, we set up Case 2 similar to Case 1 in terms of group size, numbers of covariates and number of groups, but we decrease the number of nonzero covariates in nonzero groups to increase the level of sparsity within groups. There are 5 groups with 3 covariates within each group. The true values of β_i's are taken to be $(\beta_1, \beta_2, \beta_3)^\top = (1,0,0)^\top$, $(\beta_4, \beta_5, \beta_6)^\top =$

$(0,1.5,0)^\top$, $(\beta_7,\ldots,\beta_9)^\top=(0,\ldots,0)^\top$, $(\beta_{10},\ldots,\beta_{12})^\top=(0,\ldots,0)^\top$, $(\beta_{13},\ldots,\beta_{15})^\top=(0,\ldots,0)^\top$. The 15 covariates $(z_1,\ldots,z_{15})^\top$ are generated in the same manner as in Case 1.

Case 3. There are 10 groups with 4 covariates within each group. The true values of β_i's are taken to be $(\beta_1,\ldots,\beta_4)^\top=(1,0.5,1,0.5)^\top$, $(\beta_5,\ldots,\beta_8)^\top=(0,\ldots,0)^\top$, $(\beta_9,\ldots,\beta_{12})^\top$ $=(-1.5,-1.5,-1.5,-1.5)^\top$, $(\beta_{13},\ldots,\beta_{16})^\top=(1.5,-1,-1.5,1)^\top$, $(\beta_{17},\ldots,\beta_{40})^\top=(0,\ldots,0)^\top$. The 40 covariates $(z_1,\ldots,z_{40})^\top$ are generated in a similar manner as in Case 1.

Case 4. There are 6 groups with 3 of larger size and 3 of smaller size. The true values of β_i's are taken to be $(\beta_1,\ldots,\beta_{10})^\top=(0,1,0,1,0,0,1,0,1,0)^\top$, $(\beta_{11},\ldots,\beta_{20})^\top=(0,\ldots,0)^\top$, $(\beta_{21},\ldots,\beta_{30})^\top$ $=(0,\ldots,0)^\top$, $(\beta_{31},\ldots,\beta_{34})^\top=(0,0,-1,1)^\top$, $(\beta_{35},\ldots,\beta_{38})^\top=(-1,1,0,0)^\top$, $(\beta_{39},\ldots,\beta_{42})^\top$ $=(0,\ldots,0)^\top$. The 42 covariates $(z_1,\ldots,z_{42})^\top$ are generated in a similar manner as in Case 1.

Case 5. There are 6 groups with 3 of larger size and 3 of smaller size. The true values of β_i's are taken to be $(\beta_1,\ldots,\beta_{10})^\top=(0,1,0,1,0,\ldots,0)^\top$, $(\beta_{11},\ldots,\beta_{20})^\top=(-1,-1,0,\ldots,0)^\top$, $(\beta_{21},\ldots,\beta_{30})^\top=(0,\ldots,0)^\top$, $(\beta_{31},\ldots,\beta_{34})^\top=(0,0,0,1)^\top$, $(\beta_{35},\ldots,\beta_{38})^\top=(-1,0,0,0)^\top$, $(\beta_{39},\ldots,\beta_{42})^\top=(0,\ldots,0)^\top$. The 42 covariates $(z_1,\ldots,z_{42})^\top$ are generated in a similar manner as in Case 4.

The simulation results for Cases 1, 2, 3, 4 and 5 are presented in Tables 2.1, 2.2, 2.3, 2.4 and 2.5 respectively. All the results are based on 400 repetitions with sample size $n=100$. Results for different sample sizes such as $n=200$ are similar and thus omitted to save space. For each simulated dataset we apply group bridge (GB), adaptive group bridge (AGB) and composite group bridge (CGB) approaches for parameter estimation and group/variable selection. For simplicity we use the same $\gamma=0.5$ in (2.9), (2.10) and (2.11) for the three methods, and set $w_k=|\tilde{\beta}_k|^{-0.5}$ in (2.10) and $\mu=0.5$ in (2.11) without any particular preference. In the five tables, the columns #Group and #Variable stand for average number of groups selected and average number of variables selected, respectively. The tables also present %Corr.G and %Corr.V, where %Corr.G gives the portion of occasions on which the model produced contains exactly the same groups as the underlying true model, while %Corr.V calculates the portion of occasions on which the model produced contains

exactly the same variables as the underlying true model. All the numbers in parenthesis are the corresponding standard errors based on 400 repetitions.

From Tables 2.1-2.5 we observe that all the three bi-level variable selection penalties perform well when applied to partial log-likelihood method. They consistently reduce model complexity and select correct groups and variables. More specifically, we observe

- In general, all the three bi-level variable selection methods perform satisfactorily, even when the dimension of covariates is relatively high. This indicates that applying those bi-level selection approaches will greatly alleviate model complexity.

- When significant covariates within each group is highly sparse (Cases 2 and 5), in general, AGB and CGB tend to select smaller number of covariates than GB does, and AGB and CGB achieve better results in terms of the percentage of selecting correct variables than that of GB. However, GB performs better than AGB and CGB in Cases 1 and 3 when the covariates in each group are either all zero or all nonzero, i.e. the nonsparsity is present at within group level.

- From the comparison of Cases 4 and 5, higher sparsity lead to higher probability of selecting important covariates, especially for AGB. Cases 4 and 5 have the same number of groups, number of covariates, and grouping structure. The true model of Case 5 is more sparse than that of Case 4, and all the three methods give better variable selection for Case 5 that Case 4. Similar phenomenon is observed when we compare Cases 1 and 2. This is possibly resulted from their enhanced individual level selection capability.

- When the three tuning parameter selection criteria AIC, BIC and GCV are compared, if number of groups and number of variables selected are concerned, GCV always produces slightly smaller sizes than AIC while BIC produces significantly larger sizes than both. If percentages of selecting correct groups and variables are

Table 2.1: Simulation results for Case 1.

r	Penalty	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	GB	AIC	2.00(0.472)	5.17(1.374)	0.934(0.159)	0.903(0.138)
		BIC	2.19(0.707)	5.67(1.758)	0.900(0.174)	0.900(0.144)
		GCV	2.00(0.459)	5.13(1.356)	0.933(0.159)	0.902(0.138)
	AGB	AIC	2.08(0.370)	4.72(1.016)	0.971(0.090)	0.895(0.079)
		BIC	2.26(0.706)	5.30(1.683)	0.920(0.135)	0.890(0.103)
		GCV	2.06(0.331)	4.68(0.988)	0.975(0.084)	0.895(0.078)
	CGB	AIC	2.09(0.560)	3.97(1.330)	0.934(0.154)	0.831(0.101)
		BIC	2.21(0.831)	4.54(2.067)	0.892(0.173)	0.831(0.123)
		GCV	2.06(0.529)	3.89(1.262)	0.936(0.152)	0.829(0.101)
0.3	GB	AIC	2.08(0.610)	5.05(1.642)	0.898(0.186)	0.869(0.153)
		BIC	2.33(0.850)	5.73(0.193)	0.852(0.193)	0.852(0.159)
		GCV	2.07(0.603)	5.04(1.625)	0.899(0.186)	0.868(0.153)
	AGB	AIC	2.19(0.551)	4.70(1.343)	0.944(0.116)	0.871(0.088)
		BIC	2.45(0.871)	5.44(2.095)	0.879(0.164)	0.856(0.117)
		GCV	2.18(0.540)	4.67(1.308)	0.946(0.115)	0.871(0.088)
	CGB	AIC	2.18(0.675)	3.87(1.427)	0.900(0.172)	0.804(0.108)
		BIC	2.34(0.949)	4.45(2.225)	0.846(0.191)	0.796(0.122)
		GCV	2.18(0.660)	3.82(1.443)	0.901(0.170)	0.802(0.107)
0.5	GB	AIC	2.07(0.625)	4.91(1.594)	0.899(0.179)	0.870(0.141)
		BIC	2.34(0.866)	5.65(2.205)	0.852(0.192)	0.851(0.154)
		GCV	2.06(0.624)	4.88(1.592)	0.900(0.179)	0.870(0.140)
	AGB	AIC	2.18(0.575)	4.56(1.325)	0.934(0.133)	0.863(0.093)
		BIC	2.55(0.887)	5.59(2.085)	0.860(0.170)	0.844(0.119)
		GCV	2.16(0.561)	4.52(1.301)	0.934(0.133)	0.862(0.093)
	CGB	AIC	2.18(0.720)	3.77(1.560)	0.886(0.183)	0.793(0.106)
		BIC	2.41(1.004)	4.56(2.452)	0.829(0.196)	0.785(0.125)
		GCV	2.16(0.698)	3.71(1.486)	0.888(1.486)	0.792(0.107)
0.7	GB	AIC	2.17(0.762)	5.04(1.877)	0.871(0.197)	0.844(0.157)
		BIC	2.49(0.981)	5.93(2.455)	0.814(0.205)	0.823(0.164)
		GCV	2.15(0.741)	5.00(1.832)	0.872(0.196)	0.845(0.157)
	AGB	AIC	2.37(0.742)	4.82(1.540)	0.898(0.156)	0.840(0.108)
		BIC	2.76(1.019)	5.92(2.500)	0.822(0.190)	0.812(0.137)
		GCV	2.35(0.723)	4.77(1.492)	0.902(0.153)	0.840(0.107)
	CGB	AIC	2.30(0.852)	3.87(1.730)	0.860(0.194)	0.776(0.114)
		BIC	2.59(1.096)	4.82(2.595)	0.803(0.205)	0.766(0.129)
		GCV	2.26(0.822)	3.79(1.650)	0.863(0.193)	0.776(0.113)
1	GB	AIC	2.28(0.779)	5.17(1.825)	0.862(0.186)	0.839(0.142)
		BIC	2.76(1.078)	6.43(2.687)	0.775(0.206)	0.789(0.162)
		GCV	2.27(0.775)	5.13(1.795)	0.863(0.186)	0.840(0.142)
	AGB	AIC	2.44(0.805)	4.91(1.790)	0.880(0.157)	0.826(0.111)
		BIC	3.00(1.075)	6.34(2.645)	0.772(0.195)	0.778(0.139)
		GCV	2.39(0.771)	4.79(1.627)	0.886(0.153)	0.828(0.108)
	CGB	AIC	2.55(0.903)	4.67(2.104)	0.842(0.193)	0.778(0.124)
		BIC	2.94(1.118)	5.91(2.906)	0.772(0.203)	0.749(0.144)
		GCV	2.53(0.904)	4.62(2.102)	0.846(0.195)	0.779(0.123)
True model			2	6	1	1

Table 2.2: Simulation results for Case 2.

r	Penalty	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	GB	AIC	1.56(0.666)	2.11(1.376)	0.859(0.130)	0.915(0.063)
		BIC	1.38(0.656)	1.86(1.464)	0.823(0.115)	0.908(0.063)
		GCV	1.54(0.648)	2.08(1.328)	0.860(0.130)	0.917(0.061)
	AGB	AIC	1.50(0.675)	1.73(1.078)	0.853(0.122)	0.933(0.056)
		BIC	1.22(0.556)	1.38(0.965)	0.810(0.095)	0.923(0.047)
		GCV	1.49(0.664)	1.71(1.032)	0.852(0.123)	0.933(0.051)
	CGB	AIC	1.50(0.680)	1.59(0.851)	0.850(0.124)	0.937(0.053)
		BIC	1.21(0.573)	1.28(0.767)	0.802(0.097)	0.925(0.043)
		GCV	1.49(0.679)	1.59(0.851)	0.849(0.123)	0.937(0.053)
0.3	GB	AIC	1.58(0.667)	2.22(1.383)	0.850(0.146)	0.902(0.074)
		BIC	1.34(0.742)	1.98(1.717)	0.805(0.129)	0.892(0.087)
		GCV	1.57(0.665)	2.20(1.377)	0.849(0.146)	0.902(0.074)
	AGB	AIC	1.49(0.649)	1.80(1.115)	0.845(0.137)	0.919(0.067)
		BIC	1.26(0.670)	1.55(1.440)	0.797(0.111)	0.906(0.080)
		GCV	1.48(0.641)	1.78(1.083)	0.846(0.137)	0.920(0.065)
	CGB	AIC	1.54(0.681)	1.69(0.900)	0.838(0.149)	0.927(0.060)
		BIC	1.26(0.712)	1.46(1.452)	0.786(0.113)	0.908(0.082)
		GCV	1.54(0.674)	1.69(0.888)	0.838(0.148)	0.927(0.060)
0.5	GB	AIC	1.54(0.707)	2.21(1.467)	0.827(0.156)	0.891(0.077)
		BIC	1.44(0.808)	2.10(1.910)	0.790(0.150)	0.880(0.098)
		GCV	1.53(0.707)	2.20(1.465)	0.827(0.156)	0.891(0.078)
	AGB	AIC	1.52(0.732)	1.88(1.224)	0.824(0.144)	0.909(0.066)
		BIC	1.31(0.718)	1.64(1.465)	0.782(0.132)	0.899(0.078)
		GCV	1.52(0.729)	1.88(1.215)	0.824(0.145)	0.910(0.066)
	CGB	AIC	1.52(0.782)	1.71(1.095)	0.808(0.151)	0.911(0.069)
		BIC	1.25(0.662)	1.41(1.107)	0.773(0.136)	0.905(0.069)
		GCV	1.51(0.776)	1.69(1.078)	0.810(0.151)	0.911(0.069)
0.7	GB	AIC	1.70(0.881)	2.19(1.510)	0.813(0.160)	0.893(0.086)
		BIC	1.45(0.995)	1.91(2.004)	0.760(0.139)	0.881(0.109)
		GCV	1.70(0.880)	2.18(1.509)	0.812(0.162)	0.893(0.086)
	AGB	AIC	1.70(0.881)	2.19(1.510)	0.813(0.160)	0.893(0.086)
		BIC	1.45(0.995)	1.91(2.004)	0.760(0.139)	0.881(0.109)
		GCV	1.70(0.880)	2.18(1.509)	0.812(0.162)	0.893(0.086)
	CGB	AIC	1.70(0.925)	2.00(1.387)	0.801(0.164)	0.898(0.082)
		BIC	1.38(0.950)	1.70(1.752)	0.758(0.134)	0.886(0.097)
		GCV	1.69(0.918)	1.97(1.359)	0.800(0.165)	0.899(0.080)
1	GB	AIC	1.94(0.952)	3.12(2.030)	0.783(0.176)	0.834(0.111)
		BIC	1.99(1.224)	3.43(2.978)	0.730(0.167)	0.802(0.153)
		GCV	1.93(0.946)	3.09(2.013)	0.783(0.177)	0.835(0.110)
	AGB	AIC	1.93(0.986)	2.62(1.676)	0.775(0.172)	0.861(0.097)
		BIC	1.89(1.271)	2.85(2.707)	0.721(0.166)	0.829(0.141)
		GCV	1.92(0.982)	2.61(1.675)	0.776(0.172)	0.861(0.097)
	CGB	AIC	1.97(1.058)	2.50(1.755)	0.778(0.180)	0.865(0.100)
		BIC	2.02(1.442)	2.95(2.976)	0.700(0.181)	0.817(0.155)
		GCV	1.97(1.058)	2.50(1.755)	0.778(0.180)	0.865(0.100)
True model			2	2	1	1

concerned, all the three perform competitively while AIC and GCV tend to produce similar results that are slightly better than those of BIC.

- In general, the three proposed bi-level variable selection approaches perform better when the transformation index in the semiparametric transformation model is close to zero (the Cox PH model). One possible reason for this is that the higher transformation index means higher complexity brought into the model compared to the Cox PH model. It influences the objective function for obtaining penalized estimators.

Combining these observations we conclude that for the semiparametric transformation models (2.2), adaptive group bridge and composite group bridge penalized estimators give better overall performance than group bridge penalized estimator, especially when the sparsity level within each group is high which agrees with the fact that GB doesn't have individual level selection consistency. Among the three tuning parameter selection criteria, AIC and GCV performs competitively and better than BIC.

2.6 Real Data Applications

In this section we demonstrate the use of proposed bi-level variable selection methods by analyzing two well-known datasets, i.e., the primary biliary cirrhosis (PBC) data and the breast cancer data, described below.

2.6.1 Primary Biliary Cirrhosis (PBC) Data

PBC is a fatal chronic liver disease. An investigation of PBC was conducted between 1974 and 1984 in the Mayo Clinic trials study. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the dataset participated in the randomized trial and most of which complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost

Table 2.3: Simulation results for Case 3.

r	Penalty	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	GB	AIC	2.67(0.623)	8.76(2.074)	0.947(0.057)	0.911(0.050)
		BIC	3.96(1.450)	12.69(3.560)	0.889(0.138)	0.912(0.082)
		GCV	2.61(0.595)	8.55(2.045)	0.946(0.556)	0.907(0.050)
	AGB	AIC	2.93(0.722)	8.55(2.012)	0.952(0.070)	0.901(0.050)
		BIC	4.34(1.596)	12.36(3.319)	0.857(0.155)	0.890(0.081)
		GCV	2.89(0.700)	8.46(2.001)	0.953(0.068)	0.900(0.050)
	CGB	AIC	2.97(0.810)	7.37(2.089)	0.942(0.078)	0.869(0.052)
		BIC	4.48(1.653)	11.16(3.389)	0.837(0.160)	0.865(0.076)
		GCV	2.91(0.772)	7.21(2.053)	0.942(0.075)	0.867(0.052)
0.3	GB	AIC	2.74(0.777)	8.61(2.325)	0.934(0.070)	0.898(0.050)
		BIC	4.32(1.738)	13.24(4.702)	0.841(0.162)	0.874(0.098)
		GCV	2.67(0.746)	8.39(2.231)	0.934(0.068)	0.895(0.049)
	AGB	AIC	3.01(0.890)	8.20(2.204)	0.933(0.081)	0.881(0.050)
		BIC	4.85(1.879)	13.19(4.506)	0.797(0.178)	0.850(0.095)
		GCV	2.96(0.856)	8.06(2.134)	0.933(0.079)	0.880(0.050)
	CGB	AIC	3.14(1.028)	7.05(2.311)	0.919(0.097)	0.848(0.050)
		BIC	4.95(1.939)	11.60(4.127)	0.783(0.183)	0.834(0.081)
		GCV	3.07(0.959)	6.91(2.201)	0.923(0.090)	0.848(0.050)
0.5	GB	AIC	2.85(0.913)	8.57(2.426)	0.913(0.080)	0.884(0.057)
		BIC	4.95(1.875)	14.70(5.150)	0.780(0.177)	0.837(0.113)
		GCV	2.80(0.870)	8.42(2.354)	0.914(0.078)	0.882(0.057)
	AGB	AIC	3.34(1.176)	8.68(2.614)	0.900(0.101)	0.868(0.052)
		BIC	5.29(1.874)	13.82(4.612)	0.756(0.177)	0.822(0.099)
		GCV	3.28(1.153)	8.50(2.525)	0.904(0.100)	0.868(0.052)
	CGB	AIC	3.43(1.288)	7.24(2.556)	0.886(0.117)	0.832(0.055)
		BIC	5.59(1.939)	12.56(4.399)	0.719(0.180)	0.798(0.086)
		GCV	3.30(1.219)	6.95(2.435)	0.893(0.113)	0.831(0.054)
0.7	GB	AIC	3.28(1.240)	9.37(2.967)	0.886(0.108)	0.872(0.064)
		BIC	5.34(1.832)	15.47(5.269)	0.735(0.172)	0.806(0.114)
		GCV	3.17(1.189)	9.08(2.814)	0.888(0.105)	0.872(0.064)
	AGB	AIC	3.55(1.408)	8.92(3.167)	0.872(0.121)	0.851(0.064)
		BIC	5.91(1.929)	15.10(4.982)	0.690(0.185)	0.785(0.107)
		GCV	3.44(1.332)	8.61(2.969)	0.881(0.117)	0.853(0.062)
	CGB	AIC	3.79(1.506)	7.79(2.907)	0.851(0.133)	0.821(0.060)
		BIC	6.01(1.904)	13.38(4.538)	0.676(0.178)	0.777(0.091)
		GCV	3.69(1.427)	7.51(2.680)	0.858(0.128)	0.820(0.059)
1	GB	AIC	3.70(1.607)	10.13(4.143)	0.845(0.142)	0.842(0.080)
		BIC	6.24(1.930)	18.04(5.786)	0.651(0.180)	0.741(0.128)
		GCV	3.50(1.442)	9.56(3.580)	0.857(0.127)	0.848(0.072)
	AGB	AIC	4.23(1.682)	10.04(3.711)	0.816(0.157)	0.825(0.077)
		BIC	6.77(1.921)	17.32(5.402)	0.608(0.182)	0.728(0.111)
		GCV	4.07(1.571)	9.63(3.383)	0.827(0.147)	0.828(0.073)
	CGB	AIC	4.51(1.801)	8.92(3.651)	0.785(0.164)	0.794(0.072)
		BIC	6.83(1.973)	15.28(5.280)	0.596(0.182)	0.728(0.102)
		GCV	4.34(1.729)	8.52(3.423)	0.796(0.157)	0.797(0.069)
True model			3	12	1	1

