STATISTICAL LEARNING METHODS FOR SUBGROUP DISCOVERY WITH SURVIVAL OUTCOME

Beilin Jia

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

Chapel Hill
2021

Approved by:

Donglin Zeng

Joseph G. Ibrahim

Jason J.Z. Liao

Xianming Tan

Quefeng Li

ABSTRACT

Beilin Jia: Statistical Learning Methods for Subgroup Discovery with Survival Outcome
(Under the direction of Donglin Zeng and Joseph G. Ibrahim)

In clinical trials, it is important to understand and characterize disease and treatment response heterogeneity among patients so that precision medicine can particularly target certain subsets of patients, defined by baseline characteristics. Feature variables, such as demographic characteristics, genetic, genomic and environmental information, combined with a patient's survival outcome, can be used to explore such latent heterogeneity.

In the first project, we propose a mixture model to explore each patient's latent survival pattern, where the mixing probabilities for latent groups are modeled through a multinomial distribution. The Bayesian information criterion (BIC) is used for selecting the number of latent groups. Furthermore, we incorporate variable selection with the adaptive lasso into inference so that only a few feature variables will be selected to characterize the latent heterogeneity. We show that our adaptive lasso estimator has oracle properties when the number of parameters diverges with the sample size. The finite sample performance is evaluated by simulation studies under different scenarios, and the proposed method is illustrated by the data from a breast cancer clinical trial (IBCSG) and the data of the assay of free light chain.

In the second project, we develop a mixture survival tree model for direct risk classification. We assume that the patients can be classified into a pre-specified number of risk groups, where each group has distinct survival profile. Our proposed tree-based method is devised to estimate latent group membership using the Expectation Maximization (EM) algorithm. The observed data log-likelihood function is used as the splitting criterion in recursive partitioning. We examine the monotone likelihood property of the proposed algorithm. The finite sample performance is

evaluated by extensive simulation studies and the proposed method is illustrated by a case study in breast cancer.

In the third project, we study the unobserved heterogeneity in patient's treatment response. We consider a semi-parametric approach to directly classify patients into different latent subgroups where each subgroup of patients demonstrates a distinct average treatment effect. A random forest algorithm is developed to learn how the baseline covariates determine the unobserved heterogeneity in patients. In each individual tree, the EM algorithm is incorporated to handle the unobserved subgroup membership. The observed data log-likelihood function is used as the splitting criterion in recursive partition. A variable importance measurement is derived to facilitate identifying important features related to subgroup membership assignment. We evaluate the numeric performance of our proposed random forest model via extensive simulation studies and provide an application to a Phase III randomized clinical trial in patients with hematological malignancies.

To my family.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitudes to my advisors, Dr. Donglin Zeng and Dr. Joseph G. Ibrahim for their guidance, encouragement and support throughout my PhD study. Their enthusiasm for research always encourages me to explore unknown fields. Their guidance and support are invaluable to me. I'm proud of being one of their students.

I would like to express my sincere appreciation to my committee members, Dr. Jason J.Z. Liao, Dr. Xianming Tan and Dr. Quefeng Li for their inspiration, suggestion and help throughout the development of this dissertation. I would also like to thank my collaborators, Dr. Guanghan F. Liu and Dr. Guoqing Diao, for their instructions and constructive questions in our collaboration, and insightful comments and advice in the preparation of the manuscripts.

My genuine gratitudes also go to all other faculty members, students and staff in the Department of Biostatistics at University of North Carolina at Chapel Hill for providing the helpful resources and creating great academic enviroment.

Last but not least, I would like dedicate this work to my beloved family. Their unconditional support and endless love helped me get through tough times. To my dearest fiancé, thank you for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In the fight against diseases such as cancers, one drug hardly demonstrates benefits for a large population of patients. With patients' different survival experiences and/or an overall non-significant treatment effect observed in practice, an exploration in the overall population for potential heterogeneity is fundamental in clinical trials. Particularly, the discovery of subpopulations with different survival risks or distinct treatment responses enables the development of tailored therapy for a subgroup of patients. With the rapid growth in science and technology, more health-related data become available. The development of statistical learning methods to identify latent groups of patients and study the homogeneity of patients from the same group defined by individual characteristics attracts more attention. The goal of this dissertation is to discover latent subgroups using survival outcome and baseline characteristics. By the three research topics, we propose various approaches to assist the latent subgroup discovery under different scenarios.

First, we focus on variable selection procedure in the latent subgroup identification using survival outcome. The survival distribution is modeled parametrically. We assume that patients in each subgroup have different survival experience and the latent subgroup membership is determined by baseline covariates. The EM algorithm is applied to deal with the unobserved subgroup membership. Important variables are selected by introducing adaptive lasso penalty. We show the oracle property for our proposed estimator when the number of covariates diverges with the sample size. Simulation studies are used to evaluate the performance of our proposed methodology and our method is illustrated using two data examples.

Second, we explore the latent prognosis groups in a nonparametric manner. We develop a tree-based algorithm to explore different survival risk groups. We aim to provide a direct classification on patients to avoid concerns related to the lack of statistical power. More specifically,

we recursively partition the covariate space to learn the relationship between latent subgroup membership assignment and baseline covariates. The EM algorithm is embedded in each splitting and the observed data loglikelihood function is used as the splitting criterion. The survival risk classification is based on the posterior probabilities of a patient belonging to each latent subgroup, which is calculated along the tree growth. We examine the monotone likelihood property of the proposed algorithm. Extensive simulation studies are performed to evaluate our method. We apply the proposed method on a clinical trial of breast cancer.

Lastly, we extend the tree-based method for survival risk classification to the scenario that heterogeneous treatment effects are present in the data. The survival distribution is modeled semi-parametrically through a cox model. The latent subgroup are defined by treatment and baseline covariates. The differences among subgroups are due to patients' different treatment responses. A random forest model based on ensemble of individual trees is proposed to enhance the predictive performance. We derive a variable importance measurement to identify importance variables that are predictive of latent subgroup membership. We use simulations and an analysis on a phase III clinical trial in patients with hematological malignancies to illustrate the application of our proposed method.