Table 2.4: Simulation results for Case 4.

r	Penalty	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	GB	AIC	3.08(0.920)	11.59(3.687)	0.793(0.154)	0.796(0.061)
		BIC	3.71(1.049)	15.24(5.045)	0.828(0.156)	0.775(0.073)
		GCV	3.04(0.924)	11.44(3.736)	0.790(0.154)	0.796(0.061)
	AGB	AIC	2.25(0.525)	8.16(2.166)	0.682(0.085)	0.800(0.052)
		BIC	2.92(1.044)	11.53(4.655)	0.713(0.137)	0.777(0.072)
		GCV	2.24(0.526)	8.09(2.173)	0.680(0.084)	0.799(0.052)
	CGB	AIC	3.31(0.989)	7.23(3.019)	0.800(0.153)	0.805(0.056)
		BIC	4.05(1.282)	11.02(4.902)	0.784(0.158)	0.789(0.078)
		GCV	3.29(1.002)	7.13(3.035)	0.798(0.154)	0.805(0.058)
0.3	GB	AIC	3.04(0.981)	11.57(3.886)	0.768(0.159)	0.766(0.060)
		BIC	3.81(1.154)	15.81(5.445)	0.794(0.152)	0.733(0.085)
		GCV	3.00(0.985)	11.36(3.951)	0.763(0.162)	0.766(0.060)
	AGB	AIC	2.29(0.576)	8.18(2.397)	0.675(0.093)	0.775(0.054)
		BIC	3.20(1.166)	12.85(5.183)	0.694(0.139)	0.731(0.079)
		GCV	2.27(0.574)	8.08(2.425)	0.673(0.092)	0.775(0.054)
	CGB	AIC	3.36(1.137)	7.16(3.208)	0.763(0.158)	0.776(0.059)
		BIC	4.22(1.338)	11.72(5.060)	0.758(0.162)	0.751(0.078)
		GCV	3.27(1.109)	6.82(3.054)	0.761(0.158)	0.777(0.057)
0.5	GB	AIC	3.16(0.978)	11.75(3.910)	0.768(0.161)	0.753(0.067)
		BIC	4.15(1.096)	17.59(5.512)	0.781(0.155)	0.701(0.091)
		GCV	3.09(0.975)	11.46(3.899)	0.762(0.162)	0.755(0.066)
	AGB	AIC	2.41(0.676)	8.51(2.585)	0.671(0.101)	0.763(0.059)
		BIC	3.47(1.244)	14.40(5.676)	0.687(0.139)	0.701(0.086)
		GCV	2.37(0.652)	8.38(2.519)	0.669(0.099)	0.763(0.058)
	CGB	AIC	3.53(1.078)	7.67(3.087)	0.768(0.162)	0.768(0.063)
		BIC	4.48(1.224)	12.78(5.085)	0.748(0.148)	0.730(0.085)
		GCV	3.45(1.059)	7.41(3.032)	0.765(0.164)	0.768(0.062)
0.7	GB	AIC	3.30(1.050)	12.01(4.145)	0.752(0.165)	0.737(0.067)
		BIC	4.42(1.037)	18.92(5.580)	0.764(0.144)	0.665(0.097)
		GCV	3.25(1.033)	11.80(3.973)	0.750(0.168)	0.739(0.066)
	AGB	AIC	2.62(0.814)	9.39(3.057)	0.663(0.114)	0.745(0.068)
		BIC	3.73(1.186)	15.59(5.659)	0.670(0.141)	0.673(0.090)
		GCV	2.56(0.773)	9.11(2.897)	0.667(0.110)	0.749(0.066)
	CGB	AIC	3.73(1.134)	8.29(3.504)	0.753(0.163)	0.752(0.066)
		BIC	4.80(1.166)	14.20(5.354)	0.723(0.145)	0.698(0.088)
		GCV	3.66(1.117)	7.97(1.650)	0.755(0.163)	0.754(0.063)
1	GB	AIC	3.64(1.074)	13.26(4.317)	0.757(0.161)	0.710(0.072)
		BIC	4.76(1.143)	21.21(6.402)	0.732(0.145)	0.612(0.107)
		GCV	3.59(1.041)	12.93(4.090)	0.760(0.162)	0.714(0.070)
	AGB	AIC	2.89(0.923)	10.24(3.422)	0.659(0.136)	0.722(0.072)
		BIC	4.18(1.138)	18.00(5.437)	0.664(0.146)	0.632(0.090)
		GCV	2.81(0.891)	9.84(3.154)	0.656(0.134)	0.725(0.070)
	CGB	AIC	4.19(1.122)	9.82(3.688)	0.743(0.162)	0.726(0.067)
		BIC	5.16(1.059)	16.37(5.463)	0.701(0.135)	0.657(0.089)
		GCV	4.11(1.133)	9.35(3.487)	0.741(0.166)	0.730(0.066)
True model			4	12	1	1

Table 2.5: Simulation results for Case 5.

r	Penalty	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	GB	AIC	3.07(1.368)	8.65(5.16)	0.739(0.205)	0.816(0.069)
		BIC	3.33(1.593)	10.29(6.583)	0.743(0.207)	0.790(0.094)
		GCV	3.00(1.358)	8.41(5.081)	0.733(0.210)	0.818(0.067)
	AGB	AIC	2.48(1.043)	5.86(3.087)	0.682(0.154)	0.856(0.060)
		BIC	2.35(1.343)	5.93(4.843)	0.644(0.162)	0.842(0.070)
		GCV	2.46(1.030)	5.76(3.043)	0.681(0.154)	0.857(0.060)
	CGB	AIC	3.10(1.390)	4.77(3.033)	0.745(0.203)	0.871(0.052)
		BIC	2.99(1.803)	5.47(4.650)	0.683(0.205)	0.850(0.067)
		GCV	3.08(1.388)	4.71(3.012)	0.744(0.203)	0.872(0.052)
0.3	GB	AIC	2.86(1.346)	7.89(5.012)	0.700(0.205)	0.808(0.073)
		BIC	3.42(1.609)	11.24(7.143)	0.717(0.194)	0.757(0.112)
		GCV	2.82(1.338)	7.74(4.931)	0.696(0.204)	0.810(0.071)
	AGB	AIC	2.37(0.961)	6.07(3.226)	0.659(0.151)	0.839(0.061)
		BIC	2.54(1.445)	7.24(5.626)	0.637(0.159)	0.814(0.084)
		GCV	2.34(0.949)	5.97(3.191)	0.657(0.149)	0.841(0.060)
	CGB	AIC	3.08(1.428)	4.76(3.055)	0.706(0.180)	0.734(0.046)
		BIC	3.38(1.842)	7.03(5.508)	0.675(0.182)	0.711(0.060)
		GCV	3.03(1.404)	4.62(2.952)	0.703(0.181)	0.735(0.046)
0.5	GB	AIC	2.82(1.42)	7.95(5.216)	0.668(0.202)	0.798(0.076)
		BIC	3.53(1.717)	12.12(7.854)	0.700(0.192)	0.731(0.124)
		GCV	2.80(1.415)	7.78(5.104)	0.665(0.202)	0.801(0.073)
	AGB	AIC	2.51(1.062)	6.76(3.580)	0.643(0.153)	0.820(0.066)
		BIC	2.93(1.538)	9.18(6.353)	0.633(0.155)	0.775(0.099)
		GCV	2.46(1.044)	6.53(3.471)	0.640(0.154)	0.823(0.063)
	CGB	AIC	3.53(1.078)	7.67(3.087)	0.768(0.162)	0.788(0.060)
		BIC	4.48(1.224)	12.78(5.085)	0.748(0.148)	0.711(0.093)
		GCV	3.45(1.059)	7.41(3.032)	0.765(0.164)	0.791(0.058)
0.7	GB	AIC	3.04(1.384)	8.908(5.294)	0.666(0.188)	0.775(0.084)
		BIC	3.89(1.671)	14.17(8.126)	0.683(0.174)	0.685(0.136)
		GCV	3.00(1.374)	8.67(5.170)	0.665(0.187)	0.779(0.083)
	AGB	AIC	2.65(1.063)	7.37(3.686)	0.637(0.154)	0.801(0.070)
		BIC	3.26(1.532)	11.30(6.858)	0.640(0.160)	0.731(0.114)
		GCV	2.61(1.054)	7.18(3.544)	0.635(0.155)	0.803(0.067)
	CGB	AIC	3.47(1.449)	6.11(3.658)	0.678(0.183)	0.709(0.060)
		BIC	4.13(1.768)	10.00(6.096)	0.666(0.163)	0.673(0.079)
		GCV	3.39(1.442)	5.84(3.480)	0.675(0.183)	0.710(0.059)
1	GB	AIC	3.35(1.376)	10.14(5.377)	0.678(0.196)	0.749(0.091)
		BIC	4.50(1.494)	18.16(8.076)	0.683(0.174)	0.607(0.145)
		GCV	3.30(1.368)	9.93(5.188)	0.675(0.197)	0.753(0.087)
	AGB	AIC	2.93(1.131)	8.91(3.982)	0.625(0.163)	0.766(0.078)
		BIC	3.88(1.410)	14.89(6.642)	0.638(0.160)	0.659(0.117)
		GCV	2.90(1.120)	8.70(3.839)	0.625(0.162)	0.769(0.076)
	CGB	AIC	3.92(1.509)	8.01(4.307)	0.682(0.184)	0.783(0.075)
		BIC	4.82(1.519)	13.82(6.335)	0.671(0.152)	0.682(0.116)
		GCV	3.83(1.508)	7.63(4.082)	0.681(0.189)	0.789(0.071)
True model			4	6	1	1

to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. Among the 312 randomized patients with PBC, 152 were assigned to the drug D-penicillanmine, while the rest of them were assigned to a control group with placebo drug. 140 patients died from PBC disease during the ten years follow up study. Censoring was due to liver transplantation. It was of primary interest to study the effectiveness of D-penicillanmine in curing PBC disease. The PBC data have been investigated by many authors, e.g., Zhang and Lu (2007) and Huang et al. (2014).

To compare with the analysis of this data in Huang et al. (2014), we focus on the same observed 17 risk factors as in Huang et al. (2014) and their effects on the failure times of the 276 complete cases in the full model. The observed failure time is the number of days between registration and the earlier of death, liver transplantation, or study analysis time in July 1986. The dataset is available in R package 'survival'. The description of the covariates are summarized in Table 2.6. All these 17 risk factors are naturally categorized into 9 different groups according to different aspect measured. Under the transformation model assumption, we consider a grid of r over the range $[0,6]$ with increment of 0.1 for logarithmic transformation indexes. For each r the EM algorithm is implemented and the r that yields the largest log-likelihood is chosen as the best transformation model for fitting the data and to proceed for the variable selection procedure.

Table 2.7 presents the results for different combinations of the three group/variable selection methods GB, AGB and CGB and the two tuning parameter selection criteria AIC and BIC. The analysis results with GCV are very similar to those with AIC and thus omitted in the table. From the table we can see that all the methods suggest that groups G_1, G_4 and G_9 should be excluded from the reduced model. In group G_3, the variable *Spiders* should also be removed from the model. This table also indicates that BIC tends to select more variables than AIC which is consistent with our observation in simulation studies. According to the comparison of different methods in our simulation studies, one may use the results based on GB-AIC or CGB-AIC as the reduced model. Considering the fact that the coefficients of those extra significant covariates selected of GB-AIC

Table 2.6: PBC data: dictionary of covariates.

Group	Variable	Type	Definition
Age(G_1)	Z_1	C	Age(years)
Gender(G_2)	Z_2	D	Female gender(0 male and 1 female)
Phynotype(G_3)	Z_3	D	Ascites(0 absence)
	Z_4	D	Hepatomegaly (0 absence and 1 presence)
	Z_5	D	Spiders(0 absence and 1 presence)
	Z_6	D	Edemaoed(0 no edema, 0.5 untreated or successfully treated and 1 unsuccessfully treated)
Liver function damage(G_4)	Z_7	C	Alkaline phosphatase(units/litre)
	Z_8	C	Sgot(liver enzyme units/ml)
Excretory function of the liver(G_5)	Z_9	C	Serum bilirubin(mg/dl)
	Z_{10}	C	Serum cholesterol(mg/dl)
	Z_{11}	C	Triglyserides(mg/dl)
Liver function damage(G_6)	Z_{12}	C	Albumin(g/dl)
	Z_{13}	C	Prothrombin time(seconds)
Treatment(G_7)	Z_{14}	D	Penicillamine v.s. placebo (1 control and 2 treatment)
Reflection(G_8)	Z_{15}	D	Stage (histological stage of disease, graded 1,2,3, or 4)
	Z_{16}	C	Urine copper(mcg/24hrs)
Haematology(G_9)	Z_{17}	C	Platelets(per cubic ml/1000)

compared with CGB-AIC are very close to zero, one may use the simpler model selected by CGB-AIC for further study of the data.

2.6.2 Breast Cancer Data

A breast cancer data containing the metastasis-free survival times is available in the R package 'penalized'. It contains 144 independent lymph node positive breast cancer patients on metastasis-free survival, 5 clinical risk factors and expression measurements of 70 genes found to be prognostic for metastasis-free survival in an earlier study. These 75 variables, along with observed event time and event indicator, are listed in Table 2.8. This data has censoring rate 66%.

We follow the data pre-processing done in Huang et al. (2014). Specifically, we first reduce the model dimension to 50 by screening out the 25 most unimportant variables. Applying dynamic clustering, the remaining 50 variables could be classified into 8 groups. Based on this grouping structure, the estimation results of the three variable/group selection methods are displayed in

Table 2.7: PBC data: estimates of coefficients.

Group	Covariates	MLE	GB-AIC	GB-BIC	AGB-AIC	AGB-BIC	CGB-AIC	CGB-BIC
G_1	age	0.0064	0	0	0	0	0	0
G_2	gender	-0.5085	-0.4909	-0.5912	-0.8741	-0.8741	-1.1047	-0.8670
G_3	asc	0.1479	0	0.0441	0	0	0	0
	hep	0.0804	0	0.0080	0	0	0	0
	spid	0.0163	0	0	0	0	0	0
	edema	1.0139	0	0.5901	0.7355	0.7355	0.4971	0.7685
G_4	alk	0.0000	0	0	0	0	0	0
	sgot	0.0030	0	0	0	0	0	0
G_5	bil	0.0816	0.0839	0.0921	0.0873	0.0873	0.0949	0.0998
	chol	0.0005	0	0.0003	0	0	0	0
	trig	0.0014	-0.0010	-0.0018	0	0	0	0
G_6	alb	-0.7625	-1.0038	-0.9170	-1.0031	-1.0031	-1.3751	-1.2171
	prot	0.1282	0.0362	0.0342	0	0	0	0
G_7	trt	-0.1854	0	-0.2017	-0.1435	-0.1435	0	-0.1220
G_8	stage	0.4910	0.2657	0.3680	0.3311	0.3311	0.2965	0.3886
	cop	0.0031	0.0022	0.0025	0	0	0	0.0024
G_9	plat	0.0014	0	0	0	0	0	0

Table 2.8: Breast cancer data: dictionary of variables.

Variable	Type	Definition
time	C	Metastasis-free follow-up time (months)
event	D	Event indicator. ($1 =$ metastasis or death; $0 =$ censored)
Diam	D	Diameter of the tumor (1 for $\geq$ 2cm and 0 for$<$ 2cm)
N	D	Number of affected lymph nodes (1 for $1-3$ and 0 for ≥ 4)
ER	D	Estrogen receptor status (1 for positive and 0 for negative)
Grade	D	Grade of the tumor (three ordered levels)
Age	C	Patient age at diagnosis (years)
TSPYL5 . . . C20orf46	C	Gene expression measurements of 70 prognostic genes (Netherlands Cancer Institute 70 gene signature)

Figure 2.1 with subfigure (a) based on AIC and (b) based on BIC. The results for GCV is omitted for the same reason that it gives similar results to those of AIC. In the figure, each block represents a group. From Figure 2.1 we observe that all the three penalties select the same 7 important groups (except for the first group). In terms of number of important variables selected out of the total 50, group bridge selects the same 17 variables, by either AIC or BIC adaptive group bridge selects 29 variables by AIC and 31 by BIC, and composite group bridge selects 10 by AIC criterion and 11 by BIC criterion. Apparently BIC tends to retain slightly more variables than AIC.

2.7 Conclusion and Discussion

With the rapid development in modern evolving technology, increasing numbers of potential predictors could be easily gathered with low cost to avoid missing critical factors for statistical modelling. This promotes the fact that high dimensional data are becoming increasingly prevalent in various fields including, but not limited to, finance, clinical trials, neuroimaging and genomics. To improve model interpretability, stability and prediction accuracy, selecting a subset of significant variables plays a pivotal role in building statistical models. However, many traditional variable selection approaches, such as stepwise procedures and best subset selection, are no longer feasible due to the nature of high dimensionality and the expensive computation cost. In this chapter, we consider a large class of transformation models including the Cox PH model and proportional odds model as special cases. Literatures on variable selection for this class of models are only at individual variable level and our work fills the gap of variable selection for transformation models at both group and individual levels. Nevertheless, in this work we didn't consider the high-dimension situation where number of covariates is diverging or larger than sample size. Further investigation is needed in this direction.

In some clinical follow-up studies, a positive proportion of patients who respond favourably to treatment appear subsequently to be free of any signs or symptoms of a considered disease and may be considered as 'cured'. In Chapter 3, we will extend our work to incorporate a cure rate into

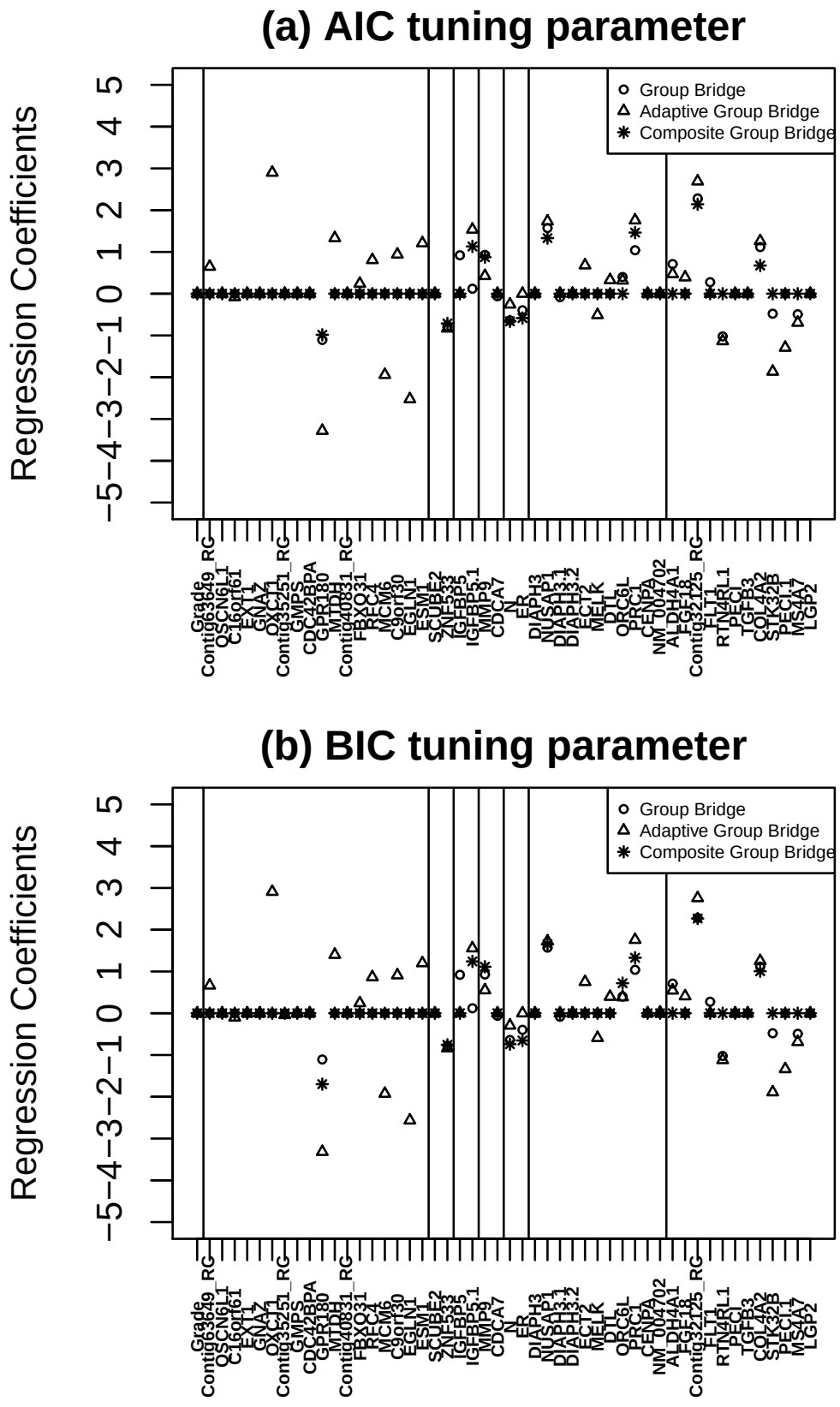


Figure 2.1: Plot of variable and group selection in the breast cancer data by using group bridge, adaptive group bridge and composite group bridge with AIC and BIC.

transformation models, resulted in mixture cure models. The unknown baseline cumulative hazard function could be estimated or piece-wise constants. Other bi-level variable selection approaches, such as group MCP, composite MCP and group SCAD could also be considered for mixture cure models.

2.8 Proof of Theorems

We define the counting process $N_i(s) = \delta_i I(T_i \leqslant s)$, where $s \in [0, \tau]$. Then $\ell_n(\beta)$ in (2.7) can be written as

$$\ell_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \beta^\top Z_i(s) \mathrm{d}N_i(s) - \sum_{i=1}^{n} \int_0^{\tau} \log \left[\sum_{k=1}^{n} I(T_k \geqslant s) \tilde{c}_k \exp\left\{ \beta^\top Z_k(s) \right\} \right] \mathrm{d}N_i(s),$$

where the weights $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_n)$ based on the MLEs $\tilde{\beta}$ and $\delta\tilde{\Lambda}(T)$ are

$$\begin{aligned} \tilde{c}_k &= G^{'} \left[\int_0^{\tau} I(s \leqslant T_k) \exp\left\{ \tilde{\beta}^\top Z_k(s) \right\} \mathrm{d}\tilde{\Lambda}(s) \right] - \int_0^{\tau} \frac{G^{''} \left[\int_0^s \exp\left\{ \tilde{\beta}^\top Z_k(t) \right\} \mathrm{d}\tilde{\Lambda}(t) \right]}{G^{'} \left[\int_0^s \exp\left\{ \tilde{\beta}^\top Z_k(t) \right\} \mathrm{d}\tilde{\Lambda}(t) \right]} \mathrm{d}N_k(s) \\ &\equiv c_k \left\{ T_k, Z_k(\cdot), \delta_k, \tilde{\beta}, \tilde{\Lambda} \right\}. \end{aligned}$$

To facilitate the proofs of the theorems, we first establish the following lemmas under *Conditions 1-4*.

LEMMA 2.8.1. *Denote by $U_n(\beta)$ the first-order derivative of $\ell_n(\beta)$ with respect to β. Then $n^{-1/2} U_n(\beta_0) = O_p(1)$.*

LEMMA 2.8.2. *Denote by $V_n(\beta)$ the second-order derivative of $-\ell_n(\beta)$ with respect to β. Then $V_n(\beta)/n$ converges uniformly to a positive definite matrix $V(\beta)$ almost surely, which does not depend on the data; as a result, $\ell_n(\beta)$ is a strictly concave function when n is large.*

Proof of Lemma 2.8.1. Write the true weight c_{0k} in the vector $c_0 = (c_{01}, \ldots, c_{0n})^\top$ as $c_{0k} = c_k \{T_k, Z_k(\cdot), \delta_k, \beta_0, \Lambda_0\}$, where β_0 and Λ_0 are the respective true value of β and Λ. Let P_n be the empirical measure, with P being the expectation. Then $U_n(\beta)$ at $\beta = \beta_0$ can be further expressed

as

$$
\begin{aligned}
& n^{-1/2}U_n(\beta_0) \\
=& n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\left[Z_i(s)-\frac{\sum_{k=1}^{n}I(T_k\geqslant s)Z_k(s)\tilde{c}_k\exp\left\{\beta_0^{\top}Z_k(s)\right\}}{\sum_{k=1}^{n}I(T_k\geqslant s)\tilde{c}_k\exp\left\{\beta_0^{\top}Z_k(s)\right\}}\right]\mathrm{d}N_i(s) \\
=& n^{1/2}(P_n-P)\int_0^{\tau}\left(Z(s)-\frac{E\left[I(T\geqslant s)Z(s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}\right)\mathrm{d}N(s) \\
& +n^{1/2}E\left\{\int_0^{\tau}\left(Z(s)-\frac{E\left[I(T\geqslant s)Z(s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}\right)\mathrm{d}N(s)\right\} \qquad \text{(A1)} \\
& -n^{1/2}\int_0^{\tau}\left(\frac{E\left[I(T\geqslant s)Z(s)\tilde{c}\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)\tilde{c}\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}-\frac{E\left[I(T\geqslant s)Z(s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}\right)P_n\mathrm{d}N(s) \qquad \text{(A2)} \\
& -n^{1/2}\int_0^{\tau}\left(\frac{P_n\left[I(T\geqslant s)Z(s)\tilde{c}\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{P_n\left[I(T\geqslant s)\tilde{c}\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}-\frac{E\left[I(T\geqslant s)Z(s)\tilde{c}\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)\tilde{c}\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}\right)P_n\mathrm{d}N(s). \qquad \text{(A3)}
\end{aligned}
$$

For term (A1), since the intensity for $N_k(s)$ is $I(T_k\geqslant s)\exp\left\{\beta_0^{\top}Z_k(s)\right\}\Lambda_0'(s)G'\left[\int_0^s\exp\left\{\beta_0^{\top}Z_k(t)\right\}\mathrm{d}\Lambda_0(t)\right]$, (A1) is equal to

$$
n^{1/2}E\left[\int_0^{\tau}\left\{Z(s)-\frac{E\left(I(T\geqslant s)Z(s)\exp\left\{\beta_0^{\top}Z(s)\right\}G'\left[\int_0^s\exp\left\{\beta_0^{\top}Z(t)\right\}\mathrm{d}\Lambda_0(t)\right]\right)}{E\left(I(T\geqslant s)\exp\left\{\beta_0^{\top}Z(s)\right\}G'\left[\int_0^t\exp\left\{\beta_0^{\top}Z(s)\right\}\mathrm{d}\Lambda_0(t)\right]\right)}\right\}\mathrm{d}N(s)\right]=0.
$$

For term (A2), by the mean value theorem, the integrand of (A2) is equal to

$$
\begin{aligned}
& -n^{1/2}\left[\nabla_{\beta}\frac{E\left[I(T\geqslant s)Z(s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}\left(\tilde{\beta}-\beta_0\right)\right. \\
& \qquad \left.+\nabla_{\Lambda}\frac{E\left[I(T\geqslant s)Z(s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}{E\left[I(T\geqslant s)c_0\exp\left\{\beta_0^{\top}Z(s)\right\}\right]}\left(\tilde{\Lambda}-\Lambda_0\right)+o_p(n^{-1/2})\right] \\
& \qquad =-n^{1/2}\mathscr{L}(s)\left(\tilde{\beta}-\beta_0,\tilde{\Lambda}-\Lambda_0\right)^{\top}+o_p(1),\text{say},
\end{aligned}
$$

where ∇_{β} denotes the derivative with respect to β and ∇_{Λ} denotes the Hadamard derivative with respect to Λ. So (A2) is equal to $-n^{1/2}(P_n-P)\int_0^{\tau}\mathscr{L}(s)\left(S_{\beta},S_{\Lambda}\right)P\mathrm{d}N(s)+o_p(1)$, where $\mathscr{L}$ is the linear operator, $o_p(1)$ converges to zero in probability uniformly for all $s\in[0,\tau]$, and S_{β} and S_{Λ} are efficient influence functions for β_0 and Λ_0, respectively.

For (A3), using the asymptotic consistency results for $\tilde{\beta}$ and $\tilde{\Lambda}$ from Zeng and Lin (2006), and the mean value theorem we have

$$\sup_{k=1,\ldots,n} |\tilde{c}_k - c_{0k}| \leqslant \sup_{k=1,\ldots,n} |\nabla_\beta c_k \{T_k, Z_k(\cdot), \delta_k, \beta^*, \Lambda^*\}| \|\tilde{\beta} - \beta_0\| + \sup_{k=1,\ldots,n} |\nabla_\Lambda c_k \{T_k, Z_k(\cdot), \delta_k, \beta^*, \Lambda^*\}| \|\tilde{\Lambda} - \Lambda_0\|_{l^\infty[0,\tau]} \xrightarrow{a.s.} 0$$

almost surely, where $\|\cdot\|_{l^\infty[0,\tau]}$ denotes the supremum norm over $[0,\tau]$, β^* is between $\tilde{\beta}$ and β_0, and Λ^* is between $\tilde{\Lambda}$ and Λ_0 uniformly in t. Based on Theorems 2.10.3 and 2.10.6 of van der Vaart and Wellner (1997), the weight $c\{T, Z(\cdot), \delta, \beta, \Lambda\}$ is a bounded Donsker class. Hence, by the Glivenko-Cantelli theorem, the integrand of (A3) is equal to

$$-n^{1/2} \left(\frac{(P_n - P)\left[I(T \geqslant s) Z(s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]}{E\left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]} - \frac{E\left[I(T \geqslant s) Z(s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]}{\left(E\left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]\right)^2} (P_n - P)\left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right] \right) + o_p(1);$$

so (A3) is equal to $-n^{1/2}(P_n - P)\int_0^\tau S_1 P \mathrm{d}N(s) + o_p(1)$, where

$$S_1 = \frac{\left[I(T \geqslant s) Z(s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]}{E\left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]} - \frac{E\left[I(T \geqslant s) Z(s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]}{\left(E\left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]\right)^2} \left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]$$

is the influence function and $o_p(1)$ converges to zero in probability uniformly over all $s \in [0,\tau]$.

Therefore, the normalized derivative $n^{-1/2} U_n(\beta_0)$ can be written as

$$n^{1/2}(P_n - P) \left\{ \int_0^\tau \left(Z(s) - \frac{E\left[I(T \geqslant s) Z(s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]}{E\left[I(T \geqslant s) c_0 \exp\left\{\beta_0^\top Z(s)\right\}\right]} \right) \mathrm{d}N(s) - \int_0^\tau \mathscr{L}(s)\left(S_\beta, S_\Lambda\right) P \mathrm{d}N(s) - \int_0^\tau S_1 P \mathrm{d}N(s) \right\} + o_p(1).$$

By the Donsker theorem, $n^{-1/2} U_n(\beta_0) = O_p(1)$. □

Proof of Lemma 2.8.2. We have

$$n^{-1} V_n(\beta) = \int_0^\tau \left\{ \frac{P_n\left[I(T \geqslant s) \tilde{c} Z^{\otimes 2}(s) \exp\{\beta^\top Z(s)\}\right]}{P_n\left[I(T \geqslant s) \tilde{c} \exp\{\beta^\top Z(s)\}\right]} - \left(\frac{P_n\left[I(T \geqslant s) \tilde{c} Z(s) \exp\{\beta^\top Z(s)\}\right]}{P_n\left[I(T \geqslant s) \tilde{c} \exp\{\beta^\top Z(s)\}\right]} \right)^{\otimes 2} \right\} P_n \mathrm{d}N(s),$$

where $Z^{\otimes 2} = ZZ^\top$. By the Glivenko-Cantelli theorem, the integrand converges uniformly to its asymptotic limit

$$A(s;\beta) \equiv \frac{E\left[I(T \geqslant s)c_0 Z^{\otimes 2}(s)\exp\{\beta^\top Z(s)\}\right]}{E\left[I(T \geqslant s)c_0\exp\{\beta^\top Z(s)\}\right]} - \left(\frac{E\left[I(T \geqslant s)c_0 Z(s)\exp\{\beta^\top Z(s)\}\right]}{E\left[I(T \geqslant s)c_0\exp\{\beta^\top Z(s)\}\right]}\right)^{\otimes 2}.$$

Define

$$V(\beta) = \int_0^\tau A(s;\beta)P\mathrm{d}N(s).$$

Then $\sup_\beta |n^{-1}V_n(\beta) - V(\beta)| \xrightarrow{a.s.} 0$.

To show that $V(\beta)$ is positive definite, we substitute the intensity for $N(s)$; then $V(\beta)$ can be written as

$$\begin{aligned}\int_0^\tau E\Bigg\{ &\left[Z(s) - \frac{E\left(I(T \geqslant s)Z(s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\right)}{E\left(I(T \geqslant s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\right)}\right]^{\otimes 2} \\ &\times I(T \geqslant s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\Bigg\} \\ &\times \frac{E\left\{I(T \geqslant s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\right\}}{E\left\{I(T \geqslant s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\right\}}\mathrm{d}\Lambda_0(s).\end{aligned}$$

Thus, $V(\beta)$ is semipositive definite. If $V(\beta)$ is not positive definite, then there exists a vector α such that $\alpha \neq 0$ and $\alpha^\top V(\beta)\alpha = 0$. This implies that for all $s \in [0,\tau]$,

$$\begin{aligned} 0 &= \alpha^\top Z(s) - \alpha^\top \frac{E\left(I(T \geqslant s)Z(s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\right)}{E\left(I(T \geqslant s)\exp\{\beta^\top Z(s)\}G'\left[\int_0^s \exp\{\beta_0^\top Z(t)\}\mathrm{d}\Lambda_0(t)\right]\right)} \\ &= \alpha_0(s) + \alpha^\top Z(s), \end{aligned}$$

which contradicts *Condition 2*. Therefore $V(\beta)$ is positive definite, and so $\ell_n(\beta)$ is strictly concave when n is large. □

Proof of Theorem 2.4.1. Consider the penalized objective function

$$Q_n(\beta) = \ell_n(\beta) - n\sum_{j=1}^J p_{\lambda_n}^{(j)}(|\beta_{A_j}|).$$

If the penalty term $\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\beta_{A_j}|)$ is strictly convex, it follows from Lemma 2.8.2 that $Q_n(\beta)$ is strictly concave a.s. when n is large. Thus, there exists a unique maximizer $\hat{\beta}_n$ of $Q_n(\beta)$ for large n. It is sufficient to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$\mathbb{P}\left(\sup_{\|u\|=C} Q_n(\beta_0 + n^{-1/2}u) < Q_n(\beta_0) \right) \geqslant 1 - \varepsilon. \tag{A4}$$

This implies that, with probability at least $1-\varepsilon$, there exists a local maximizer in the ball $\left\{\beta_0 + n^{-1/2}u : \|u\| \leqslant C\right\}$.

To prove (A4), we define

$$\begin{aligned} D_n(u) &\equiv n^{-1}\left[Q_n(\beta_0 + n^{-1/2}u) - Q_n(\beta_0)\right] \\ &= n^{-1}\left[\ell_n(\beta_0 + n^{-1/2}u) - \ell_n(\beta_0)\right] + \sum_{j=1}^{J}\left[p_{\lambda_n}^{(j)}(|\beta_0^{(j)}|) - p_{\lambda_n}^{(j)}(|\beta_0^{(j)} + n^{-1/2}u^{(j)}|)\right]. \end{aligned} \tag{A5}$$

For the first term in (A5), since for any β^* between $\tilde{\beta}$ and β_0 we have $\|\beta^* - \beta_0\| \leqslant \|\tilde{\beta} - \beta_0\|$, it follows from Lemma A2 that $\|n^{-1}V_n(\beta^*) - V(\beta_0)\| \leqslant \|n^{-1}V_n(\beta^*) - V(\beta^*)\| + \|V(\beta^*) - V(\beta_0)\| \to 0$ almost surely; that is, $n^{-1}V_n(\beta^*) = V(\beta_0) + o_p(1)$. Then, by Taylor expansion and Lemma A1, there exists a β^* between β_0 and $\beta_0 + n^{-1/2}u$ such that the first term equals

$$\begin{aligned} n^{-1}u^\top\left[n^{-1/2}U_n(\beta_0)\right] - (2n)^{-1}u^\top\left[n^{-1}V_n(\beta^*)\right]u \\ = n^{-1}O_p(1)\sum_{j=1}^{J}|u^{(j)}| - (2n)^{-1}u^\top\left[V(\beta_0) + o_p(1)\right]u. \end{aligned}$$

Next we consider the second term in (A5) under different penalty.

With group bridge penalty, the second term in (A5) is equal to $\lambda_n\sum_{j=1}^{J} c_j\|\beta_{0A_j}\|_1^\gamma - \lambda_n\sum_{j=1}^{J} c_j\|\beta_{A_j}\|_1^\gamma$. To find the upper bound of $D_n(u)$, it is sufficient to consider the case that $\|\beta_{0A_j}\|_1 \geqslant \|\beta_{A_j}\|_1$. By $b^\gamma - a^\gamma \leqslant 2(b-a)b^{\gamma-1}$ for $0 \leqslant a \leqslant b$ and the Cauchy-Schwartz inequality, it follows that

$$\begin{aligned} \sum_{j=1}^{J} c_j\|\beta_{0A_j}\|_1^\gamma - \sum_{j=1}^{J} c_j\|\beta_{A_j}\|_1^\gamma &\leqslant 2\sum_{j=1}^{J} c_j\|\beta_{0A_j}\|_1^{\gamma-1}\left(|A_j|\|\beta_{A_j} - \beta_{0A_j}\|_2^2\right)^{1/2} \\ &\leqslant 2\eta_n\left(\sum_{j=1}^{J}\|\beta_{A_j} - \beta_{0A_j}\|_2^2\right)^{1/2} \\ &= 2\eta_n\|\tilde{\beta} - \beta_0\|_2 \\ &= 2\eta_n n^{-1/2}\|u\|_2 \end{aligned}$$

with $\eta_n^2 = \sum_{j=1}^J c_j^2 \|\beta_{0A_j}\|_1^{2\gamma-2}|A_j|$. Since $n^{1/2}\lambda_n = O_p(1)$, we have

$$D_n(u) \leqslant n^{-1}O_p(1)\sum_{j=1}^J |u^{(j)}| - (2n)^{-1}u^\top \left\{V(\beta_0) + o_p(1)\right\} u + O_p(1)\eta_n n^{-1}\|u\|_2.$$

The first and third terms on the right hand side of above inequality are of order C/n and the second term is of order C^2/n. Therefore, when C is large the first and third terms are dominated by the second term. Thus, the inequality (A4) holds.