The remainder of this dissertation is organized as follows. In Chapter 2, we introduce related concepts and works that are essential to the development of this dissertation. The aforementioned three research topics are covered in chapters 3, 4 and 5 respectively. Chapter 6 discusses extension and future research directions of the three topics. Technical details and references of the entire dissertation are followed.

CHAPTER 2: LITERATURE REVIEW

In this Chapter, we review related concepts and existing representative works that are essential to the development of this dissertation. In Section 2.1, we review a number of literature on the statistical methods to handle and identify heterogeneity with the time-to-event data using the mixture model framework. In Section 2.2, a large set of approaches in tackling high-dimensional data will be discussed. In Section 2.3, with the focus of tree-based methods, we introduce the main semi-parametric and non-parametric approaches in the context of survival outcomes. In Section 2.4, we go through the representative works that investigate heterogeneous treatment effects in the data.

2.1 Mixture model in Survival Analysis

Survival analysis is a popular research field in statistics, which models the expected duration of time associated with the occurrence of one or more events. It is widely used in cancer clinical research and drug/vaccine clinical trials. In the past several decades, it is of great interest to identify the unobserved heterogeneity in survival data, where the patients' heterogeneity may be associated with different survival profiles and/or different treatment responses. In the broad literature, finite mixture models is a popular direction to deal with the heterogeneity in the data.

In the field of survival analysis, a finite mixture model is directly applicable in the case that one parametric distribution cannot adequately describe the distribution of survival time. An example is from McGiffin et al. (1992), where the survival time of a patient after major cardiac surgery could be decomposed into three phases and each phase was associated different risk of death. A mixture model consisted of three components could be applied to model the distribution of the time to death.

Another case that the finite mixture model can be employed is related to competing risks where a patient is exposed to competing causes of failure. The mixture model can be regarded as

an alternative to Prentice et al. (1978). Instead of considering cause-specific hazard functions to characterize the joint distribution of the failure time and the type of failure, Larson and Dinse (1985) proposed a parametric mixture model to analyze competing risks data where the mixing parameters correspond to the marginal probabilities of various failure types. More specifically, they model the number of risk-specific failures through a multinomial distribution and the probability of failing from risk j is written via a logistic regression, given as

$$P_j(\boldsymbol{z}) = P(D = j|\boldsymbol{z}) = \frac{\exp(\mu_j + \boldsymbol{\pi_j}^T \boldsymbol{z})}{\sum_{j=1}^{J} \exp(\mu_j + \boldsymbol{\pi_j}^T \boldsymbol{z})},$$

where D indicates the type of failure and z denotes the vector of covariates. The survival function, given the type of failure j, is modelled as

$$Q_j(t|\boldsymbol{z}) = P(T > t|\boldsymbol{z}, D = j) = \exp\left\{ - \int_0^t \lambda_j(x) \exp(\boldsymbol{\beta_j}^T \boldsymbol{z}) dx \right\}$$

where $\lambda_j(x)$ is the null hazard function for failure j with covariates taking value of 0.

The finite mixture model can also be utilized to model the mixing proportion as a function of the covariates. For example, Farewell (1982) analyzed a toxicological experiment data using mixture models by assuming a fraction of long-term survivors, where a logistic function of covariates were used to model the mixing components (i.e., a long-term survivor vs. early death in Farewell (1982)).

More recently, mixture models are widely used to identify and analyze latent subgroups in the data. Altstein and Li (2013) studied a semiparametric accelerated failure time mixture model on a latent subgroup with time-to-event data in randomized clinical trials, where the unobservable membership in one arm of the clinical trial introduced the latency in the data. Shen and He (2015) performed a confirmatory statistical test to examine the existence of subgroups and considered a structured Logistic-Normal mixture model to identify a subgroup with enhanced treatment effect. Bussy et al. (2019) proposed a Quasi-Newton Expectation Maximization algorithm to detect patients subgroups based on discrete survival data. As technology advances, more machine learning approaches in analyzing survival data with heterogeneity emerged and enriches the fruitful

literature. Bennis et al. (2020) proposed a neural network architecture to estimate a finite mixture of two-parameter Weibull distributions with right-censored data.

2.2 Variable Selection

Baseline characteristics are often considered to be predictive of the latent subgroup membership. A large number of covariates, such as demographic characteristics, genetic, genomic and environmental information, may be involved in the procedure of latent subgroup identification. It is natural to believe that only a few covariates are truly predictive of latent subgroup membership. Therefore, variables selection in latent subgroup identification is necessary to allow the final model to possess good predictability and interpretability.

Various approaches of variable selection have been greatly discussed in the literature. The most classical way to deal with high-dimensional data is to construct statistical tests. Stepwise regression (Breaux, 1967) is the first model selection strategy. Its idea is simple and straightforward: the subset of important variables are selected by retrieving insignificant variables or adding significant variables based on statistical criterion. Later, many model-based statistics, including Mallow's C_p (Mallows, 1973), the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (Schwarz et al., 1978) and autometrics (Hendry and Richard, 1987), are developed to choose the best subset of variables among candidates.