With adaptive group bridge penalty, the second term in (A5) is equal to $\lambda_n \sum_{j=1}^J c_j \|\beta_{0A_j}\|_{1,w}^\gamma - \lambda_n \sum_{j=1}^J c_j \|\beta_{A_j}\|_{1,w}^\gamma$. To find the upper bound of $D_n(u)$, it is sufficient to consider the case that $\|\beta_{0A_j}\|_1 \geqslant \|\beta_{A_j}\|_1$. By $b^\gamma - a^\gamma \leqslant 2(b-a)b^{\gamma-1}$ for $0 \leqslant a \leqslant b$ and the Cauchy-Schwartz inequality, it follows that

$$\begin{aligned}
\sum_{j=1}^J c_j \|\beta_{0A_j}\|_{1,w}^\gamma - \sum_{j=1}^J c_j \|\beta_{A_j}\|_{1,w}^\gamma &\leqslant 2\sum_{j=1}^J c_j \|\beta_{0A_j}\|_{1,w}^{\gamma-1} \sum_{k=1}^{d_j} w_k \left(|\beta_{0k}| - |\beta_{nk}|\right) \\
&\leqslant 2\sum_{j=1}^J c_j \|\beta_{0A_j}\|_{1,w}^{\gamma-1} \left(\sum_{k=1}^{d_j} w_k^2\right)^{1/2} \left[\sum_{k=1}^{d_j} \left(|\beta_{0k}| - |\beta_{nk}|\right)^2\right]^{1/2} \\
&\leqslant 2\sum_{j=1}^J c_j \|\beta_{0A_j}\|_{1,w}^{\gamma-1} \left(\sum_{k=1}^{d_j} w_k^2\right)^{1/2} \|\beta_{A_j} - \beta_{0A_j}\|_2 \\
&\leqslant 2\left[\sum_{j=1}^J c_j^2 \|\beta_{0A_j}\|_{1,w}^{2\gamma-2} \sum_{k=1}^{d_j} w_k^2\right]^{1/2} \left(\sum_{j=1}^J \|\beta_{A_j} - \beta_{0A_j}\|_2^2\right)^{1/2} \\
&\leqslant 2\eta_n^\star \|\tilde{\beta} - \beta_0\|_2 \\
&= 2\eta_n^\star n^{-1/2}\|u\|_2
\end{aligned}$$

with $\eta_n^{\star 2} = \sum_{j=1}^J c_j^2 \|\beta_{0A_j}\|_{1,w}^{2\gamma-2} \sum_{k=1}^{d_j} w_k^2$. Since $n^{1/2}\lambda_n = O_p(1)$, we have

$$D_n(u) \leqslant n^{-1}O_p(1)\sum_{j=1}^J |u^{(j)}| - (2n)^{-1}u^\top \left\{V(\beta_0) + o_p(1)\right\} u + 2O_p(1)\eta_n^\star n^{-1}\|u\|_2.$$

The first and third terms on the right hand side of above inequality are of order C/n and the second term is of order C^2/n. Therefore, when C is large the first and third terms are dominated by the second term. Thus, the inequality (A4) holds.

With composite group bridge penalty, the second term in (A5) is equal to $\lambda_n \sum_{j=1}^J c_j \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu\right)^\gamma - \lambda_n \sum_{j=1}^J c_j \left(\sum_{k=1}^{d_j} |\beta_{nk}|^\mu\right)^\gamma$. To find the upper bound of $D_n(u)$, it is sufficient to consider the case

where $\|\beta_{0A_j}\|_1 \geqslant \|\beta_{A_j}\|_1$. By $b^\gamma - a^\gamma \leqslant 2(b-a)b^{\gamma-1}$ for $0 \leqslant a \leqslant b$ and the Cauchy-Schwartz inequality, it follows that

$$
\begin{aligned}
&\sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^\gamma - \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_{nk}|^\mu \right)^\gamma \\
\leqslant\ & 2 \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^{\gamma-1} \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu - \sum_{k=1}^{d_j} |\beta_{nk}|^\mu \right) \\
\leqslant\ & 4 \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^{\gamma-1} \sum_{k=1}^{d_j} |\beta_{0k}|^{\mu-1} \left(|\beta_{0k}| - |\beta_{nk}| \right) \\
\leqslant\ & 4 \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^{\gamma-1} \left[\sum_{k=1}^{d_j} |\beta_{0k}|^{2\mu-2} \sum_{k=1}^{d_j} \left(|\beta_{0k}| - |\beta_{nk}| \right)^2 \right]^{1/2} \\
\leqslant\ & 4 \sum_{j=1}^{J} c_j \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^{\gamma-1} \left(\sum_{k=1}^{d_j} |\beta_{0k}|^{2\mu-2} \right)^{1/2} \|\beta_{A_j} - \beta_{0A_j}\|_2 \\
\leqslant\ & 4 \left[\sum_{j=1}^{J} c_j^2 \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^{2\gamma-2} \sum_{k=1}^{d_j} |\beta_{0k}|^{2\mu-2} \right]^{1/2} \left(\sum_{j=1}^{J} \|\beta_{A_j} - \beta_{0A_j}\|_2^2 \right)^{1/2} \\
\leqslant\ & 4\eta_n^\dagger \|\tilde{\beta} - \beta_0\|_2 \\
=\ & 4\eta_n^\dagger n^{-1/2} \|u\|_2
\end{aligned}
$$

with $\eta_n^{\dagger 2} = \sum_{j=1}^{J} c_j^2 \left(\sum_{k=1}^{d_j} |\beta_{0k}|^\mu \right)^{2\gamma-2} \sum_{k=1}^{d_j} |\beta_{0k}|^{2\mu-2}$. Since $n^{1/2}\lambda_n = O_p(1)$, we have

$$
D_n(u) \leqslant n^{-1} O_p(1) \sum_{j=1}^{J} |u^{(j)}| - (2n)^{-1} u^\top \left\{ V(\beta_0) + o_p(1) \right\} u + 2O_p(1)\eta_n^\dagger n^{-1} \|u\|_2.
$$

The first and third terms on the right hand side of above inequality are of order C/n and the second term is of order C^2/n. Therefore, when C is large the first and third terms are dominated by the second term. Thus, the inequality (A4) holds. This completes the proof of Theorem 1. □

Proof of Theorem 2.4.2. Recall that we assume the predictors form J groups. Without loss of generality, we assume that only the first J_1 groups are relevant, i.e., $\beta_{A_j} \neq 0$ for $j = 1, \ldots J_1$ and $\beta_{A_j} = 0$ for $j = J_1 + 1, \ldots, J$. We further assume in each of the rest J_1 groups, only a subset of the predictors are important. For each $A_j, j = 1, \ldots, J_1$, let $A_{1j} \subseteq A_j, \beta_{A_{1j}} \neq 0$, denote the set all nonzero β's in A_j and $\mathscr{A} = \cup_{j=1,\ldots,J_1} A_{1j}$. Note that the J groups may overlap with each other and

their union is allowed to be a proper subset of all the predictors.

First we prove the sparsity: $Pr(\hat{\beta}_{n,jk}=0)\to 1$ as $n\to\infty$ if $\beta_{0,jk}=0$. Using Taylor expansion, we have from (2.8) that

$$-\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{jk}} = \left(n^{-1}\frac{\partial \ell_n(\beta_0)}{\partial \beta_{jk}} + \sum_{j',k'} n^{-1}\frac{\partial^2 \ell_n(\beta^*)}{\partial \beta_{jk}\partial \beta_{j'k'}}\left(\hat{\beta}_{n,j'k'}-\beta_{0,j'k'}\right)\right) - \frac{\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|\right)}{\partial|\beta_{jk}|}\mathrm{sgn}(\hat{\beta}_{n,jk}), \tag{A6}$$

where β^* lies between $\hat{\beta}_n$ and β_0. By Theorem 3.2 in (Andersen and Gill, 1982) and the fact that $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimator, the first two terms on right-hand side of (A6) are both $O_p(n^{-1/2})$. Hence we have

$$-\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{jk}} = n^{-1/2}\left[O_p(1) - n^{1/2}\frac{\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|\right)}{\partial|\beta_{jk}|}\mathrm{sgn}(\hat{\beta}_{n,jk})\right].$$

If $n^{1/2}\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|\right)/\partial|\beta_{jk}|\to\infty$ with probability tending to 1 as $n\to\infty$, then for an arbitrary $\varepsilon>0$ and $j=J_1+1,\ldots,J$, when n is large we have

$$\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{jk}}<0,\quad -\varepsilon<\hat{\beta}_{n,jk}<0,\quad \frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{jk}}>0,\quad 0<\hat{\beta}_{n,jk}<\varepsilon.$$

Therefore, $Pr(\hat{\beta}_{n,jk}=0)\to 1$ as $n\to\infty$.

Second, we prove the asymptotic normality. Based on Theorem 1 and the sparsity property that we just have shown, there exists a $\sqrt{n}$ consistent estimator $\hat{\beta}_n=\left(\hat{\beta}_{n,\mathscr{A}},0\right)^\top$ that satisfies the equation

$$\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{jk}}=0,\quad (j,k)\in\mathscr{A}.$$

Hence we have

$$
\begin{aligned}
0 =& n^{-1}\frac{\partial \ell_n(\hat{\beta}_{n,\mathscr{A}},0)}{\partial \beta_{\mathscr{A}}} - \sum_{j=1}^{J_1}\left\{\frac{\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|,0\right)}{\partial|\beta_{\mathscr{A}}|}\operatorname{sgn}(\hat{\beta}_{n,\mathscr{A}})\right\} \\
=& n^{-1}\frac{\partial \ell_n(\beta_{0,\mathscr{A}},0)}{\partial \beta_{\mathscr{A}}} + n^{-1}\frac{\partial^2 \ell_n(\beta^*_{\mathscr{A}},0)}{\partial \beta^2_{\mathscr{A}}}\left(\hat{\beta}_{n,\mathscr{A}}-\beta_{0,\mathscr{A}}\right) \\
& - \sum_{j=1}^{J_1}\left\{\frac{\partial p_{\lambda_n}^{(j)}\left(|\beta_{0,j1}|,\ldots,|\beta_{0,jd_j}|,0\right)}{\partial|\beta_{\mathscr{A}}|}\operatorname{sgn}(\beta_{0,\mathscr{A}}) + \frac{\partial^2 p_{\lambda_n}^{(j)}\left(|\beta_{0,j1}|,\ldots,|\beta_{0,jd_j}|,0\right)}{+o_p(n^{-1/2})}(\hat{\beta}_{n,\mathscr{A}}-\beta_{0,\mathscr{A}})\right\} \\
& + o_p(n^{-1/2}),
\end{aligned}
$$

where $\beta^*_{\mathscr{A}}$ lies between $\hat{\beta}_{n,\mathscr{A}}$ and $\beta_{0,\mathscr{A}}$. If $b_n \to 0$ and $n^{1/2}\partial p_{\lambda_n}^{(j)}\left(|\beta_{0,j1}|,\ldots,|\beta_{0,jd_j}|,0\right)/\partial|\beta_{jk}| \to 0$ as $n\to\infty$, it follows by Theorem 3.2 in Andersen and Gill (1982) and Slutsky's Theorem that $n^{1/2}(\hat{\beta}_{\mathscr{A}}-\beta_{0,\mathscr{A}})$ converges in distribution to a normal random variable with mean zero and variance $V(\beta_{0,\mathscr{A}})^{-1}$.

Now we check the corresponding sparsity condition in Theorem 2.4.2 for different penalties. For group bridge penalty $p_{\lambda_n}^{(j)} = \lambda_n\left(|\beta_{j1}|+\ldots+|\beta_{jd_j}|\right)^{\gamma}$,

$$
\begin{aligned}
\frac{\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|\right)}{\partial|\beta_{jk}|} &= \lambda_n\gamma\left(|\hat{\beta}_{n,j1}|+\ldots+|\hat{\beta}_{n,jd_j}|\right)^{\gamma-1} \\
&= \lambda_n\gamma\left[|\beta_{0,j1}|+\ldots+|\beta_{0,jd_j}|+O_p(n^{-1/2})\right]^{\gamma-1} \\
&\geqslant \lambda_n\gamma\left[O_p(n^{-1/2})\right]^{\gamma-1} \\
&= \lambda_n\gamma\left(M_1 n^{-1/2}\right)^{\gamma-1}.
\end{aligned}
$$

Since $n^{1/2}\lambda_n = O_p(1)$ and $\gamma\in(0,1)$, we have $\sqrt{n}\lambda_n\gamma\left(M_1 n^{-1/2}\right)^{\gamma-1}\to\infty$.

For adaptive group bridge penalty $p_{\lambda_n}^{(j)} = \lambda_n\left(\frac{|\beta_{j1}|}{|\tilde{\beta}_{n,j1}|^{\mu}}+\ldots+\frac{|\beta_{jd_j}|}{|\tilde{\beta}_{n,jd_j}|^{\mu}}\right)^{\gamma}$,

$$\begin{aligned}
\frac{\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|\right)}{\partial|\beta_{jk}|} &= \lambda_n\gamma\left(\frac{|\hat{\beta}_{n,j1}|}{|\tilde{\beta}_{n,j1}|^{\mu}}+\ldots+\frac{|\hat{\beta}_{n,jd_j}|}{|\tilde{\beta}_{n,jd_j}|^{\mu}}\right)^{\gamma-1}\frac{1}{|\tilde{\beta}_{n,jk}|^{\mu}} \\
&= \lambda_n\gamma\frac{1}{\tilde{\beta}_{n,jk}^{\mu}}\left[\frac{|\beta_{0,j1}|}{|\tilde{\beta}_{n,j1}|^{\mu}}+\ldots+\frac{|\beta_{0,jd_j}|}{|\tilde{\beta}_{n,jd_j}|^{\mu}}+O_p(n^{-1/2})\right]^{\gamma-1} \\
&= \lambda_n\gamma\frac{1}{|\tilde{\beta}_{n,jk}|^{\mu}}\left[O_p(n^{-1/2})\right]^{\gamma-1} \\
&= \lambda_n\gamma\frac{1}{|\tilde{\beta}_{n,jk}|^{\mu}}\left(M_2n^{-1/2}\right)^{\gamma-1}.
\end{aligned}$$

Since $n^{1/2}\lambda_n = O_p(1), \gamma\in(0,1)$ and $\mu\in(0,1)$, we have $\sqrt{n}\lambda_n\gamma\frac{1}{\tilde{\beta}_{n,jk}^{\mu}}\left(M_2n^{-1/2}\right)^{\gamma-1}\to\infty$.

For composite group bridge penalty $p_{\lambda_n}^{(j)} = \lambda_n\left(|\beta_{j1}|^{\mu},\ldots,|\beta_{jd_j}|^{\mu}\right)^{\gamma}$,

$$\begin{aligned}
\frac{\partial p_{\lambda_n}^{(j)}\left(|\hat{\beta}_{n,j1}|,\ldots,|\hat{\beta}_{n,jd_j}|\right)}{\partial|\beta_{jk}|} &= \lambda_n\gamma\mu\left(|\hat{\beta}_{n,j1}|^{\mu}+\ldots+|\hat{\beta}_{n,jd_j}|^{\mu}\right)^{\gamma-1}|\hat{\beta}_{n,jk}|^{\mu-1} \\
&= \lambda_n\gamma\mu\left[|\beta_{0,j1}|^{\mu}+\ldots+|\beta_{0,jd_j}|^{\mu}+O_p(n^{-1/2})\right]^{\gamma-1}|\beta_{0,jk} \\
&\quad +O_p(n^{-1/2})|^{\mu-1} \\
&\geqslant \lambda_n\gamma\mu\left[O_p(n^{-1/2})\right]^{\gamma-1}\left[O_p(n^{-1/2})\right]^{\mu-1} \\
&= \lambda_n\gamma\mu\left[O_p(n^{-1/2})\right]^{\gamma+\mu-2} \\
&= \lambda_n\gamma\mu\left(M_3n^{-1/2}\right)^{\gamma+\mu-2}.
\end{aligned}$$

Since $n^{1/2}\lambda_n = O_p(1), \gamma\in(0,1)$ and $\mu\in(0,1)$, we have $\sqrt{n}\lambda_n\gamma\mu\left(M_3n^{-1/2}\right)^{\gamma+\mu-2}\to\infty$. □

Chapter 3

BI-LEVEL VARIABLE SELECTION IN SEMIPARAMETRIC TRANSFORMATION MIXTURE CURE MODELS WITH RIGHT-CENSORED DATA

Survival data with a proportion of cured subjects commonly occur in medical study. In this chapter, we investigate semiparametric transformation mixture cure models (STMCMs) to incorporate a cure fraction. This chapter is organized as follow: In Section 3.1 we present a brief review on mixture cure models; In Section 3.2, we propose to use a semiparametric transformation model for the latency part and a logistic regression model for the incidence part of the mixture cure model, and develop an EM-based algorithm for numerical calculation of the maximum likelihood estimation. In Section 3.3, we use the group bridge penalty to conduct bi-level variable selection for both parts in the mixture cure model. The corresponding computational algorithm is shown in Section 3.4. To demonstrate the implementation of the proposed procedures, simulation studies and real data examples are presented in Section 3.5 and Section 3.6 respectively. The conclusion and discussion of this work are presented in Section 3.7.

3.1 Introduction

In standard survival analysis, there is a general underlying assumption behind standard survival models, such as Cox proportional hazards (PH) model, proportional odds (PO) model or accelerated failure time (AFT) model, that all subjects would eventually experience the event of interest if the follow-up time is long enough. However, as the evolving medical technology becomes more advanced, it is more frequentLY encountered in some medical studies that a group of subjects may respond favourably to the treatment and will be immune from the event. Those risk-free subjects

are regarded as "cured" from the event and their survival times are considered as infinite. For such data, Kaplan-Meier (K-M) estimator of the survival function for the time to event levels off to a nonzero proportion referred as the cure rate. Thus, survival models that take a cure rate into account are more plausible for such data, which are named as cure models. The cure models broaden the scope of standard survival models in applications, with the capability of allowing the possibility that a portion of subjects will never experience the event of interest. Applications of cure models can be found in, but not limited to, biomedical sciences, engineering and economics (Ross and Xian, 1996).

A commonly used approach to incorporate the cure rate in survival models is by assuming that the underlying population is a mixture of susceptible and nonsusceptible subjects. A prominent application of mixure cure models is to model patients with long-term censored survival times that appear in many cancer clinical trials studies, such as Farewell (1986) which analysed a clinical study of breast cancer and Sy and Taylor (2000) and Peng et al. (2007) which studied tonsil cancer using mixture cure models. Under mixture cure rate models (Farewell, 1982), we can separately model the fraction of cured individuals (the incidence part) and the failure time distribution for susceptible ones (the latency part). We are interested in the estimation of both the proportion of subjects who are cured and the failure time distribution of uncured subjects. There is a wide variety of modeling approaches on mixture cure models, from fully parametric to completely nonparametric. To slack the restrictions imposed by parametric survival models, semiparametric mixture cure models are often of greater interest than parametric models since the assumptions of parametric models are hardly met in practice. Most of the literatures in this area focus on different modeling of the latency while keep the logistic regression form for the incidence part. Kuk and Chen (1992), Sy and Taylor (2000), and Peng and Dear (2000) studied a logistic/Cox PH mixture cure model, which assumes a logistic regression for the incidence part and a Cox PH model for the conditional survival function in the latency part with the baseline survival function totally unspecified. Later, Lu and Ying (2004) relaxed the Cox PH model by replacing it with a semiparametric linear trans-

formation model, a more general model with Cox PH model and PO model as special cases. As an extension of the work of Lu and Ying (2004), we use the even more general semiparametric transformation models proposed Zeng and Lin (2006) for the conditional survival function in the latency part and a logistic regression model for the incidence component.

In multivariate regression setting, it is critical to determine which risk factors affect the probability of being susceptible and/or the failure time to an the event. When risk factors are grouped, we are interested in selecting the groups and individual risk factors that have impact on survival probability and/or cure rate. Liu et al. (2012) first established a variable selection method based on a penalized likelihood for both parts in the logistic/Cox PH mixture cure model. Based on the complete-data likelihood for this model, they developed a penalized EM algorithm, which is equivalent to maximizing a logistic model with penalization and separately a Cox PH model with penalization. The LASSO (Tibshirani, 1996) and SCAD (Fan et al., 2004) penalty functions are employed in their work, without taking consideration of the inherent interconnections among predictors, and these methods perform variable selection at individual level only. However, grouping structure arises naturally in many real cases. For example, groups of medical measurements of patients may be taken in a clinic study; in gene expression analysis, genes belonging to the same biological pathway can be considered as a group; and in genetic association studies, genetic markers from the same gene or haplotype can be grouped. To incorporate the grouping structure, a series of group selection approaches has been developed. For instance, Yuan and Lin (2006) proposed group LASSO penalty; like the group LASSO, but imposing either a SCAD penalty, bridge penalty or MCP on the norm of each group respectively, Fan et al. (2004) developed group SCAD penalty, and Huang et al. (2009) introduced group bridge penalty and Breheny and Huang (2009) investigated group MCP method. For the purpose of bi-level variable selection, Breheny and Huang (2009) studied composite MCP approach and summarized a general hierarchical framework for constructing bi-level variable selection penalties, while Breheny (2015) recently proposed the group exponential LASSO penalty. To the best of our knowledge, there is no literature that ad-

dresses group selection or bi-level variable selection in semiparametric mixture cure models. To fill this gap, we study the bi-level selection in semiparametric transformation mixture cure models with right-censored data. Group bridge penalty is adopted for bi-level variable selection for both latency part and incidence part.

The aim of this work is to investigate the bi-level variable selection problem under semiparametric transformation mixture cure models in order to identify important variables and groups that contribute to cure proportion and survival function of the uncured subjects.