In the last several decades, more advanced approaches, such as regularization approaches, have been developed and commonly used in practice. Methods in this category involve a penalty on parameters to introduce sparsity of covariates. The most common method is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). A L_1 type penalty is imposed to the negative log-likelihood function. Friedman et al. (2010) introduced the coordinate descent algorithm to successively minimize along coordinate directions to find the minimum of the objective function. Meinshausen and Bühlmann (2006) pointed out that it is possible that LASSO selects noise variables using an optimal shrinkage parameter. So many alternative methods have been developed, including the Bridge (Knight et al., 2000), the smoothly slipped absolute deviation penalty (SCAD) (Fan

and Li, 2001), the adaptive LASSO (Zou, 2006), the adaptive elastic-net (Zou and Zhang, 2009). These penalty functions are designed to impose the regression parameter estimates for unimportant variables to zero, as a result, excluding these unimportant variables from the model. Therefore, predicted values of the response of interest could be written as a function of a potentially smaller number of variables. Favorable theoretical properties, including the oracle properties, for these methods have been well established. The oracle properties refer to the consistency of selection and asymptotic normality, with the asymptotic covariance matrix being the same as that which would be obtained if we know the true underlying model. Thus, for large samples, oracle procedures perform as well as if the true underlying model were known in advance.

Another category of variables selection approaches is based on screening, such as the sure independence screening (Fan and Lv, 2008) and its extension in Fan et al. (2009). The methods in this category is very efficient in handling very high dimensional data. The choice of the ranking measure for screening methods plays a key role in reducing the set of candidate variables. To deliver computational efficient and powerful results, screening methods are often combined with other variable selection procedures: the most popular choice is regularization method in the literature.

In the context of unlabeled data or mixture models, variable selection is also greatly studied. Law et al. (2003, 2004) discussed feature selection in Gaussian mixture-based clustering and proposed two approaches. One is to estimate feature saliencies with the help of expectation-maximization algorithm, and the other is carried out by a backward search scheme on the basis of Koller and Sahami's mutual-information-based feature relevance criterion. Raftery and Dean (2006) investigated feature selection for model-based clustering by addressing the nested model comparison via approximate Bayes factors. Khalili and Chen (2007) proposed a class of penalty functions to be used for variable selection, which are counterparts of LASSO, HARD and SCAD in the context of finite mixture regression models, labelled as MIXLASSO, MIXHARD and MIXSCAD penalties respectively.

For data with time-to-event outcomes, Tibshirani (1997) proposed the LASSO method for variable selection and shrinkage in Cox proportional hazards model. Fan and Li (2002) generalized

the nonconcave penalized likelihood approach to the Cox proportional hazards model and the Cox proportional hazards frailty model. Liu et al. (2012) extended the adaptive LASSO approach to a Cox mixture cure model, which assumes that the subjects consist of two subpopulations: the cured one refers to subjects who never experience the event of interest, and the other one, by contrast, named ”non-cured”. The mixing probabilities are usually assumed to follow a logistic regression model. The authors state that a mixture cure model, in which a Cox proportional hazard is assumed in the latency, can be estimated iteratively in two parts: the Cox model and the logistic regression. Hence, the adaptive LASSO procedure can be easily applied in this context.

2.3 Survival trees

Numerous studies with time-to-event data arise in various research areas. Among these studies, the Cox proportional hazard regression model and its extensions are the most classical and widely used methods because they simply interpret the effect of covariates and are easily employed for inference. However, a specific link between the covariates and the response is required for the Cox proportional hazard model. Moreover, the interaction between covariates can be incorporated into such models but the functional form should be first specified by the user. Sometimes it is infeasible to impose a link function and specify the functional form of covariates to build the Cox proportional hazard model.

In this case, more flexible approaches are needed. As a result, more efforts have been devoted to develop tree-based methods by virtual of their nonparametric nature. Tree-based methods can automatically detect interactions based on recursive partition and offer great flexibility and interpretability. Therefore, survival trees and forests become popular alternatives to parametric and semi-parametric models. Additionally, decision trees can be regarded as another approach to discover subgroups and conduct variables selection. A number of regions of the covariate space (i.e. subgroups) are created by performing a series of binary splits. The regions of the covariate space contain individuals who are similar with respect to the outcome of interest, i.e., their survival profile and/or treatment responses. In this perspective, decision trees can naturally group subjects by their

survival experience and/or treatment response given some baseline covariates, so that subgroups can be easily discovered afterwards. In the meantime, these regions are usually defined using only a subset of the available variables.

Classification and regression tree (CART) (Breiman et al., 1984) is proposed to handle categorical or continuous response variables using a set of covariates. The tree partition is realized by recursively splitting the parent node into two child nodes based on some entropy measures of impurity, and the subjects in the same node will eventually have desirable similarity in terms of the outcome of interest. A pruning method to prevent overfitting is applied after the tree is fully grown according to a stopping criterion.

The survival tree (Ciampi et al., 1981; Marubini et al., 1983; Gordon and Olshen, 1985) was first developed, aiming to extend existing tree-based methods to the time-to-event data. There are a number of discussions focusing on the splitting criterion for tree partition. The basic idea is to maximize the within-node homogeneity and the between-node heterogeneity. Gordon and Olshen (1985) discussed the use of the logrank statistic and a parametric likelihood ratio statistic to measure how different between two child nodes. Many later works (Ciampi et al., 1986, 1987; Davis and Anderson, 1989; Ciampi et al., 1988; LeBlanc and Crowley, 1993) further studied these two splitting criteria for some specific models including the exponential model and the Cox proportional hazard model.

A new splitting criterion based on a node deviance measure between the log-likelihoods from a saturated model and a maximized model is introduced by LeBlanc and Crowley (1992). This work adopted CART algorithm to estimate the full likelihood of the learning sample for a tree $\mathcal{T}$,

$$L = \prod_{h \in \mathcal{N}_{\mathcal{T}}} \prod_{i \in S_h} \lambda_h(t_i)^{\delta_i} \exp(-\Lambda_h(t_i)),$$

where $\mathcal{N}_{\mathcal{T}}$ denotes the set of terminal nodes of tree $\mathcal{T}$; S_h is the set of observation labels, $\{i : x_i \in \mathcal{X}_h\}$, for observations in the region $\mathcal{X}_h$ corresponding to node h; t_i and δ_i are the observed time and the event status for individual i, respectively; and $\lambda_h(t)$ and $\Lambda_h(t)$ are the hazard and cumulative

hazard functions for node h. $\lambda_h(t)$ can be written as $\lambda_h(t) = \theta_h \lambda_0(t)$, where θ_h is a nonnegative parameter and $\lambda_0(t)$ is the baseline hazard. The Breslow estimator is used to approximate the baseline cumulative hazard function. They considered the full likelihood deviance to measure the goodness-of-fit of current tree. The deviance for node h is given by

$$R(h) = 2\left\{L_h(\text{saturated}) - L_h(\tilde{\theta}_h)\right\}$$

where $L_h(\text{saturated})$ is the log-likelihood for the saturated model and $L_h(\tilde{\theta}_h)$ is the maximized log-likelihood when the baseline cumulative hazard is known. Their algorithm maximized the reduction in deviance realized by the split by recursively partitioning the data.

In recent years, survival trees for time-to-event data still receive much attention. Molinaro et al. (2004) established a unified approach to construct and select trees with censoring. Their approach is driven by the choice of a loss function for the full (uncensored) data structure. The median survival tree based on L_1 loss function is investigated by Cho and Hong (2008). The use of integrated absolute difference between the two children nodes survival functions as the splitting criterion is discussed by Moradian et al. (2017). Numerous research extended the squared error loss in regression trees to survival data with censoring (Molinaro et al., 2004; Steingrimsson et al., 2016, 2019). Sun et al. (2019) incorporated the time-dependent receiver operating characteristics (ROC) curves into survival trees. With this method, the ROC curves is utilized to guide the tree-building algorithm and evaluate the performance of survival trees.

2.4 Subgroup Analysis

With the thriving of the biology, pharmacology and technology, more personalized medicine and targeted treatments are favorable. Subgroup analysis aims to reveal potential variation in treatment effect in different subgroups of individuals. It is of great interest to explore such heterogeneity and identify different subgroups of patients who respond to the treatment differently. So personalized treatment could be applied afterwards. Increasing discussions and attention have been received in

the area of subgroup analysis using statistical methods. Pocock et al. (2002) investigated the use of patient characteristics in clinical trials by surveying 50 reports of clinical trials. The difficulties in conducting subgroup analysis and the need for the appropriate statistical procedures in medical decision making are well discussed. Tanniou et al. (2016) defined four purposes for subgroup analyses in phase III clinical trials through a comprehensive review of 1857 papers in this field. They argued that subgroup analysis plays a fundamental role in the investigation of the consistency of treatment effects across subgroups, the exploration of the treatment effect across different subgroups within an overall non-significant trial, the evaluation of the safety profiles in a limited number of subgroups, and the establishment of the efficacy in the targeted subgroup.

A large number of studies focus on identifying and assessing the heterogeneity of treatment effect in clinical trials based on statistical hypothesis testing. For example, Song and Chi (2007) proposed a general statistical testing procedure, which offers the optimal power and strong control of the familywise Type I error rate. Their proposed framework is applicable to the cases that certain subgroups in clinical trials are already discovered. Michiels et al. (2011) studied treatment-effect-modifying-biomarkers in a phase III clinical trial with a survival endpoint to identify different treatment effects in subgroups that are associated with multiple specific biomarkers. Five different permutation test procedures were considered in their work to ascertain the existence of a subgroup of patients responding to treatment differently. Many other works, including Alosh and Huque (2009); Chen and Beckman (2009); Millen et al. (2012); Kovalchik et al. (2013); Krisam and Kieser (2014), are also devoted to develop statistical methodology for evaluating heterogeneous treatment effects in subgroups.

Unlike aforementioned methods within the scope of statistical testing, another research area in subgroup analysis, with the goal of exploring certain patterns within the overall population, has drawn wide attention for decades. This purpose can be formulated as a classification problem and established by extending the idea of recursive partitioning. One of the representative works is developed by Ciampi et al. (1995), who proposed tree-structure subgroup analysis using the Recursive Partition and amalgamation (RECPAM) algorithm. Their approach is to partition patients

into subgroups based on the similarity of their response to treatment. They considered a Cox model

$$\lambda(t|\boldsymbol{x_i}) = \lambda_0(t|s, \mathcal{T}) \exp\{\beta \cdot trt_i + \boldsymbol{\gamma}^T \cdot trt_i \cdot \boldsymbol{z_i}\}$$

to fit the data in the form of classification. In their model, $\lambda_0(t|s, \mathcal{T})$ is the baseline hazard function for the individuals in stratum s, where s is created by stratifying the terminal nodes of $\mathcal{T}$. $\boldsymbol{z_i}$ is a dummy vector, function of the predictor $\boldsymbol{x_i}$, indicating the membership to one of the classes other than the reference class. The class, or the prognosis group, is determined by $\boldsymbol{\gamma}$, the log-relative hazard with respect to the baseline, on the terminal nodes. Later, Negassa et al. (2005) investigated the model selection in tree-structured subgroup analysis based on RECPAM and proposed a two-stage computationally inexpensive model selection procedure.

Interaction trees (ITs), proposed by Su et al. (2008, 2009), show similar idea as Ciampi et al. (1995). ITs recursively partition the data with censored survival times into two subsets, aiming to obtain the greatest interaction with treatment. In other words, according to the Cox proportional hazard model

$$\lambda(t|\boldsymbol{x_i}) = \lambda_0(t) \exp\left\{\beta_1 \cdot trt_i + \beta_2 \cdot z_i^{(s)} + \beta_3 \cdot trt_i \cdot z_i^{(s)}\right\},$$

where $z_i^{(s)}$ is the indicator variable associated with split s, each split in ITs seeks for the greatest interaction effect with the treatment by testing the null hypothesis of $\beta_3 = 0$. Such hypothesis can be carried out by the partial likelihood ratio test. In such a way, ITs are capable of exploring subgroups with distinct treatment effects.