3.2 Model and Estimation Method

3.2.1 Semiparametric Mixture Cure Models

Let $\tilde{T}$ denote the failure time to the event of interest. The mixture cure model assumes the failure time $\tilde{T} = uT^* + (1-u)\infty$, where T^* denotes the failure time of susceptible subjects and u is a binary indicator of the susceptibility of event of interest. Here $u = 1$ means that a subject will eventually experience the event and $u = 0$ otherwise. Suppose that $\tilde{T}$ is subject to random right censoring with censoring time C, and assume that the censoring mechanism is noninformative. Denote the observed time $T = \min\{\tilde{T}, C\}$ and the censoring indicator $\delta = I(\tilde{T} \leqslant C)$, where $I(\cdot)$ is the indicator function. Due to the presence of right censoring, we never observe $\tilde{T}$ when it equals infinity. In fact, when $\delta = 1$ (uncensored observation), we know for sure that the individual is susceptible (uncured), whereas when $\delta = 0$ (censored observation) the subject could belong to either of the cured group or uncured roup, and the membership of which one is unknown. Let $X = (X_1, \ldots, X_p)^\top$ and $Z = (Z_1, \ldots, Z_q)^\top$ be the corresponding vector of covariates accounting for the survival outcome in the latency part and the cure rate, respectively. Assume that $\tilde{T}$ and C are conditionally independent given (X, Z), and X and Z may share some common variables. Let $S(t|X,Z)$ be the conditional survival function of $\tilde{T}$ for the entire population, $S_u(t|Z)$ be the survival function of T^* for the susceptible subjects, possibly depending on Z, $S_c(t)$ be the survival function of C for the cured subjects, and $f(t|X,Z)$, $f_u(t|Z)$ and $f_c(t)$ are the corresponding p.d.f.s of $\tilde{T}, T^*$

and C, respectively. Because the cured subjects will never experience the event of interest, their survival time is infinite. Therefore, for each finite value t, $f_c(t) = 0$ and $S_c(t) = 1$. Under such a notation, the survival function and p.d.f. of $\tilde{T}$ for the entire population are expressed as:

$$\begin{aligned} S(t|X,Z) &= \pi(X)S_u(t|Z) + [1-\pi(X)]S_c(t) = \pi(X)S_u(t|Z) + 1 - \pi(X), \\ f(t|X,Z) &= \pi(X)f_u(t|Z) + [1-\pi(X)]f_c(t) = \pi(X)f_u(t|Z). \end{aligned} \tag{3.1}$$

As t increases, $S(t|X,Z)$ goes to $1-\pi(X)$ and $1-\pi(X)$ is the probability of being cured given the covariate X. As mentioned above, $S_u(t|Z)$ and $\pi(X)$ are the "latency" part and "incidence" part respectively. Then $S_u(t|Z)$ can be also denoted as $S(t|u=1,Z)$, and $\pi(X)$ is equivalent to $P(u=1|X)$. Suppose that a random sample of n independent subjects is chosen. With right censoring, the observed data consist of $\{t_i, \delta_i, z_i, x_i\}$ for $i = 1, \ldots, n$, where t_i and δ_i denote the observed survival time and censoring indicator for the i^{th} subject, z_i and x_i are the observed values of the two covariate vectors. By (3.1), the likelihood function of the observed data from the mixture model is

$$\prod_{i=1}^{n} f(t_i|z_i,x_i)^{\delta_i} S(t_i|z_i,x_i)^{1-\delta_i} = \prod_{i=1}^{n} [\pi(x_i)f_u(t_i|z_i)]^{\delta_i} [\pi(x_i)S_u(t_i|z_i) + 1 - \pi(x_i)]^{1-\delta_i}.$$

If $u = (u_1, u_2, \ldots, u_n)$ is known, then $u_i\delta_i = \delta_i$ and thus the likelihood function for the complete data can be written as

$$\begin{aligned} &\prod_{i=1}^{n} [\pi(x_i)f_u(t_i|z_i)^{u_i}]^{\delta_i} \prod_{i=1}^{n} \{[\pi(x_i)S_u(t_i|z_i)]^{u_i}[1-\pi(x_i)]^{1-u_i}\}^{1-\delta_i} \\ &\qquad = \prod_{i=1}^{n} [1-\pi(x_i)]^{1-u_i} \pi(x_i)^{u_i} \prod_{i=1}^{n} h_u(t_i|z_i)^{\delta_i u_i} S_u(t_i|z_i)^{u_i}, \end{aligned}$$

where $h_u(\cdot)$ is the hazard function corresponding to $S_u(\cdot)$. The logarithm of the complete log-likelihood function is

$$\sum_{i=1}^{n} [(1-u_i)\log\{1-\pi(x_i)\} + u_i \log\{\pi(x_i)\}] + \sum_{i=1}^{n} [\delta_i u_i \log\{h_u(t_i|z_i)\} + u_i \log\{S_u(t_i|z_i)\}]. \tag{3.2}$$

A prominent advantage of the mixture cure models is that the proportion of cured subjects and the survival distribution of the remaining subjects are modeled separately since the incidence and latency parts interact in a multiplicative manner in the likelihood function, and the interpretation

of effects of x and z is straightforward. For the incidence term of the model, the covariates effects on the proportion of cured subjects is usually modeled by a logistic regression with a *logit* link function given by

$$\text{logit}(\pi(x)) = \gamma^\top x, \tag{3.3}$$

where γ is an unknown vector of parameters associated with the covariate vector x in order to predict the cure proportion $\pi(x)$. Other link functions can also be employed to model the probability of being susceptible, including the *probit* link

$$\Phi^{-1}(\pi(x)) = \gamma^\top x,$$

where $\Phi(\cdot)$ denotes the c.d.f. of a standard normal distribution, and the complementary *log-log* link

$$\log(-\log(1-\pi(x))) = \gamma^\top x.$$

In this work, we adopt the most commonly used logistic regression model (3.3) to depict the incident component. For the latency, we consider a family of semiparametric transformation models to formulate $S_u(t|z)$.

3.2.2 Semiparametric Transformation Models

We assume that $Z(\cdot) = (Z_1(\cdot), \ldots, Z_q(\cdot))^\top$ is a possibly time-varying covariate vector that are categorized into J groups. Assume the j^{th} group has q_j variables represented by $Z^{(j)}(\cdot) = (Z_{j1}(\cdot), \ldots, Z_{jq_j}(\cdot))^\top$. A general transformation model proposed by Zeng and Lin (2006) is adapted for the latency part which assumes that given covariates $Z(\cdot)$, the survival time T^* has cumulative hazard function is

$$\Lambda(t|Z(\cdot)) = G\left[\int_0^t \exp\{\beta^\top Z(s)\}\mathrm{d}\Lambda(s)\right], \tag{3.4}$$

where β is an unknown vector of parameters associated with the covariate vector $Z(\cdot)$, $\Lambda(\cdot)$ is an unspecified increasing function and G is a thrice continuously differentiable and strictly increasing

function with $G(0)=0$, $G'(0)>0$ and $G(\infty)=\infty$. The transformation function G is assumed to be of the form $G(x)=-\log\int_0^\infty \exp(-x\zeta)\phi(\zeta)\mathrm{d}\zeta$, where $\phi(\cdot)$ is a known density function over $[0,\infty)$. Any transformation induced by some density ϕ with support in $[0,\infty)$ can be adopted. A commonly used choice of $\phi(\cdot)$ is the gamma density with unit mean and variance ρ. Here we follow Zeng and Lin (2007b) to consider a logarithmic transformation class of functions

$$G(x)=\begin{cases}\dfrac{\log(1+rx)}{r}, & r>0\\ x, & r=0.\end{cases} \tag{3.5}$$

When $G(x)=x$, i.e. $r=0$ in (3.5), it leads to the Cox PH model. When $G(x)=\log(1+x)$, i.e. $r=1$ in (3.5), it reduces to the PO model.

The likelihood function of model (3.4) based on the observed data $\{t_i,\delta_i,z_i(\cdot)\}$ for $i=1,\ldots,n$ is

$$\begin{aligned}L_{latency}=\prod_{i=1}^{n}&\left(\Lambda'(t_i)\exp\left\{\beta^\top z_i(\cdot)\right\}G'\left[\int_0^{t_i}\exp\left\{\beta^\top z_i(s)\right\}\mathrm{d}\Lambda(s)\right]\right)^{\delta_i}\\ &\times\exp\left(-G\left[\int_0^{t_i}\exp\left\{\beta^\top z_i(s)\right\}\mathrm{d}\Lambda(s)\right]\right),\end{aligned} \tag{3.6}$$

where $\Lambda'(t_i)$ is the derivative of the cumulative hazard function Λ at time t_i. The likelihood in (3.6) includes both β and the infinite-dimensional parameter Λ, and it may not be concave in these parameters. Due to the transformation G, the partial likelihood of β does not exist. Thus, it is no longer feasible to directly apply the penalized methods in Fan and Li (2001) or Zhang and Lu (2007) for variable selection.

As a remedy of this issue, we implement the method proposed by Zeng and Lin (2007b) which has a latent variable ζ induced, which can be interpreted as the frailty in survival analysis. Then model (3.4) is equivalent to

$$\Lambda(t|Z(\cdot),\zeta)=\zeta\int_0^t\exp\left\{\beta^\top Z(s)\right\}\mathrm{d}\Lambda(s). \tag{3.7}$$

because

$$\begin{aligned} S(T|Z(\cdot)) &= \Pr(T > t|Z(\cdot)) \\ &= E\left[\Pr\left(T > t|Z(\cdot),\zeta\right)|Z(\cdot)\right] \\ &= E\left(\exp\left[-\zeta\int_0^t \exp\beta^\top Z(s)\mathrm{d}\Lambda(s)\right]|Z(\cdot)\right) \\ &= \int_0^\infty \exp\left[-\zeta\int_0^t \exp\beta^\top Z(s)\mathrm{d}\Lambda(s)\right]\phi(\zeta)\mathrm{d}\zeta \\ &= \exp\left(-G\left[\int_0^t \exp\beta^\top Z(s)\mathrm{d}\Lambda(s)\right]\right). \end{aligned}$$

For right-censored survival data, nonparametric maximum likelihood estimation (NPMLE) could be computed by treating the unknown baseline function as an infinite-dimensional parameter. In model (3.4), the cumulative baseline hazard function is an unknown infinite-dimensional parameter. The likelihood function (3.6) with $\Lambda^{'}(t_i)$ replaced by $\Delta\Lambda(t_i)$, needs to be maximized over both β and $\Delta\Lambda(t_i)$, the jump size of cumulative hazard function Λ at each observed failure time t_i, $i = 1,\ldots,n$. Zeng and Lin (2007b) showed that the NPMLEs of β and $\Delta\Lambda(t_i)$ are consistent, asymptotically normal and asymptotically efficient. The maximization process is computationally intensive, especially when number of event is large. A number of algorithms are available for the computation of NPMLEs of β and Λ_i's, such as the EM algorithm was employed by Zeng and Lin (2007b).

Thus, based on the complete data $\{t_i,\delta_i,z_i(\cdot),\zeta_i\}$ for $i = 1,\ldots,n$, the likelihood function of the latency (3.6) becomes

$$\prod_{i=1}^{n}\left(\left[\zeta_i\Delta\Lambda(t_i)\exp\left\{\beta^\top z_i(t_i)\right\}\right]^{\delta_i}\times\exp\left[-\zeta_i\int_0^{t_i}\exp\left\{\beta^\top z_i(s)\right\}\mathrm{d}\Lambda(s)\right]\times\phi(\zeta_i)\right)^{u_i}. \tag{3.8}$$

Then, a weighted version of partial log-likelihood function of β can be written as

$$\sum_{i=1}^{n}u_i\left[\delta_i\left\{\log\Delta\Lambda(t_i)+\beta^\top z_i(t_i)\right\}-\zeta_i\sum_{t_k\leqslant t_i}\exp\left\{\beta^\top z_i(t_k)\right\}\Delta\Lambda(t_k)\right]. \tag{3.9}$$

3.2.3 Estimation of Semiparametric Transformation Mixture Cure Model (STMCMs)

With model assumptions 3.1, 3.3 and 3.7 and the complete data $\{(t_i, \delta_i, z_i(\cdot), x_i, u_i, \zeta_i), i = 1, \ldots, n\}$, the log-likelihood of the semiparametric transformation mixture cure model is

$$\begin{aligned} \ell_c(\gamma, \beta, \Delta\Lambda(\cdot)) &= \sum_{i=1}^{n} u_i \log\left[\frac{\exp(\gamma^\top x_i)}{1+\exp(\gamma^\top x_i)}\right] + (1-u_i)\log\left[\frac{1}{1+\exp(\gamma^\top x_i)}\right] \\ &\quad + \sum_{i=1}^{n} u_i \left[\delta_i \left\{\log\Delta\Lambda(t_i) + \beta^\top z_i(t_i)\right\} - \zeta_i \sum_{k=1}^{n} I(t_k \leqslant t_i)\exp\left\{\beta^\top z_i(t_k)\right\}\Delta\Lambda(t_k)\right] \\ &= \ell_{c_1}(\gamma) + \ell_{c_2}(\beta, \Delta\Lambda(\cdot)). \end{aligned} \tag{3.10}$$

It involves two latent variables, u_i and ζ_i. The first one is the unobserved indicator of susceptibility and the later could be interpreted as the frailty in survival models. An embedded EM algorithm is applied to fit the semiparametric transformation mixture cure model based on the complete data.

Let $\mathscr{O}$ denote the observed data, i.e. $\mathscr{O} = \{(t_i, \delta_i, z_i(\cdot), x_i) : i = 1, \ldots, n\}$. The E-step in the EM algorithm computes the expectation of the complete log-likelihood with respect to u_i and ζ_i, conditional on both the current estimates of all the parameters and the observed data. The M-step updates the estimates by maximizing the expected log-likelihood with respect to each parameter. Specifically, it updates at the $(m+1)^{\text{th}}$ iteration the posterior expectation of latent variables, given $w_i^{(m)}$ and other parameters' values at the m^{th} iteration, by

$$\begin{aligned} v_i^{(m+1)} &= E\left\{u_i | \mathscr{O}, w_i^{(m)}, \tilde{\gamma}^{(m)}, \tilde{\beta}^{(m)}, \tilde{\Delta\Lambda}^{(m)}(t)\right\} \\ &= \delta_i + (1-\delta_i)\frac{\pi(x_i)S_u(t_i|z_i)}{1-\pi(x_i)+\pi(x_i)S_u(t_i|z_i)}. \end{aligned} \tag{3.11}$$

It is obvious that if $\delta_i = 1$, then $v_i^{(m+1)} = 1$ and if $\delta_i = 0$, then $v_i^{(m+1)} = \frac{\pi(x_i)S_u(t_i|z_i)}{1-\pi(x_i)+\pi(x_i)S_u(t_i|z_i)}$, the conditional probability of the i^{th} subject remaining susceptible. The conditional expectation of ζ_i

in the $(m+1)^{\text{th}}$ iteration, $w_i^{(m+1)}$, can be calculated by

$$\begin{aligned} w_i^{(m+1)} &= E\left\{\zeta_i|\mathscr{O}, v_i^{(m+1)}, \tilde{\gamma}^{(m)}, \tilde{\beta}^{(m)}, \tilde{\Delta\Lambda}^{(m)}(t)\right\} \\ &= \begin{cases} G'\left[\sum_{k=1}^{n} I(t_k \leqslant t_i)\exp\left\{\beta^{(m)^\top} z_i(t_k) + \log v_i^{(m+1)}\right\}\Delta\Lambda(t_k)^{(m)}\right], & \delta_i = 0, \\ -\dfrac{G''\left[\sum_{k=1}^{n} I(t_k \leqslant t_i)\exp\left\{\beta^{(m)^\top} z_i(t_k) + \log v_i^{(m+1)}\right\}\Delta\Lambda(t_k)^{(m)}\right]}{G'\left[\sum_{k=1}^{n} I(t_k \leqslant t_i)\exp\left\{\beta^{(m)^\top} z_i(t_k) + \log v_i^{(m+1)}\right\}\Delta\Lambda(t_k)^{(m)}\right]} \\ +G'\left[\sum_{k=1}^{n} I(t_k \leqslant t_i)\exp\left\{\beta^{(m)^\top} z_i(t_k) + \log v_i^{(m+1)}\right\}\Delta\Lambda(t_k)^{(m)}\right], & \delta_i = 1. \end{cases} \end{aligned} \tag{3.12}$$

Since $\delta_i \log(v_i^{(m+1)}) = 0$ and $\delta_i v_i^{(m+1)} = \delta_i$, the expectations of $\ell_{c_1}(\gamma)$ and $\ell_{c_2}(\beta, \Delta\Lambda)$ in (3.10) can be written as

$$E(\ell_{c_1}(\gamma)) = \sum_{i=1}^{n} v_i^{(m+1)} \log\left[\frac{\exp(\gamma^\top x_i)}{1+\exp(\gamma^\top x_i)}\right] + (1 - v_i^{(m+1)}) \log\left[\frac{1}{1+\exp(\gamma^\top x_i)}\right], \tag{3.13}$$

$$E(\ell_{c_2}(\beta, \Delta\Lambda)) = \sum_{i=1}^{n} v_i^{(m+1)} \left[\delta_i \left\{\log \Delta\Lambda(t_i) + \beta^\top z_i(t_i)\right\} - w_i^{(m+1)} \sum_{k=1}^{n} I(t_k \leqslant t_i)\exp\left\{\beta^\top z_i(t_k)\right\}\Delta\Lambda(t_k)\right]. \tag{3.14}$$

The M-step computes the estimates $(\tilde{\gamma}, \tilde{\beta}, \tilde{\Delta\Lambda}(\cdot))$ which maximizes the expected log-likelihood functions 3.13 and 3.14 obtained in the E-step.

Given the complete data $(T_i, \delta_i, Z_i(\cdot), u_i, \zeta_i)$, $i = 1, \ldots, n$, the estimation of the semiparametric transformation mixture cure model can be done through the embedded EM algorithm. The embedded EM algorithm applied to the log-likelihood function (3.10) can be boiled down to the following four steps.

E-Step: Compute the posterior expectation (3.11) of the latent variables u_i, at the $(m+1)^{\text{th}}$ iteration.

M-Step: Compute the estimates $\gamma^{(m+1)}$ which maximize the expected log-likelihood function (3.13), where $v_i^{(m+1)}$ is obtained in the E-step.

E-Step: Compute the posterior expectation (3.12) of the latent variables ζ_i, at the $(m+1)^{\text{th}}$ iteration.

M-Step: Compute the estimates $\beta^{(m+1)}$ and $\tilde{\Delta\Lambda}(\cdot)$ which maximize the expected log-likelihood function (3.14), where $w_i^{(m+1)}$ is obtained in the last E-step.

After convergence, we obtain the NPMLE of $\tilde{\gamma}$, $\tilde{\beta}$ and $\tilde{\Delta\Lambda}(t_i)(i=1,\ldots,n)$. Based on the NPMLEs, we compute the posterior expectation $E\left\{u_i|\mathscr{O},\tilde{\gamma},\tilde{\beta},\tilde{\Delta\Lambda}(t)\right\}$, denoted as $\tilde{u}_i$, and weight $E\left\{\zeta_i|\mathscr{O},\tilde{u}_i,\tilde{\gamma},\tilde{\beta},\tilde{\Delta\Lambda}(t)\right\}$ denoted as $\tilde{w}_i\,(i=1,\ldots,n)$. Given $\tilde{w}_i$, we differentiate the log-likelihood function (3.10) with respect to $\Delta\Lambda(t_i)$ and set it to zero, obtaining

$$\Delta\Lambda(t_i)=\frac{\delta_i}{\sum_{k=1}^{n} I\left(t_k \geqslant t_i\right)\tilde{w}_k v_k \exp\left\{\beta^{\top} z_k(t_k)\right\}}. \tag{3.15}$$

Plugging $\Delta\Lambda(t_i)$ back into function (3.9) gives a very similar form to the Cox partial log-likelihood function, a weighted version of partial log-likelihood function of β,

$$\ell_n(\beta)=\sum_{i=1}^{n}\delta_i\left(\beta^{\top} z_i(t_i)-\log\left[\sum_{k=1}^{n} I(t_k \geqslant t_i)\tilde{w}_k v_k \exp\left\{\beta^{\top} z_k(t_k)\right\}\right]\right), \tag{3.16}$$

and a pseudo log-likelihood function of γ,

$$\ell_n(\gamma)=\sum_{i=1}^{n}\tilde{u}_i\log\left[\frac{\exp(\gamma^{\top}x_i)}{1+\exp(\gamma^{\top}x_i)}\right]+(1-\tilde{u}_i)\log\left[\frac{1}{1+\exp(\gamma^{\top}x_i)}\right]. \tag{3.17}$$

We use the functions 3.16 and 3.17 to incorporate penalties for variable selection. As pointed out by Zeng and Lin (2006, 2007b), the function (3.16) inherits several good properties. For instance, it is a strict concave function and the resulted estimator $\tilde{\beta}$ is an efficient MLE of β.