Model-based recursive partitioning algorithm (MOB) (Zeileis et al., 2008) can also be employed for subgroup identification. MOB constructs a tree structure where each node is associated with a single model and a fluctuation test for parameter instability is conducted to determine whether the partitioning should be further performed. To identify subgroups in the presence of treatment-subgroup interactions, Doove et al. (2014) states the use of MOB: one can set the model in the node

equal to a regression model of the treatment outcome against the treatment type and the partitioning is then applied on the baseline characteristics.

Over the past decade, more advanced methods have been well established. Dusseldorp et al. (2010) introduced Simultaneous Threshold Interaction Modeling Algorithm (STIMA) to automatically handle higher order interaction effects that can be included into a linear regression model. STIMA simultaneously estimates a multiple regression model and a tree model, where the higher order interaction effects can be carried out by a regression tree. According to Doove et al. (2014), one can set the first split in the regression tree to be made on the treatment variable for the task of subgroup identification when the interaction between treatment and subgroup is present and of interest.

Virtual Twins (Foster et al., 2011) identifies a subgroup of patients with an enhanced treatment effect in a randomized clinical trial. The idea, inspired by counterfactual models, consists of the prediction of response probabilities for treatment and control "twin" for each subject. A regression or classification tree is applied afterwards to find a small number of covariates that have strong association with treatment effect.

Subgroup Identification based on Differential Effect Search (SIDES) (Lipkovich et al., 2011) discovers multiple subgroups with enhanced treatment effects based on recursive partitioning. This method incorporates a treatment-by-split interaction in the splitting criterion and only searches within specific regions of the covariate space to generate subgroups. SIDEScreen (Lipkovich and Dmitrienko, 2014), the extension of SIDES, adds fixed and adaptive screens to screen out non-informative biomarkers. The second step is to search subgroups based on selected biomarkers.

Qualitative interaction tree (QUINT) (Dusseldorp and Van Mechelen, 2014) partitions patients into terminal nodes based on patient characteristics and the treatment responses to two alternative treatments are evaluated subsequently. Loh et al. (2015) came up with a regression tree method to identify subgroups with differential treatment effects. Qiu and Wang (2019) combined the estimation of optimal piecewise linear individualized treatment rules (ITRs) and subgroup identification by using a composite interaction tree (CITree). To achieve the simultaneous learning of optimal ITRs

and subgroups, CITree consists of the qualitative-interaction split and the quantitative split, where the qualitative split partitions patients into homogeneous subgroups of similar optimal ITR, and the quantitative split is designed to reduce ITR benefit heterogeneity.

Recently, another group of methods are proposed to study heterogeneous treatment effects using causal tree learning. The causal tree learning is leveraged by decision tree algorithm and aims to estimate heterogeneous treatment effects based on splitting observed individuals into groups. Athey and Imbens (2016) proposed causal trees to estimate heterogeneous causal effects in experimental and observational studies and provided valid inference for average treatment effects for each identified subpopulations. With the honest estimation proposed in Athey and Imbens (2016), one sample is utilized to choose partition, and another sample is used to estimate treatment effects for each subpopulation. Later, causal forest (Wager and Athey, 2018), extended from Breiman's random forest algorithm, are developed based on the work of Athey and Imbens (2016) to investigate treatment effect heterogeneity. The analysis of bias and consistency properties for causal forest are also discussed in their paper.

CHAPTER 3: INFERRING LATENT HETEROGENEITY USING MANY FEATURE VARIABLES WITH SURVIVAL OUTCOME

3.1 Introduction

A typical clinical trial is designed to test a drug/vaccine on a large and diverse group of patients and hopefully the One-Size-Fits-All approach is successful. The benefit for this approach is the quick availability of an effect drug/vaccine to the broadly targeted population with the unmet medical need. However, with much less low-hanging fruits available, it becomes more challenging to develop a blockbuster drug/vaccine that works for all study populations. Especially in more advanced and hard-to-treat diseases such as oncology, patients often present a heterogeneous survival experience, and their disease outcomes may be early death or spontaneous progression of the tumor followed by cure. This traditional one-size-fit-all approach may not be cost and time effective due to the high heterogeneity of the study population. As technology advances, more personal clinical, genetic, genomic, and environmental information and other baseline characteristic variables are available before the clinical study. Consequently, sponsors are looking into ways to conduct study in a more homogeneous subgroup with much higher probability of success to develop new medicines effectively. Thus, a challenging statistical problem with strong scientific/clinical interest in drug discovery and development is the identification of patient subgroups with different survival experience and potential treatment response heterogeneity. Recently, Liao and Liu (2019) demonstrated that many Kaplan Meier survival curves commonly seen in oncology trials can be reconstructed using a mixture of two or three parametric survival profiles. In other words, a disease population can be approximately decomposed into two or three latent groups with unique corresponding survival behavior in each latent subgroup.

In this chapter, to predict the latent subgroup membership for future individuals and identify the important variables that are predictive of latent subgroup membership for individuals with specific survival profiles, we propose a method for variable selection in latent subgroup identification with time-to-event data. More specifically, we model the survival distribution through a mixture of Weibull distributions, where each mixture represents a latent subgroup. The latent group membership is then modelled via a multinomial distribution that may vary with feature variables. To select important feature variables for characterizing the latent groups, the EM algorithm is first applied to obtain the initial maximum likelihood estimate, and then the adaptive lasso penalty is introduced for variable selection. We show that our proposed estimator enjoys the oracle property when the number of covariates diverges with the sample size.

The rest of this chapter is organized as follows. Section 3.2 details the proposed method for variable selection in latent subgroup identification for individuals with specific survival profiles. Theoretical results are provided in Section 3.3. Section 3.4 shows the finite sample performance of our proposed method via two simulation studies. Two real data examples in Section 3.5 demonstrate the applications of the proposed method.