3.3 Bi-level Variable Selection in STMCMs

In an attempt to select significant variables to increase model interpretability, a bi-level variable selection procedure for STMCM model is carried out in the framework of penalized regression. The objective function of penalized regression can be represented as $Loss\ function + Penalty\ function$.

Residual sum of squared error (RSS) under regression model or negative log-likelihood function otherwise is usually chosen as the loss function. In general, let α denote the parameters of interest, the resulted estimator of α from penalized regression is obtained by minimizing $Loss(\alpha)+p_{\lambda_n}(\alpha)$, where $p_{\lambda_n}(\cdot)$ denotes a defined penalty function measuring the complexity of the enclosed coefficient vector and λ_n is a non-negative tuning parameter controlling the degree of penalization. Without loss of generality, in a regression model with d single variables covered by J factors (groups of variables), let A_j be a size d_j subset of $\{1,\ldots,d\}$ representing known grouping structure of the design vectors, $j=1,\ldots,J$. Here we consider the scenario that the d variables are partitioned into J disjoint groups, and thus $d=\sum_{j=1}^{J} d_j$. Let $\alpha_{A_j}\in\mathbb{R}^{d_j}$ be the coefficient vector corresponding to the variable subset A_j. Therefore, bi-level variable selection can be realized by maximizing the penalized negative log-likelihood function, i.e. by maximizing the following objective function

$$Q_n(\alpha)=-\frac{1}{n}\ell_n(\alpha)+\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\alpha^{(j)}|)=-\frac{1}{n}\ell_n(\alpha)+\sum_{j=1}^{J}\sum_{k=1}^{d_j} p_{\lambda_n}\left(\alpha_{jk}\right). \tag{3.18}$$

To chose the tuning parameter λ_n, Akaike information criterion (AIC), Bayesian information criterion (BIC) and generalized cross validation score (GCV) are reasonable and commonly used criteria. The unknown $\alpha=\left(\alpha_{A_1}^{\top},\ldots,\alpha_{A_J}^{\top}\right)^{\top}\in\mathbb{R}^d$ is the parameter of interest. We aim at identifying zero vectors $\alpha_{A_j}'s$ and further the non-zero components if α_{A_j} is a non-zero vector.

Many approaches for group selection have been studied under various statistical modeling. The group LASSO penalty, first introduced by Bakin et al. (1999), uses an L_2-norm of the coefficients associated with a group of variables as the penalty function. Later, including the group LASSO as a special case, a quite general composite absolute penalty (CAP) for group selection was proposed by Zhao et al. (2009). Huang et al. (2009) considered the problem of simultaneous group and individual variable selection, or bi-level selection, and proposed a group bridge method and a general framework for bi-level selection under generalized linear models and further derived a local coordinate descent algorithm. Huang et al. (2014) extended the group bridge approach in Huang et al. (2009) to the Cox PH model. With grouping information provided, the penalties that are able to accommodate grouping structure in covariates should be considered. To accommodate

the grouping information in the semiparametric transformation mixture cure models, we consider group bridge penalty in this work.

Under the semiparametric transformation mixture cure model, the variable selection procedure involves two parts: identifying a set of significant variables for the latency part and identifying a set of significant variables for the incidence part. In this work, the group bridge penalty (Huang et al., 2009) is utilized for both parts of the model, in order to simultaneously select important variables and important groups of variables.

3.3.1 Group Bridge Penalty

Group bridge penalty was first proposed in multiple linear regression setting which simultaneously selects variables at both group level and individual variable level. As shown in Huang et al. (2009), it enjoys the oracle property in group selection, in that it can correctly select important groups with probability tending to one. Afterwards, it has been extended to Cox PH model by Huang et al. (2014). The group bridge penalty function is defined as

$$\sum_{j=1}^{J} p_{\lambda_n}^{(j)}(|\alpha_{A_j}|) \;=\; \lambda_n \sum_{j=1}^{J} c_j \|\alpha_{A_j}\|_1^{\omega} \;=\; \lambda_n \sum_{j=1}^{J} c_j \left(\sum_{k \in A_j} |\alpha_k| \right)^{\omega}, \tag{3.19}$$

where α is the vector of paramaters of interest, the constant $\omega \in (0,1)$ and c_j's are some between-group level weights accounting for group sizes. According to Huang et al. (2009), a simple choice of c_j is $c_j \propto d_j^{1-\omega}$ adjusting for the size of each group of coefficients. The value α that minimizes $Q_n(\alpha)$ in (3.18) with the group bridge penalty (3.19) is referred as a group bridge estimator. Note that when $d_j = 1$, the group bridge penalty in (3.19) simplifies to the standard bridge penalty. If further $c_j = 1$ and $\omega = 1$, it is reduced to the standard LASSO penalty. By the last representation in (3.19), the group bridge penalty could be viewed as imposing a LASSO penalty at within-group level while employing a bridge penalty at between-group level. Due to the essence of LASSO, the resulting estimator fails to achieve the selection consistency at within-group level. This sparked more recent research on developing estimators which possess the oracle property at both levels. Since the adaptive LASSO penalty is a natural extension of LASSO penalty, a simple remedy

named adaptive group bridge penalty, was proposed by Wang et al. (2009), in which the adaptive LASSO penalty is used at within-group level.

3.3.2 Computational Algorithm

Due to the non-convexity of the penalty functions, it is challenging to directly maximize the objection function $Q_n(\alpha)$ in (3.18). Following Huang et al. (2009, 2014), we formulate an equivalent minimization problem that is easier to be solved computationally. Let $\nabla\ell_n(\alpha) = -\partial\ell_n(\alpha)/\partial\alpha$ and $\nabla^2\ell_n(\alpha) = -\partial^2\ell_n(\alpha)/\partial\alpha^2$ denote the gradient vector and the Hessian matrix of ℓ_n respectively. Consider the Cholesky decomposition of $\nabla^2\ell_n(\alpha)$, i.e., $\nabla^2\ell_n(\alpha) = X^\top X$, and set the pseudo-response vector to be $Y = (X^\top)^{-1}[\nabla^2\ell_n(\alpha)\alpha - \nabla\ell_n(\alpha)]$. Then $-\ell_n(\alpha)$ can be approximated by the quadratic form $\frac{1}{2}(Y - X\alpha)^\top(Y - X\alpha)$.

Recall for group bridge estimation, the objective function is $Q_n(\alpha) = -\frac{1}{n}\ell_n(\alpha) + \lambda_n\sum_{j=1}^J c_j\|\alpha_{A_j}\|_1^\omega$. Following Huang et al. (2014), it is computational easier and more efficient to consider an equivalent minimization of function

$$W_n(\alpha,\theta) \;=\; \frac{1}{2}\|Y - X\alpha\|_2^2 + \sum_{j=1}^J \theta_j^{1-1/\omega} c_j^{1/\omega}\|\alpha_{A_j}\|_1 + \tau_n\sum_{j=1}^J \theta_j$$

over (α,θ), where τ_n is a penalty tuning parameter. Huang et al. (2009) or Proposition 1 of Huang et al. (2014) proved that minimizing $Q_n(\alpha)$ in (3.18) is equivalent to minimizing $W_n(\alpha,\theta)$ if

$$\lambda_n = \tau_n^{1-\omega}\omega^{-\omega}(1-\omega)^{\omega-1}. \tag{3.20}$$

The minimization of W_n could be treated as a LASSO-type optimization problem, and thus the shooting algorithm and coordinate descent algorithm could be applied to solve it numerically.

We are interested in narrowing down the set of parameters in both latency part and incidence part of the STMCM. Thus, we apply group bridge penalty function (3.19) of different parameters separately to the partial log-likelihood function (3.16) of the survival model and the pseudo log-likelihood function (3.17) of the cure rate model. Note that even though the log-likelihood function (3.17) is derived from a logistic regression model, it is the log-likelihood function of

quasi-binomial model due to the fact that $\tilde{u}_i$ is not dichotomized but instead an empirical probability estimated via EM algorithm.

The iterative group bridge penalized variable selection algorithm for β, the parameters in the survival model, is given below.

Step 1. Use the EM algorithm presented in Section 3.2 to compute $\tilde{\beta}$ and $\widetilde{\Delta\Lambda}(T_i)$'s, then compute the weights $\tilde{v}_i$'s and $\tilde{w}_i$'s and formulate the weighted version of the partial log-likelihood function $\ell_n(\beta)$ given in (3.16).

Step 2. Initialize by setting $\beta^{(0)} = \tilde{\beta}$.

Step 3. At the m^{th} step, compute the pseudo-covariates-matrix X and pseudo-response Y from the Cholesky decompositions based on $\beta^{(m-1)}$ from previous step.

Step 4. Compute

$$\theta_j^{(m)} \;=\; c_j \left(\frac{1-\omega}{\tau_n \omega}\right)^{\omega} \|\beta_{A_j}^{(m-1)}\|_1^{\omega}, \;\; j = 1, \ldots, J.$$

Step 5. Use coordinate descent algorithm to calculate the updated estimate

$$\beta^{(m)} \;=\; \arg\min_{\beta} \{\frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{J} (\theta_j^{(m)})^{1-1/\omega} c_j^{1/\omega} \|\beta_{A_j}\|_1\}$$

.

Step 6. Repeat Steps 3-5 until $\|\beta^{(m)} - \beta^{(m-1)}\|_2 / \|\beta^{(m-1)}\|_2$ is small.

Since the group bridge penalty function is not convex, for a fixed τ_n value, the above algorithm may only produce a local minimizer of $W_n(\beta, \theta)$. To solve this issue, based on the relationship between τ_n and λ_n in (3.20), we consider the grid values of tuning parameter λ over the range $[0, 40]$ with increment size 0.2 in our simulation studies, and choose the one that gives the smallest value of $W_n(\beta, \theta)$ for the final group bridge estimate calculation.

The $(p+1)\times 1$ vector γ of the regression coefficients in the logistic regression includes the intercept term which should be excluded from penalization. Assuming the covariate vector $X_{[-1]} = (X_1,\ldots,X_p)^\top$ is grouped into $\tilde{J}$ groups, let B_j be a size p_j subset of $\{1,\ldots,p\}$, $j=1,\ldots,\tilde{J}$, $p_j = |B_j|$, $p = \sum_{j=1}^{\tilde{J}} p_j$. It is possible that $p \neq q$ or $\tilde{J} \neq J$. The modified computational algorithm for γ is given below.

Step 1. Use the EM algorithm presented in Section 3.2 to compute $\tilde{\gamma}, \hat{\beta}_{GB}$ and $\tilde{\Delta\Lambda}(T_i)$'s , $\tilde{w}_i$'s, then updated the weights $\hat{u}_i$'s to obtain the the partial log-likelihood function $\ell_n(\gamma)$ given in (3.17).

Step 2. Initialize by setting $\gamma^{(0)} = \tilde{\gamma}$.

Step 3. At the m^{th} iteration, compute the pseudo-covariates-matrix X and pseudo response Y in the cholesky decomposition based on $\gamma^{(m-1)}$ from previous step.

Step 4. At the m^{th} iteration, fix the intercept term γ_0 in γ as $\gamma_0^{(m-1)}$, and update $\gamma^{(m-1)}$ excluding the intercept term, denoted as $\gamma_{(-0)}^{(m-1)}$.

Step 4.1. Compute $V = X_{[-1]}$, i.e. X without the first column, and $W^{(m)} = Y - V\gamma_{(-0)}^{(m)}$,

Step 4.2. Compute

$$\theta_j^{(m)} \;=\; c_j \left(\frac{1-\omega}{\tau_n \omega}\right)^{\omega} \|\gamma_{A_j}^{(m-1)}\|_1^{\omega}, \;\; j=1,\ldots,\tilde{J}.$$

Step 4.3. Use coordinate descent algorithm to calculate the updated estimate

$$\gamma_{(-0)}^{(m)} \;=\; \arg\min_{\gamma_{(-0)}} \{\frac{1}{2}\|Y - X_{[1]}\gamma_{(0)}^{(m-1)} - V\gamma_{(-0)}\|_2^2 + \sum_{j=1}^{\tilde{J}} (\theta_j^{(m)})^{1-1/\omega} c_j^{1/\omega} \|\gamma_{A_j}\|_1\}$$

.

Step 5. At the m^{th} iteration, update the intercept term, denoted as $\gamma_{(0)}^{(m)}$.

$$\begin{aligned}
\gamma_{(0)}^{(m)} &= \arg\min_{\gamma_{(0)}} \frac{1}{2}\|Y - X\gamma^{(m-1)}\|_2^2, \\
&= \arg\min_{\gamma_{(0)}} \frac{1}{2}\|W_{(-0)}^{(m)} - X_{[1]}\gamma_{(0)}\|_2^2, \\
\gamma_{(0)}^{(m)} &= \frac{\left(W_{(-0)}^{(m)}\right)^\top X_{[1]}}{X_{[1]}^\top X_{[1]}}.
\end{aligned}$$

Step 6. Repeat Steps 3-5 until $\|\gamma^{(m)} - \gamma^{(m-1)}\|_2 / \|\gamma^{(m-1)}\|_2$ is small.

After comparing several iterative algorithms that conduct the estimation and penalization in different order, we recommend the penalization of γ inside of the EM algorithm, in order to save computing time while maintain estimation accuracy. As a result, the penalization of γ is right after obtaining the estimates of the uncured probabilities, i.e. the estimates of latent variables u_i's. That is, in the embedded EM algorithm, the objective function in the first maximization step is (3.13), instead of (3.17), plus the group bridge penalty. The estimates of γ obtained from the embedded EM algorithm are group bridge-penalized estimates. We cycle the two iteration procedures for β and γ a few times until no update on the values of all parameters.

3.3.3 Tuning Parameter Selection

A tuning parameter in variable selection balances goodness-of-fit and sparsity, and thus the choice of a tuning parameter is essential to the performance of the penalized log-likelihood procedures discussed in this chapter. In practice, it is often to set a range for a tuning parameter and find the corresponding estimator for each fixed tuning parameter value in the range. Then a tuning parameter value is selected which optimizes a certain model selection criterion. In this paper, tuning parameters are selected via the minimization of criterion AIC, BIC or GCV. Specifically, let $\hat{d}(\lambda_n)$ denote the cardinality of the set of indices of nonzero coefficients for a fixed tuning parameter λ_n. Then the AIC, BIC and GCV are given respectively by

$$AIC(\lambda_n) = \log\left[\frac{\ell_n(\hat{\beta})}{n}\right] + \frac{2\hat{d}(\lambda_n)}{n},$$

$$BIC(\lambda_n) = \log\left[\frac{\ell_n(\hat{\beta})}{n}\right] + \frac{\log(n)\hat{d}(\lambda_n)}{n},$$

$$GCV(\lambda_n) = \frac{\ell_n(\hat{\beta})}{n\left[1-\hat{d}(\lambda_n)/n\right]^2}.$$

3.4 Simulation Studies

In this section, we examine through simulation studies the variable selection performance under the semiparametric transformation mixture cure models.

In order to demonstrate the existence of a cure fraction, we generate 500 observations in each repetition case from a specified semiparametric transformation model. Of those 500, 20% are cured and in total 40% are censored, consisting of the 20% cured observations, and another 20% censored uncured observations. Failure times are generated from the semiparametric transformation model with the logarithmic transformation function G specified by (3.5). We consider three transformation models indexed by $r = 0, 0.5, 1$. Note that $r = 0$ and $r = 1$ correspond to Cox PH model and PO model respectively. A Weibull distribution $\Lambda(t) = at^b$ with some $a, b > 0$ is assumed as the baseline cumulative hazard function. Specifically, the failure times $T^* = \{[(1/U)^r - 1]\exp(-\beta^\top Z)/(ar)\}^{1/b}$, where U is a random variable generated from the uniform distribution over $(0,1)$. Censoring times are generated from a uniform distribution over $(\min\{T^*\}+c, \max\{T^*\}-c)$, where c is chosen to obtain a censoring rate around 20%. The grouping structure of covariates and the coefficient values are described below for three different cases with Case 1 for smaller number of covariates and Cases 2 and 3 for larger number of covariates. In addition, in Case 1, the coefficients within each group are either all zeros or all non-zeros, while Cases 2 and 3 consider mixed groups with some zeros and some non-zeros.

Case 1. There are 5 groups with 3 covariates within each group, and thus totally 15 covariates in the survival model and logistic regression model. The true values of β_i's are taken to be $(\beta_1, \beta_2, \beta_3)^\top = (0.5, 1, 1.5)^\top$, $(\beta_4, \beta_5, \beta_6)^\top = (1,1,1)^\top$, $(\beta_7, \ldots, \beta_9)^\top = (0, \ldots, 0)^\top$, $(\beta_{10}, \ldots, \beta_{12})^\top = (0, \ldots, 0)^\top$, $(\beta_{13}, \ldots, \beta_{15})^\top = (0, \ldots, 0)^\top$. The true values of γ_i's are taken to be $\gamma_0 = 1.8$, $(\gamma_1, \gamma_2, \gamma_3)^\top =$

$(1,1,1)^\top$, $(\gamma_4,\ldots,\gamma_6)^\top = (0,\ldots,0)^\top$, $(\gamma_7,\ldots,\gamma_9)^\top = (0,\ldots,0)^\top$, $(\gamma_{10},\ldots,\gamma_{12})^\top = (0,\ldots,0)^\top$, $(\gamma_{13},\gamma_{14},\gamma_{15})^\top = (-1,-1,-1)^\top$. The covariates $(x_1,\ldots,x_{15})^\top = (z_1,\ldots,z_{15})^\top$ are generated in the following way. We first simulate $V_1,\ldots,V_{15}$ from the standard normal distribution, and independently $Z_1,\ldots,Z_5$ from an $AR(1)$ model with initial the standard normal distribution and $\mathrm{Cov}(Z_i,Z_j) = 0.4^{|i-j|}$, $i,j = 1,\ldots,5$. Then we set the covariate $z_j = (Z_{g_j} + V_j)/4$, $j = 1,\ldots,15$, where g_j is the smallest integer greater than $(j-1)/3$, and the z_j's with the same value of g_j belong to the same group.

Case 2. There are 10 groups with 4 covariates within each group. The true values of β_i's are taken to be $(\beta_1,\ldots,\beta_4)^\top = (1,1,1,1)^\top$, $(\beta_5,\ldots,\beta_8)^\top = (0,\ldots,0)^\top$, $(\beta_9,\ldots,\beta_{12})^\top = (-1.5,-1.5,-1.5,-1.5)^\top$, $(\beta_{13},\ldots,\beta_{16})^\top = (1.5,-1,-1.5,1)^\top$, $(\beta_{17},\ldots,\beta_{40})^\top = (0,\ldots,0)^\top$. The true values of γ_i's are taken to be $\gamma_0 = 3$, $(\gamma_1,\ldots,\gamma_4)^\top = (0,\ldots,0)^\top$, $(\gamma_5,\ldots,\gamma_8)^\top = (-2,-2,-2,-2)^\top$, $(\gamma_9,\ldots,\gamma_{12})^\top = (0,\ldots,0)^\top$,$(\gamma_{13},\ldots,\gamma_{16})^\top = (1.5,1.5,1.5,1.5)^\top$, $(\gamma_{17},\ldots,\gamma_{20})^\top = (1,1,1,1)^\top$, $(\gamma_{21},\ldots,\gamma_{40})^\top = (0,\ldots,0)^\top$. The 40 covariates $(x_1,\ldots,x_{40})^\top = (z_1,\ldots,z_{40})^\top$ are generated in a similar manner as in Case 1.

Case 3. There are 6 groups with 3 of larger size and 3 of smaller size. The true values of β_i's are taken to be $(\beta_1,\ldots,\beta_{10})^\top = (0,1.5,0,1.5,0,0,1.5,0,1.5,0)^\top$,
$(\beta_{11},\ldots,\beta_{20})^\top = (-1,-1,-1,-1,1,1,1,1,1,1)^\top$, $(\beta_{21},\ldots,\beta_{30})^\top = (0,\ldots,0)^\top$, $(\beta_{31},\ldots,\beta_{34})^\top = (0,1,0,-1)^\top$, $(\beta_{35},\ldots,\beta_{38})^\top = (-1,1,1,-1)^\top$, $(\beta_{39},\ldots,\beta_{42})^\top = (0,\ldots,0)^\top$. The true values of γ_i's are taken to be $\gamma_0 = 2.5$, $(\gamma_1,\ldots,\gamma_{10})^\top = (2,0,2,0,0,0,2,0,2,0)^\top$, $(\gamma_{11},\ldots,\gamma_{20})^\top = (0,\ldots,0)^\top$, $(\gamma_{21},\ldots,\gamma_{30})^\top = (1,1,1,1,-1,-1,-1,-1,-1,-1)^\top$, $(\gamma_{31},\ldots,\gamma_{34})^\top = (-1,1,0,0)^\top$, $(\gamma_{35},\ldots,\gamma_{38})^\top = (1,-1,-1,1)^\top$, $(\gamma_{39},\ldots,\gamma_{42})^\top = (0,\ldots,0)^\top$. The 42 covariates $(x_1,\ldots,x_{42})^\top = (z_1,\ldots,z_{42})^\top$ are generated in a similar manner as in Case 1.

The results on parameter estimation and variable selection for Cases 1, 2 and 3 are presented in Tables 3.1- 3.4 respectively. For each simulated dataset we apply group bridge (GB) approach to three different semiparametric transformation models for parameter estimation and group/variable selection. For simplicity we use the same $\omega = 0.5$ in the penalty function (3.19) for both compo-

Table 3.1: Estimation results of group bridge penalized estimator for nonzero covariates in STMCM Case 1 for 100 repetitions.

ρ		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
0	Bias	-0.1289	-0.1450	-0.3411	-0.1389	-0.1619	-0.2059
	MSE	0.0618	0.0630	0.1775	0.0673	0.0682	0.0892
0.5	Bias	0.0729	-0.1430	-0.4601	-0.1405	-0.1708	-0.2075
	MSE	0.0394	0.0614	0.2430	0.0519	0.0662	0.0715
1	Bias	0.1475	-0.1596	-0.4998	-0.1530	-0.1602	-0.1916
	MSE	0.0494	0.0537	0.2780	0.0509	0.0496	0.0627
		$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}a_3$	$\hat{\gamma}_{13}$	$\hat{\gamma}_{14}$	$\hat{\gamma}a_{15}$
0	Bias	-0.2060	0.1168	0.2306	0.3876	0.2987	0.3091
	MSE	0.3014	0.4361	0.3925	0.2405	0.2567	0.2734
0.5	Bias	-0.3517	-0.0218	0.2505	0.5096	0.4528	0.3822
	MSE	0.4573	0.5643	0.4704	0.4213	0.4767	0.4487
1	Bias	-0.3595	-0.1593	0.0054	0.5292	0.4953	0.4265
	MSE	0.1844	0.2324	0.3579	0.1741	0.2156	0.2547

nents in the three cases. In the three tables, the columns #Group and #Variable stand for average number of groups selected and average number of variables selected, respectively. The tables also present %Corr.G and %Corr.V, where %Corr.G gives the portion of occasions on which the fitted model contains exactly the same groups as the underlying true model, while %Corr.V calculates the portion of occasions on which the fitted model contains exactly the same variables as the underlying true model. All the numbers in parenthesis are the corresponding standard errors based on 100 repetitions.

Tables 3.1 and 3.2 present the estimation and variable selection results for Case 1. From Table 3.1, we observe that the bias and MSE in the latency part, that is, the $\hat{\beta}$'s, are relatively low. However, the bias and MSE for the incidence part, that is, the $\hat{\gamma}$'s, can be large, especially when the semiparametric transformation index $r = 0.5$, compared with $r = 0, 1$. We also compare the estimation performance of when $n = 500$ with smaller sample sizes, such as $n = 100$ and $n = 200$, and observe that the bias and MSE decrease with sample size increases. This is not surprising as we need enough cured observations to obtain enough information to distinguish susceptible subjects from cured subjects and to proceed downstream estimation and variable selection in the

Table 3.2: Variable selection results in STMCM Case 1 for 100 repetitions.

ρ	Parameter	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	β	AIC	2.02(0.141)	5.98(0.147)	0.996(0.028)	0.996(0.016)
		BIC	2.14(0.253)	6.12(0.551)	0.987(0.131)	0.984(0.198)
		GCV	2.02(0.141)	5.98(0.147)	0.996(0.028)	0.996(0.016)
	γ	AIC	2.64(0.722)	6.64(1.535)	0.864(0.148)	0.891(0.102)
0.5	β	AIC	2.00(0.043)	5.99(0.102)	0.999(0.002)	0.999(0.007)
		BIC	2.09(0.177)	6.10(0.578)	0.988(0.151)	0.990(0.144)
		GCV	2.00(0.043)	5.99(0.102)	0.999(0.002)	0.999(0.007)
	γ	AIC	2.68(0.777)	6.70(1.904)	0.856(0.153)	0.869(0.115)
1	β	AIC	2.01(0.071)	6.01(0.071)	0.999(0.141)	0.997(0.005)
		BIC	2.10(0.207)	6.13(0.611)	0.988(0.174)	0.982(0.251)
		GCV	2.01(0.071)	6.01(0.071)	0.999(0.141)	0.997(0.005)
	γ	AIC	2.58(0.753)	6.37(1.717)	0.867(0.149)	0.875(0.111)
True model			2	6	1	1

Table 3.3: Variable selection results in STMCM Case 2 for 100 repetitions.

ρ	Parameter	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	β	AIC	3.24(0.453)	12.31(0.979)	0.976(0.044)	0.987(0.021)
		BIC	3.26(0.462)	12.63(1.185)	0.971(0.052)	0.984(0.037)
		GCV	3.24(0.453)	12.31(0.979)	0.976(0.044)	0.987(0.021)
	γ	AIC	6.63(1.899)	22.21(8.452)	0.637(0.190)	0.703(0.188)
0.5	β	AIC	2.83(0.326)	10.67(1.599)	0.985(0.035)	0.966(0.040)
		BIC	2.85(0.372)	11.22(1.437)	0.987(0.037)	0.980(0.037)
		GCV	2.83(0.326)	10.67(1.599)	0.985(0.035)	0.966(0.040)
	γ	AIC	5.82(1.673)	18.04(6.794)	0.712(0.148)	0.776(0.136)
1	β	AIC	2.76(0.441)	9.41(1.231)	0.977(0.042)	0.937(0.031)
		BIC	2.73(0.423)	10.23(1.517)	0.974(0.038)	0.956(0.033)
		GCV	2.76(0.441)	9.41(1.231)	0.977(0.042)	0.937(0.031)
	γ	AIC	5.85(1.345)	18.46(5.734)	0.717(0.135)	0.768(0.127)
True model			3	12	1	1

Table 3.4: Variable selection results in STMCM Case 3 for 100 repetitions.

ρ	Parameter	Tuning	#Group	#Variable	%Corr.G	%Corr.V
0	β	AIC	3.78(0.996)	20.56(6.472)	0.920(0.172)	0.871(0.104)
		BIC	3.82(0.875)	22.84(6.293)	0.922(0.127)	0.842(0.106)
		GCV	3.78(0.996)	20.56(6.472)	0.920(0.172)	0.871(0.104)
	γ	AIC	3.86(1.125)	19.71(6.993)	0.707(0.176)	0.617(0.126)
0.5	β	AIC	3.75(1.023)	18.92(3.567)	0.858(0.124)	0.771(0.072)
		BIC	3.78(1.167)	21.60(3.658)	0.858(0.123)	0.805(0.067)
		GCV	3.75(1.023)	18.92(3.567)	0.858(0.124)	0.771(0.072)
	γ	AIC	4.14(1.119)	20.07(6.541)	0.677(0.153)	0.559(0.125)
1	β	AIC	3.15(1.089)	14.65(4.091)	0.792(0.142)	0.687(0.084)
		BIC	3.21(1.217)	17.72(4.624)	0.792(0.167)	0.731(0.095)
		GCV	3.15(1.089)	14.65(4.091)	0.792(0.142)	0.687(0.084)
	γ	AIC	3.93(1.253)	19.44(6.718)	0.637(0.159)	0.551(0.104)
True model			4	20	1	1

incident part. When sample size is small with relatively low cure rate, the computational algorithm for the logistic regression may not converge. Due to this intrinsic property of mixture cure model, there is a noticeable difference on estimation performance between the two components, and this difference is passed along to the variable selection procedure, and as a consequence the results in Tables 3.1 and 3.2 agree. In Cases 1 and 2, the variable selection performance for the latency part is very good, no matter which evaluation index is considered. The selected number of groups and variables are very close to the true model in latency part, with high correct selection rate at both levels. The variable selection performance for γ is satisfactory in Case 1, where all are zero or nonzero in each group and group size is small. In Case 2, when the number of covariates in full model increases from 15 to 40 and the sparsity level increases from 60% to 70%, we observe a slight drop in the percentages of correctly selecting groups or variables. The resulted estimates of γ tends to keep large number of groups and variates than the true model. In Case 3, we explore our proposed method in varying group sizes setting, that is, sparsity exists at within group level. The selection performance of latency part is not as good as that in Cases 1 and 2, but still acceptable. However, for the incidence part, the selected number of groups and variables are very close to the true model but with relatively low percentage in selection precision for both at group level

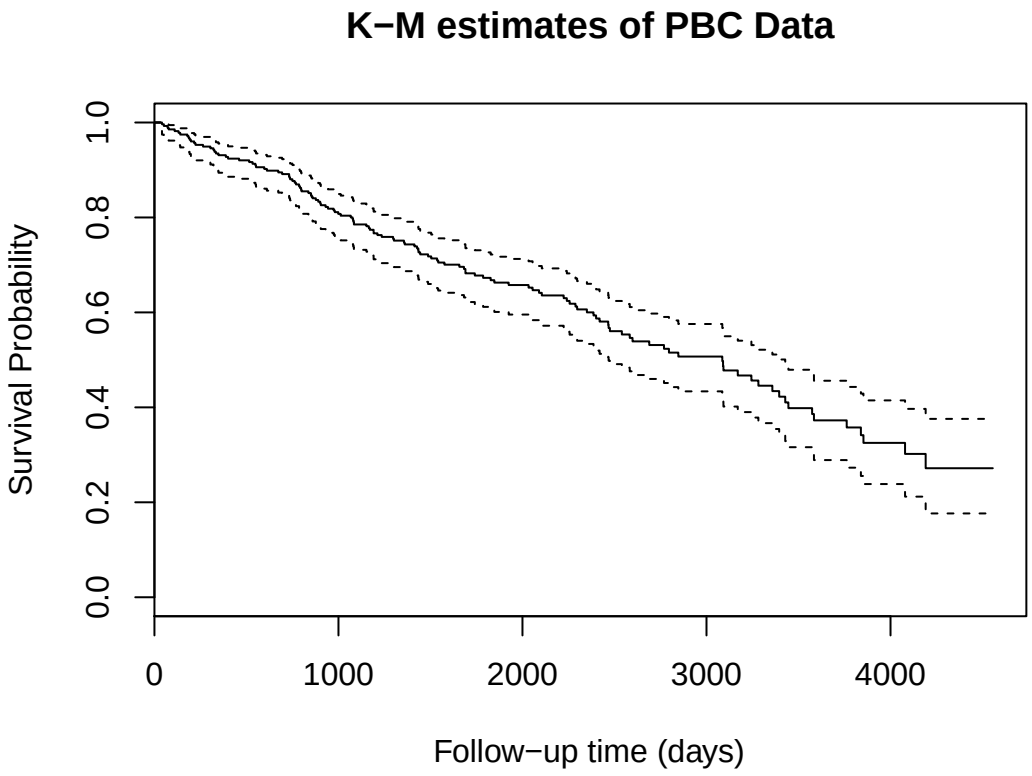
and within group level. One possible reason is that the group bridge penalty does not possess the selection consistency at within group level. Further investigation could be done by replacing the group bridge penalty with some other bi-level variable selection penalties which enjoy the oracle property.

3.5 Real Data Analysis

In this section we demonstrate the use of proposed bi-level variable selection methods by analyzing two well-known datasets, i.e., the primary biliary cirrhosis (PBC) data and the Wisconsin Prognostic Breast Cancer (WPBC) data. In Figure 3.1, we show the plots of K-M estimators of survival function for those two datasets. Both plots show a clear plateau in the right tail, of which the height is an estimator of the cure proportion $1-u$. However, due to the right censoring, it could happen that some of the observations in the plateau correspond to censored uncured observations. The K-M curve of PBC data levels off around 0.3, while that of WPBC data levels off above 0.6. It suggests that the mixture cure models is plausible to fit these two data sets.

3.5.1 Primary Biliary Cirrhosis (PBC) Data

PBC is a fatal chronic liver disease. An investigation of PBC was conducted between 1974 and 1984 in the Mayo Clinic trials study. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the dataset participated in the randomized trial and most of which complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. Among the 312 randomized patients with PBC, 152 were assigned to the drug D-penicillanmine, while the rest of them were assigned to a control group with placebo drug. 140 patients died from PBC disease during the ten years follow up study. Censoring was due



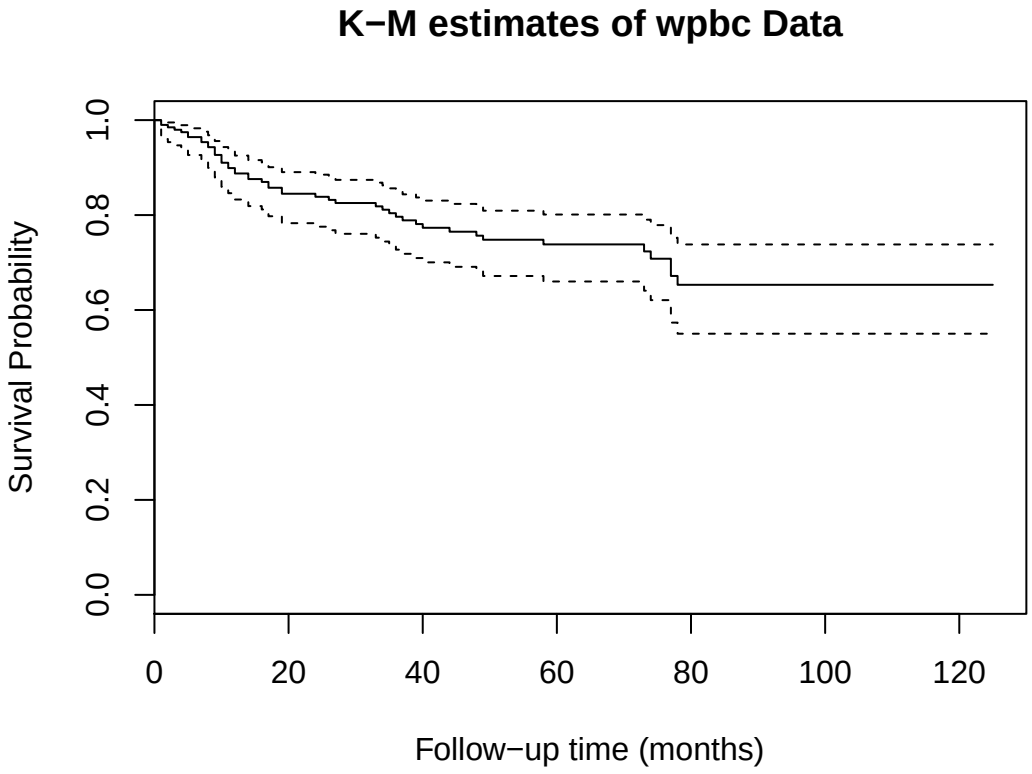


Figure 3.1: Plots of Kaplan-Meier estimates of survival probabilities in primary biliary cirrhosis dataset and Wisconsin prognostic breast cancer dataset.

to liver transplantation. It was of primary interest to study the effectiveness of D-penicillanmine in curing PBC disease. The PBC data have been investigated by many authors, e.g., Zhang and Lu (2007) and Huang et al. (2014).

Due to addressing missing data is beyond the scope of this work, we focus on the observed 17 risk factors and their effects on the observed failure times of the 276 complete cases in the full model. The censoring rate is 60%. The observed failure time is the number of days between registration and the earlier of death, liver transplantation, or study analysis time in July 1986. The dataset is available in R package 'survival'. The description of the covariates are summarized in Table 3.5. All these 17 risk factors are naturally categorized into 9 different groups by the measurements of different aspects. We fit the PBC data in a class of group bridge penalized semiparametric transformation mixture cure models. We consider a grid of r over the range $[0,6]$ with increment of 0.1 for logarithmic transformation indexes in the semiparametric transformation model. For each r the EM algorithm is implemented and the r that yields the largest log-likelihood is chosen as the best transformation model for fitting the data and to proceed for the variable selection procedure.

Table 3.6 presents the variable selection results for both of the two parts in the mixture cure model with $\rho = 0.6$ selected from a grid search based on the maximum likelihood function of the proposed mixture model. From the table we can see that the groups $G_1 - G_4$, G_7 and G_9 should be excluded from the reduced latency part no matter which tuning parameter selection criteria is used. For the incidence part, besides the same groups, i.e., G_5, G_6, and G_8, as in the latency, it includes G_2 : *gender* for modeling the cure fraction. At the individual variable level, *bil*, *trig*, *alb*, *prot*, *stage* and *cop* are selected under AIC and GCV as significant variables for modeling the survival distribution for susceptible observations. BIC returns similar results but exclude *prot* in G_6. By AIC, *gender*, *chol*, *alb* and *stage* are identified as important to estimates the cure rate.

3.5.2 Wisconsin Prognostic Breast Cancer (WPBC) Data

There are 198 records of breast cancer patients in the WPBC data. Each record represents follow-up data (in months) for one breast cancer case aLONG with 34 attributes, i.e., ID, outcome, 32

Table 3.5: PBC data: dictionary of covariates.

Group	Variable	Type	Definition
Age(G_1)	Z_1	C	Age(years)
Gender(G_2)	Z_2	D	Female gender(0 male and 1 female)
Phynotype(G_3)	Z_3	D	Ascites(0 absence)
	Z_4	D	Hepatomegaly (0 absence and 1 presence)
	Z_5	D	Spiders(0 absence and 1 presence)
	Z_6	D	Edemaoed(0 no edema, 0.5 untreated or successfully treated and 1 unsuccessfully treated)
Liver function damage(G_4)	Z_7	C	Alkaline phosphatase(units/litre)
	Z_8	C	Sgot(liver enzyme units/ml)
Excretory function of the liver(G_5)	Z_9	C	Serum bilirubin(mg/dl)
	Z_{10}	C	Serum cholesterol(mg/dl)
	Z_{11}	C	Triglyserides(mg/dl)
Liver function damage(G_6)	Z_{12}	C	Albumin(g/dl)
	Z_{13}	C	Prothrombin time(seconds)
Treatment(G_7)	Z_{14}	D	Penicillamine v.s. placebo (1 control and 2 treatment)
Reflection(G_8)	Z_{15}	D	Stage (histological stage of disease, graded 1,2,3, or 4)
	Z_{16}	C	Urine copper(mcg/24hrs)
Haematology(G_9)	Z_{17}	C	Platelets(per cubic ml/1000)

Table 3.6: PBC data: estimates of coefficients.

Group	Covariates	$\hat{\beta}_{\text{MLE}}$	$\hat{\beta}_{\text{GB-AIC}}$	$\hat{\beta}_{\text{GB-BIC}}$	$\hat{\beta}_{\text{GB-GCV}}$	$\hat{\gamma}_{\text{GB-AIC}}$
G_1	age	1.963	0	0	0	0
G_2	gender	-0.1492	0	0	0	-0.9551
G_3	asc	0.3762	0	0	0	0
	hep	0.0863	0	0	0	0
	spid	0.0262	0	0	0	0
	edema	0.7405	0	0	0	0
G_4	alk	-0.0044	0	0	0	0
	sgot	0.1845	0	0	0	0
G_5	bil	0.0581	0.0321	0.0551	0.0321	0
	chol	0.0113	0	0	0	0.5928
	trig	-0.2301	-0.9669	-0.1170	-0.9669	0
G_6	alb	-0.8086	-0.9876	-0.9406	-0.9876	-0.1261
	prot	-1.0084	0.3948	0	0.3948	0
G_7	trt	-0.1046	0	0	0	0
G_8	stage	0.2274	0.2034	0.2172	0.2034	0.2070
	cop	0.3765	0.4152	0.4114	0.4152	0
G_9	plat	0.0694	0	0	0	0

real-valued input features. These are collected from consecutive patients seen by Dr. Wolberg since 1984, including only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The WPBC data can be found in the R package 'TH.data', and the data and related detailed information are originally available from the UCI machine learning repository, see *http://www.ics.uci.edu/ mlearn/databases/breast-cancer-wisconsin/*. The censoring rate of WPBC data is about 76.3%.