3.2 Methodology

3.2.1 Model

We assume that the whole population consists of K different subgroups. Each group of patients will follow a specific survival profile. More specifically, we assume that the kth group has a survival distribution $S(t, \eta_{\mathbf{k}})$, which has a parametric form with unknown parameters $\eta_{\mathbf{k}}$, for $k = 1, \ldots, K$. In this paper, we assume that the survival outcome for each latent subgroup follows a Weibull distribution, which is a commonly used distribution in survival analysis due to its flexibility and reliability (Liao and Liu, 2019). The functional form of the Weibull distribution for the kth latent subgroup is givens by $S(t, \eta_{\mathbf{k}}) = \exp\left\{-(\frac{\mathbf{t}}{\lambda_{\mathbf{k}}})^{\kappa_{\mathbf{k}}}\right\}$, where $\eta_{\mathbf{k}} = (\kappa_{\mathbf{k}}, \lambda_{\mathbf{k}})^{\mathbf{T}}$, κ_k is the shape parameter and λ_k is the scale parameter.

We let T denote the time to event and X denote all the baseline covariates, which the number of baseline covariates could be large. To classify each patient into one of the survival groups using the baseline covariates

X (X contains constant 1), we introduce a latent group membership B and assume

$$P(T > t|B = k, X) = S(t, \eta_{\mathbf{k}})$$

and

$$P(B = k|X) = \frac{\exp\{\beta_{\mathbf{k}}^{\mathbf{T}}\mathbf{X}\}}{\sum_{k=1}^{K}\exp\{\beta_{\mathbf{k}}^{\mathbf{T}}\mathbf{X}\}} = \pi_k(X, \beta)$$

for $k = 1, \ldots, K$, where $\beta_{\mathbf{1}} = \mathbf{0}$ and $\beta_{\mathbf{2}}, \ldots, \beta_{\mathbf{K}}$ are unknown parameters. Therefore, the latent group membership determines which group the patient should belong to and this membership depends on the baseline covariates through a multinomial distribution. Clearly, the proposed model implies that the marginal survival distribution for T takes a mixture form:

$$P(T > t|X) = \sum_{k=1}^{K} S(t, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X}, \beta),$$

where $\beta = (\beta_{\mathbf{2}}^{\mathbf{T}}, \ldots, \beta_{\mathbf{K}}^{\mathbf{T}})^{\mathbf{T}}$. To conduct a future trial, for any new patient with baseline covariates $X = \mathbf{x}$, we then classify this patient into group k with maximal value $\beta_{\mathbf{k}}^{\mathbf{T}}\mathbf{x}$, i.e., the most likely group membership.

3.2.2 Initial Estimate

Suppose that we have right-censored observations from n i.i.d patients, denoted by

$$\{Y_i = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), X_i, i = 1, \ldots, n\},$$

where C_i is the censoring time. Assuming that the censoring time is independent of T_i given X_i, we obtain the observed data log-likelihood function as

$$\begin{aligned} l_{n,obs}(\theta) = \sum_{i=1}^{n} \Bigg[& \Delta_i \log \left\{ \sum_{k=1}^{K} f(Y_i, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X_i}; \beta) \right\} \\ & + (1 - \Delta_i) \log \left\{ \sum_{k=1}^{K} S(Y_i, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X_i}; \beta) \right\} \Bigg], \end{aligned} \tag{3.1}$$

where $\theta = (\eta^{\mathbf{T}}, \beta^{\mathbf{T}})^{\mathbf{T}}$, $\eta^{\mathbf{T}} = (\eta_{\mathbf{1}}^{\mathbf{T}}, \ldots, \eta_{\mathbf{K}}^{\mathbf{T}})^{\mathbf{T}}$, and $f(t, \eta_{\mathbf{k}}) = -\mathbf{S}'(\mathbf{t}, \eta_{\mathbf{k}})$.

To estimate β, we introduce $B_1, \ldots, B_n$ as the latent group membership for each subject and use the EM algorithm to compute the maximum likelihood estimators, treating the B's as missing data. In the E-step, at the kth iteration, we compute the expected log-likelihood based on the current estimates of all parameters, conditional on the observed data, which is equivalent to calculating the posterior probability of $B_i = k$ given the observed data for $k = 1, \ldots, K$ and $i = 1, \ldots, n$. More specifically, this posterior probability is

$$q_{ik} = \frac{f(Y_i, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X_i}; \beta)}{\sum_{k=1}^{K} f(Y_i, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X_i}; \beta)}$$

if $\Delta_i = 1$, and it is

$$q_{ik} = \frac{S(Y_i, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X_i}; \beta)}{\sum_{k=1}^{K} S(Y_i, \eta_{\mathbf{k}})\pi_{\mathbf{k}}(\mathbf{X_i}; \beta)}$$

if $\Delta_i = 0$. In the M-step, we compute the estimates that maximize the expected log-likelihood obtained in the E-step,

$$\begin{aligned} l_n(\eta, \beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}[&\Delta_i \log\left(f(Y_i, \eta_{\mathbf{k}})\right) \\ &+ (1 - \Delta_i)\log\left(S(Y_i, \eta_{\mathbf{k}})\right) + \log\left(\pi_k(X_i; \beta)\right)]. \end{aligned} \tag{3.2}$$

To estimate the survival distribution parameter η, we implement the Newton-Raphson algorithm to update the estimate based on the expected log-likelihood function (3.2). The expected log-likelihood function (3.2) is essentially a weighted multinomial regression. To obtain the maximum likelihood estimate $\tilde{\beta}$, for each iteration, we apply a one-step Newton-Raphson in the M-step to update the estimate. After convergence, we obtain the maximum likelihood estimates $\tilde{\eta}$ and $\tilde{\beta}$. It is easy to see that the expected log-likelihood function (3.2) in the M-step increases at each iteration, which implies that the algorithm is guaranteed to converge and will stay unchanged once converged.