Attribute information:

1) ID number
2) Outcome (R = recur, N = nonrecur) With 151 N and 47 R in the dataset
3) Time (recurrence time if field $2 = R$, disease-free time if field $2 = N$)
4-33) Ten real-valued features are computed for each cell nucleus:
    a) radius (mean of distances from center to points on the perimeter)
    b) texture (standard deviation of gray-scale values)
    c) perimeter
    d) area
    e) smoothness (local variation in radius lengths)
    f) compactness (perimeter$^2/area - 1.0$)
    g) concavity (severity of concave portions of the contour)
    h) concave points (number of concave portions of the contour)
    i) symmetry
    j) fractal dimension ("coastline approximation"-1)
34) Tumor size - diameter of the excised tumor in centimeters
35) Lymph node status - number of positive axillary lymph nodes observed at time of surgery

For the ten real-valued features, the mean, standard error, and "worst" or largest (mean of the three largest values) of each feature were computed for each image, resulting in 30 features. For instance, the 4^{th} attribute listed in the dataset is Mean Radius, the 14^{th} is Radius SE and the 24^{th} is Worst Radius.

When we preprocessing the dataset, we found that the value of *Lymph node status* is missing in 4 cases. We limited our study to the 194 patients who had complete records. We categorized the 32 covariates into 12 groups, with the first ten groups classified by that ten features of a breast

mass. There are three covariates in the each of the first ten groups and *tumor size* and *lymph node status* belong to the last two groups respectively. Based on this grouping structure, the results of the three variable/group selection criteria are displayed in Figure 3.2. The results for GCV is omitted for the reason that it gives same results to those of AIC for $\hat{\beta}$ in latency. In the figure, each block represents a group. From Figure 3.2 we observe that, for $\hat{\beta}$ in latency, AIC and BIC select the same 5 important groups. In terms of number of important variables selected out of the total 32, based on AIC criterion the group bridge selects 10 variables, while BIC retains three more variables than AIC does. It suggests that *tumor size* and *lymph node status* are not important for the estimation of failure time distribution of the patients who experienced recurrence. *tumor size* is selected in incidence part, along with other 22 covariates. In the logistic model, only 9 covariates are eliminated from the full model.

3.6 Conclusion and Discussion

Semiparmetric mixture cure models have been proved to be an important class of survival models in analyzing time-to-event data, in situations where a portion of subjects are free of the disease risk. In some of the clinical trial studies, a series of measurements are collected and investigators are interested in identifying important covariates for the survival function of uncured patients and for the cure fraction. In this chapter, we proposed a group bridge penalized bi-level variable selection procedure for the semiparmetric transformation mixture cure model. It bases on a proposed embedded EM algorithm and penalized partial likelihood methods for a class of semiparametric transformation models and logistic regression, which alleviates model complexity and achieves promising finite sample performance in our empirical studies.

We applied the proposed method to analyze real data from clinical trial studies of primary biliary cirrhosis and breast cancer and obtained interpretable results. The real data examples well illustrate that the proposed approach greatly simplifies the model components when a large number of grouped variables available for modeling. Built on earlier attempts on variable selection in the

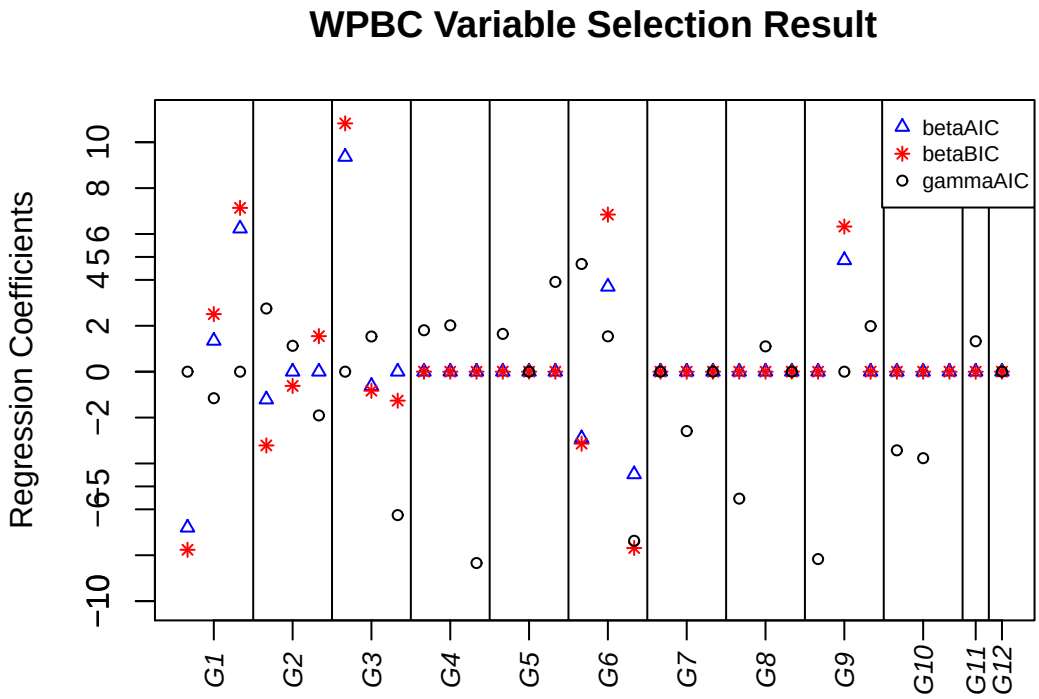


Figure 3.2: Plot of group bridge penalized coefficients in semiparametric transformation mixture cure model of WPBC data.

mixture cure model by Liu et al. (2012), this work further extends to variable selection when the covariates are grouped. The embedded EM algorithm makes the estimation be easily calculated. However, the proposed EM algorithm fails to directly obtain the asymptotic variances of estimators using the inverse of information matrix, and its convergence is slow. The penalized variable selection methods in nonlinear models tend to produce biases in the estimation of parameters. The magnitude of the bias depends on the choice of tuning parameter or weights. This leads to the fact that Hessian matrix-based variance estimation is no longer good for inference. Further work is in need to address this challenge.

3.7 Appendix

The likelihood function of the complete data for the semiparametric mixture cure model is

$$\begin{aligned}
&\prod_{i=1}^{n} h(t_i|z_i,x_i)^{\delta_i} S(t_i|z_i,x_i) \\
=&\prod_{i=1}^{n} \Big[h(t_i,\zeta_i|u_i=1)^{\delta_i} S(t_i,\zeta_i|u_i=1)\Big]^{u_i} f(u_i)\phi(\zeta_i) \\
=&\prod_{i=1}^{n} \Big[h(t_i,\zeta_i|u_i=1)^{\delta_i} S(t_i,\zeta_i|u_i=1)\Big]^{u_i} \pi(x_i)^{u_i}\left[1-\pi(x_i)\right]^{1-u_i}\phi(\zeta_i),
\end{aligned}$$

where $h(\cdot)$ is a hazard function and $S(\cdot)$ be the correspond survival function, $f(u_i)$ and $\phi(\zeta_i)$ are the probability density function of u_i and ζ_i, respectively.

Since $\delta_i u_i = \delta_i$, then the marginal probability density function of u_i and ζ_i satisfy

$$\begin{aligned}
f(u_i;t_i,\delta_i,x_i,z_i,\zeta_i) &\propto \prod_{i=1}^{n} S(t_i,\zeta_i|u_i=1)^{u_i}\pi(x_i)^{u_i}\left[1-\pi(x_i)\right]^{1-u_i}, \\
f(\zeta_i;t_i,\delta_i,x_i,z_i,u_i) &\propto \prod_{i=1}^{n} \Big[h(t_i,\zeta_i|u_i=1)^{\delta_i} S(t_i,\zeta_i|u_i=1)\Big]^{u_i}\phi(\zeta_i) \\
&\propto \prod_{i=1}^{n} h(t_i,\zeta_i|u_i=1)^{\delta_i} S(t_i,\zeta_i|u_i=1)^{u_i}\phi(\zeta_i) \\
&\propto \prod_{i=1}^{n} \zeta_i^{\delta_i} \exp\left[-\zeta_i \sum_{j=1}^{n} I\left(t_j \leqslant t_i\right) \exp\left\{\beta^{\top} z_i(t_j)\right\} \Delta\Lambda(t_j)\right]^{u_i} \phi(\zeta_i).
\end{aligned}$$

Since u_i is a binary variable taking values either 0 or 1, the conditional expectation of u_i can be calculated as

$$\begin{aligned}
\mathrm{E}[u_i|t_i,\delta_i,x_i,z_i,\zeta_i] &= \mathrm{P}(u_i=1|t_i,\delta_i,x_i,z_i,\zeta_i)\\
&= \mathrm{P}(u_i=1|t_i,\delta_i=0,x_i,z_i,\zeta_i)+\mathrm{P}(u_i=1|t_i,\delta_i=1,x_i,z_i,\zeta_i)\\
&= 0+[\delta_i+(1-\delta_i)\mathrm{P}(u_i=1|t_i,x_i,z_i,\zeta_i)]\\
&= \delta_i+(1-\delta_i)\frac{\mathrm{P}(u_i=1,t_i,x_i,z_i,\zeta_i)}{\mathrm{P}(t_i,x_i,z_i,\zeta_i)}\\
&= \delta_i+(1-\delta_i)\frac{\mathrm{P}(u_i=1,t_i,x_i,z_i,\zeta_i)}{\mathrm{P}(u_i=1,t_i,x_i,z_i,\zeta_i)+\mathrm{P}(u_i=0,t_i,x_i,z_i,\zeta_i)}\\
&= \delta_i+(1-\delta_i)\frac{\pi(x_i)S(t_i,\zeta_i|u_i=1)}{\pi(x_i)S(t_i,\zeta_i|u_i=1)+1-\pi(x_i)}\\
&= \delta_i+(1-\delta_i)\frac{\pi(x_i)S_u(t_i,\zeta_i)}{\pi(x_i)S_u(t_i,\zeta_i)+1-\pi(x_i)}.
\end{aligned}$$

Following the spirit of Zeng and Lin (2007a), for all commonly used transformations, including the classes of Box–Cox transformations and logarithmic transformations, $\exp\{-G(x)\}$ is the Laplace transformation of some function $\phi(x)$ such as

$$\exp\{-G(x)\}=\int_0^\infty \exp(-xt)\phi(t)\mathrm{d}t.$$

Clearly, $\int_0^\infty \phi(t)\mathrm{d}t=1$. The latent variable ζ_i has the density function $\phi(\cdot)$.

Since

$$G'(x)\exp\{-G(x)\}=\int_0^\infty t\exp(-xt)\phi(t)\mathrm{d}t,$$

the conditional expectation of the frailty ζ_i can be obtained as

$$\begin{aligned}
\mathrm{E}[\zeta_i|t_i,\delta_i,x_i,z_i,u_i] &= \frac{\int_0^\infty \zeta_i f(\zeta_i;t_i,\delta_i,x_i,z_i,u_i)\mathrm{d}\zeta_i}{\mathrm{P}(t_i,\delta_i,x_i,z_i,u_i)}\\
&= \frac{\int_0^\infty \zeta_i\zeta_i^{\delta_i}\left(\exp\left[-\zeta_i\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)\right\}\Delta\Lambda(t_j)\right]\right)^{u_i}\phi(\zeta_i)\mathrm{d}\zeta_i}{\int_0^\infty \zeta_i^{\delta_i}\left(\exp\left[-\zeta_i\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)\right\}\Delta\Lambda(t_j)\right]\right)^{u_i}\phi(\zeta_i)\mathrm{d}\zeta_i}\\
&= \frac{\int_0^\infty \zeta_i\zeta_i^{\delta_i}\exp\left[-u_i\zeta_i\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)\right\}\Delta\Lambda(t_j)\right]\phi(\zeta_i)\mathrm{d}\zeta_i}{\int_0^\infty \zeta_i^{\delta_i}\exp\left[-u_i\zeta_i\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)\right\}\Delta\Lambda(t_j)\right]\phi(\zeta_i)\mathrm{d}\zeta_i}\\
&= \begin{cases}
G'\left[\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)+\log u_i\right\}\Delta\Lambda(t_j)\right], & \delta_i=0,\\
-\dfrac{G''\left[\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)+\log u_i\right\}\Delta\Lambda(t_j)\right]}{G'\left[\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)+\log u_i\right\}\Delta\Lambda(t_j)\right]}\\
+G'\left[\sum_{j=1}^n I\left(t_j\leqslant t_i\right)\exp\left\{\beta^\top z_i(t_j)+\log u_i\right\}\Delta\Lambda(t_j)\right], & \delta_i=1.
\end{cases}
\end{aligned}$$

Chapter 4

SUMMARY

In survival analysis, one of the most critical tasks is to construct an interpretable survival model with high prediction accuracy. The advent of high dimensional data has made the task even more challenging. It is pivotal to choose a manageable set of risk factors out of all the collected data that are presumably related to survival outcome. In this dissertation, we investigated the problem of variable selection and survival prediction under two different types of semiparametric survival models with right-censored data. Due to the grouping structure that naturally in many real world applications, we considered bi-level variable selection approaches which can identify not only important individual variables but also important groups of variables and can simultaneously estimate the corresponding coefficients in survival models. Specifically, in Chapter 2 we employed the group bridge penalty, adaptive group bridge penalty and composite group bridge penalty for a class of semiparametric transformation models in the penalized regression fashion. The Laplace transformation and EM algorithm were proposed in order to obtain a weighted version of partial log-likelihood function of the coefficients in semiparametric transformation models. Then we used the negative weighted partial log-likelihood function as the loss function to which a penalty was added in the variable selection procedure. In Chapter 3, we developed an estimation procedure for semiparametric transformation mixture cure models. Based on an embedded EM algorithm, we derived the nonparametric estimators of the membership of susceptible subjects, their corresponding frailty and baseline hazards function. For the purpose of bi-level variable selection, a group bridge penalization approach was utilized for both latency and incidence components.

We established theoretically the variable selection consistency and asymptotic normality properties of the estimators resulted from adaptive group bridge and composite group bridge regularization approach under the semiparametric transformation models. In Chapter 2 we proved the

group selection consistency and derived the asymptotic convergence rate of group bridge penalized estimators for semiparametric transformation models. The mathematical proofs used the modern empirical process theory. We also conducted extensive simulation studies to explore the finite sample performance of all the proposed methods under the two types of survival models. The superiority and comparability of our proposed methods to existing ones was illustrated through different simulation settings. We analyzed two real datasets, i.e., the primary biliary cirrhosis data from Mayo Clinic trial of primary biliary liver cirrhosis and metastasis-free survival data from breast cancer, in order to demonstrate the implementation of the proposed methods.

Chapter 5

FUTURE RESEARCH

The proposed methods in this dissertation can be extended in several directions for future research. In Chapter 2, we studied the semiparametric transformation model with right-censored data. Other than right-censored data, the semeiparametric transformation model is also capable to handle recurrent events data and terminal events data (Zeng and Lin, 2009). The proposed methods may be extended to these types of data. The objective function we used to incorporate penalties is a negative weighted version of partial log-likelihood function. We adopted an EM algorithm to obtain estimation of the nonparametric baseline hazards function in survival function, however, other nonparametric estimation methods such as Breslow-typed estimator, spline estimator or kernel estimator could also be considered. Based on these possible alternatives of baseline hazards function estimation, the partial likelihood approach could be replaced by a kernel-smoothed profile likelihood method. Since this thesis has considered a class of semiparametric transformation models, we could further examine the performance of the proposed bi-level selection procedures when the true model is misspecified in a certain sense. This scrutiny may be difficult in theory but should be feasible through simulation studies.

In Chapter 3, we developed the bi-level variable selection procedure under a mixture cure model with the latency part modelled by a semiparametric transformation model. As an alternative to the mixture cure model, a promotion time cure model could be used to analyze survival data with a cure fraction. The survival distributions for both cured subjects and non-cured subjects can be integrated into one single formulation with the cured subjects being assumed to have survival time of infinity. We may modify the proposed methods for the promotion time cure model, parametric mixture cure models, or for other types of survival models and their extensions.

In this thesis, we restrained ourselves to the group-bridge-type penalties for bi-level variable

selection. For the aim of bi-level selection, we could also consider other existing penalization methods depending on research interests. For example, the group exponential LASSO penalty(Breheny, 2015) and group MCP penalty (Huang et al., 2012) also have the capability to conduct bi-level variable selection, although their theoretical properties under extended models need further investigation. On the other side, the grouping structure considered in this dissertation doesn't include the overlapping scenarios, while it is possible in real world applications that some risk factors could be included into more than one group. The proposed bi-level variable selection procedure could be further extended to complications associated with overlapping covariates in groups. Variable selection is of primary interest in this thesis, but in future research we could take a closer look at estimation efficiency, covariance estimation and hypothesis testing under semiparametric transformation models and mixture cure models.

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Bakin, S. et al. (1999). Adaptive regression and model selection in data mining problems. *Ph.D Thesis, Open Access Theses*.

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

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

Chen, Y. Q. and Wang, M.-C. (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association*, 95(450):608–618.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220.

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

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Farewell, V. (1977). The combined effect of breast cancer risk factors. *Cancer*, 40(2):931–936.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gu, Y., Sinha, D., and Banerjee, S. (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime Data Analysis*, 17(1):123–134.

Gui, J. and Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 27(4):481–449.

Huang, J., Liu, L., Liu, Y., and Zhao, X. (2014). Group selection in the cox model with a diverging number of covariates. *Statistica Sinica*, 24(4):1787–1810.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, 57(2):383–388.

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kuk, A. Y. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Li, H. and Luan, Y. (2002). Kernel cox regression models for linking gene expression profiles to censored survival data. In *Biocomputing 2003*, pages 65–76. World Scientific.

Li, J. and Gu, M. (2012). Adaptive lasso for general transformation models with right censored data. *Computational Statistics & Data Analysis*, 56(8):2583–2597.

Liu, X., Peng, Y., Tu, D., and Liang, H. (2012). Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in Medicine*, 31(24):2882–2891.

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika*, 100(4):859–876.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, 91(2):331–343.

Lu, W. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine*, 26(20):3771–3781.

Park, M. Y., Hastie, T., and Tibshirani, R. (2006). Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227.

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243.

Peng, Y., Taylor, J. M., and Yu, B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis*, 13(3):351–369.

Ross, M. and Xian, Z. (1996). Survival analysis with long term survivors.

Seetharaman, I. (2013). *Consistent bi-level variable selection via composite group bridge penalized regression*. PhD thesis, Kansas State University.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Taylor, J. M. and Liu, N. (2007). Statistical issues involved with extending standard models. In *Biostatistics: Advances in Statistical Modeling and Inference*, volume 3, pages 299–311. World Scientific.

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. et al. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508–1516.

Tsodikov, A., Ibrahim, J., and Yakovlev, A. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078.

van der Vaart, A. W. and Wellner, J. A. (1997). Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608.

Wang, C. (2016). *Variable selection through adaptive elastic net for proportional odds model*. PhD thesis, The University of Texas at San Antonio.

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Wang, S., Nan, B., Zhu, N., and Zhu, J. (2009). Hierarchically penalized cox regression with grouped variables. *Biometrika*, 96(2):307–322.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic models of tumor latency and their biostatistical applications*. World Scientific.

Yin, G. and Ibrahim, J. G. (2005). A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics*, 61(2):403–412.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.

Zeng, D. and Lin, D. (2007a). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):507–564.

Zeng, D. and Lin, D. (2007b). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102(477):167–180.

Zeng, D. and Lin, D. (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics*, 65(3):746–752.

Zeng, D., Yin, G., and Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101(474):670–684.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106(495):1099–1112.

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

Zhang, J. and Peng, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15(4):455.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.