To determine the best number of latent subgroups in the data, we consider several choices of the number of latent subgroups. For each potential number of latent subgroups, we apply a similar procedure as stated in this section to obtain the initial estimates. The value of the log-likelihood based on the initial estimates is

computed afterwards. The BIC, as suggested by Nylund et al. (2007), is then calculated to determine the best number of latent subgroups for the data Nylund et al. (2007) evaluated the performance of several information criteria for correctly identifying the number of groups. The performance of BIC for determining the best number of latent subgroups is also evaluated in Section 3.4.

3.2.3 Variable Selection for Latent Groups

The objective function (3.2) in the M-step is essentially a weighted multinomial regression, with weights being the posterior probability of $B_i = k$ given the observed data for $k = 1, \ldots, K$ and $i = 1, \ldots, n$. We use this objective function to accommodate penalties for variable selection. Because of the strict concavity of the objective function (3.2), we can derive nice theoretical properties for the estimator after variable selection.

Among many penalty functions, we apply the convex adaptive lasso penalty to the objective function (3.2). The weight for each coefficient in the adaptive lasso penalty is related to the importance of the corresponding covariate and helps to adaptively penalize each coefficient by tuning each coefficient with a different parameter. Zou (2006) shows that the adaptive lasso enjoys the oracle properties by inflating the weights for zero-coefficient covariates and enables the weights of nonzero-coefficient covariates to converge to a finite constant. The data-dependent adapting weights can be the reciprocal of any consistent estimator of β (Zou, 2006). Here we consider the maximum likelihood estimator $\tilde{\beta}$. The penalized objective function becomes

$$- l_n(\tilde{\eta}, \beta) + \lambda \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{|\beta_{kj}|}{|\tilde{\beta}_{kj}|^{\gamma}}, \tag{3.3}$$

where γ is a prespecified positive constant and the commonly used value is $\gamma = 1$, and $\beta = (\beta_{11}, \beta_{12}, \ldots, \beta_{1d}, \beta_{21}, \ldots, \beta_{Kd})^T$. Here, we do not introduce a penalty on η, and $\tilde{\eta}$ is the maximum likelihood estimator. Hence, minimizing (3.3) is equivalent to applying the adaptive lasso penalty to a weighted multinomial regression.

To obtain the adaptive lasso estimates $\hat{\beta}$, we minimize the penalized objective function (3.3) via a two-step strategy. The first step is to calculate the maximum likelihood estimates $(\tilde{\eta}, \tilde{\beta})$ that optimize (3.2) by an iterative Newton-Raphson update. Denote $\theta = (\eta, \beta)$. Define the gradient vector $\nabla l_n(\theta) = \frac{\partial l_n(\theta)}{\partial \theta}$ and the Hessian matrix $\nabla^2 l_n(\theta) = \frac{\partial^2 l_n(\theta)}{\partial \theta \partial \theta^T}$. The Newton-Raphson update is

$$\theta^{(t+1)} = \theta^{(t)} - \left(\nabla^2 l_n(\theta)|_{\theta=\theta^{(t)}}\right)^{-1} \nabla l_n(\theta)|_{\theta=\theta^{(t)}}. \tag{3.4}$$

The second step is to obtain the adaptive lasso estimates $\hat{\beta}$ by minimizing (3.3) via a coordinate descent algorithm, where the coefficients are iterated over to minimize (3.3).

Hence, to minimize the penalized objective function (3.3) for any fixed γ, we use the following procedure.

Step 1. Use the EM algorithm and the Newton-Raphson update (3.4) to compute the maximum likelihood estimates $\tilde{\eta}$ and $\tilde{\beta}$.

Step 2. Calculate the weights in the adaptive lasso penalty, $\tilde{w}_i$ for $i = 1, \ldots, n$, by using $\tilde{\beta}$.

Step 3. Compute the weights, q_{ik} for $i = 1, 2, \ldots, n$, and $k = 1, 2, \ldots, K$, in the weighted multinomial regression $l_n(\tilde{\eta}, \beta)$ by using $\tilde{\eta}$ and $\tilde{\beta}$.

Step 4. Apply the coordinate descent algorithm to minimize the penalized objective function (3.3) until the convergence criterion is met.

In Step 3, the weights in the weighted multinomial regression are obtained by plugging in the estimates $\tilde{\eta}$ and $\tilde{\beta}$, since $\tilde{\eta}$ and $\tilde{\beta}$ are consistent maximum likelihood estimates, and the weights are fairly close to the true weights, which is shown in the Appendix. The minimization in Step 4 is based on the coordinate descent algorithm, which can be implemented via a statistical package such as *glmnet* in R.

To select the data-dependent tuning parameter λ in the proposed algorithm, we use V-fold cross validation. We consider λ from a set of grid points and partition the data into V subsets with equal size. For each point λ, we compute the coefficients using $V - 1$ subsets and obtain the deviance residual on the Vth subset by using these coefficients V times. Averaging over V deviance residuals, we have an average deviance residual associated with one point of λ. We then select among the average deviance residuals and have the best choice for the tuning parameter λ that yields the smallest average deviance residual. After variable selection, we reapply the EM algorithm and maximize (3.1) by including the selected important covariates. We then classify patients to their most likely latent subgroup based on the post-selection maximum likelihood estimates.

3.3 Theoretical Properties

In this section, we describe the asymptotic properties of our estimators when the number of parameters grows with the sample size. With a slight abuse of notation, we write $\beta_{\mathbf{n}} = (\beta_{\mathbf{n1}}^{\mathbf{T}}, \ldots, \beta_{\mathbf{nK}}^{\mathbf{T}})^{\mathbf{T}} = (\beta_{\mathbf{n1}}, \ldots, \beta_{\mathbf{np_n}})^{\mathbf{T}}$, where p_n is the number of variables, and $\theta_{\mathbf{n}} = (\eta^{\mathbf{T}}, \beta_{\mathbf{n}}^{\mathbf{T}})^{\mathbf{T}}$. We consider the penalized

objective function based on n samples,

$$Q_n(\theta_{\mathbf{n}}) = \mathbf{l_n}(\theta_{\mathbf{n}}) - \mathbf{n}\lambda_{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{p_n}} |\beta_{\mathbf{nj}}|/|\tilde{\beta}_{\mathbf{nj}}|^{\gamma}.$$

Denote the true values of $\theta_{\mathbf{n}}$ by $\theta_{\mathbf{n0}}$. We write $\theta_{\mathbf{n0}}$ as $(\eta^{\mathbf{T}}, \beta_{\mathbf{n10}}^{\mathbf{T}}, \beta_{\mathbf{n20}}^{\mathbf{T}})^{\mathbf{T}}$, where

$$\beta_{\mathbf{n10}} = (\beta_{\mathbf{n10}}, \beta_{\mathbf{n20}}, \ldots, \beta_{\mathbf{nq0}})^{\mathbf{T}}$$

consists of all q nonzero components and

$$\beta_{\mathbf{n20}} = (\beta_{\mathbf{n(q+1)0}}, \beta_{\mathbf{n(q+2)0}}, \ldots, \beta_{\mathbf{np_n0}})^{\mathbf{T}}$$

consists of the remaining zero components. Correspondingly, we have the adaptive lasso estimator $\hat{\theta}_{\mathbf{n}} = (\hat{\eta}^{\mathbf{T}}, \hat{\beta}_{\mathbf{n1}}^{\mathbf{T}}, \hat{\beta}_{\mathbf{n2}}^{\mathbf{T}})^{\mathbf{T}}$.

We require the following regularity conditions.

(C1) The function $S(t, \eta_{\mathbf{k}})$ for $k = 1, 2, \ldots, K$ is non-increasing and continuously differentiable.

(C2) Let $g(X_i, Y_i, \Delta_i, \theta_{\mathbf{n}})$ Denote the probability density for observation $\{X_i, Y_i, \Delta_i\}$, for $i = 1, 2, \ldots, n$. The observations $\{X_i, Y_i, \Delta_i, i = 1, 2, \ldots, n\}$ are independent and identically distributed. Let $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the minimum and maximum eigenvalues of a positive definite matrix A, respectively. Assume that, for all i, the Fisher information matrix

$$I_n(\theta_{\mathbf{n}}) = \mathbf{E}\left[\left(\frac{\partial \log \mathbf{g}(\mathbf{X_i}, \mathbf{Y_i}, \mathbf{\Delta_i}, \theta_{\mathbf{n}})}{\partial \theta_{\mathbf{n}}}\right)\left(\frac{\partial \log \mathbf{g}(\mathbf{X_i}, \mathbf{Y_i}, \mathbf{\Delta_i}, \theta_{\mathbf{n}})}{\partial \theta_{\mathbf{n}}}\right)^{\mathbf{T}}\right]$$

satisfies

$$C_1 \leq \lambda_{min}\{I_n(\theta_{\mathbf{n}})\} \leq \lambda_{\mathbf{max}}\{\mathbf{I_n}(\theta_{\mathbf{n}})\} \leq \mathbf{C_2}$$

and, for $j, l = 1, 2, \ldots, p_n$,

$$E\left[\left(\frac{\partial \log g(X_i, Y_i, \Delta_i, \theta_{\mathbf{n}})}{\partial \eta}\right)^T \left(\frac{\partial \log g(X_i, Y_i, \Delta_1, \theta_{\mathbf{n}})}{\partial \eta}\right)\right]^2 \leq C_3$$

and

$$E\left[\frac{\partial \log g(X_i, Y_i, \Delta_i, \theta_{\mathbf{n}})}{\partial \beta_{nj}} \frac{\partial \log g(X_i, Y_i, \Delta_i, \theta_{\mathbf{n}})}{\partial \beta_{nl}}\right]^2 \leq C_4,$$

where C_1, C_2, C_3 and C_4 are positive constants.

(C3) $\theta_{\mathbf{n0}}$ is contained in a large enough open set. For all $\theta_{\mathbf{n}}$ within this open set, the third derivatives of $g(X_i, Y_i, \Delta_i, \theta_{\mathbf{n}})$ with respect to $\beta_{\mathbf{n}}$ satisfy

$$\left|\frac{\partial^3 \log g(X_i, Y_i, \Delta_i, \theta_{\mathbf{n}})}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nm}}\right| \leq M_{njlm}(X_i, Y_i, \Delta_i)$$

and

$$E[M^2_{njlm}(X_i, Y_i, \Delta_i)] \leq C_5, \text{ where } C_5 \text{ is a positive constant}$$

for $j, l, m = 1, 2, \ldots, p_n$.

(C4) Assume that

$$\min_{1 \leq j \leq q} \frac{|\beta_{nj0}|}{\lambda_n} \to \infty, \text{ as } n \to \infty.$$

Condition (C1) requires $S(t, \eta_{\mathbf{k}}),\ \mathbf{k} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{K}$ to be a valid survival distribution. Conditions (C2) and (C3) are similar to conditions (F) and (G) in Fan et al. (2004), which assume that the likelihood function has reasonably good behavior. Condition (C4) is used to establish the oracle property of the adaptive lasso estimator and already implicitly assumed in a finite dimensional setting. This condition is exactly condition (H) in Fan et al. (2004), which allows nonzero coefficients to vanish and can be distinguished at a rate by the penalized likelihood.

Under conditions (C1) - (C4), we have the following asymptotic results for our estimators.

Theorem 3.1. *Denote the maximum likelihood estimates of $l_{n,obs}(\theta_{\mathbf{n}})$ by $\tilde{\theta}_n$, where*

$$l_{n,obs}(\theta_{\mathbf{n}}) = \sum_{i=1}^{n} \left[\Delta_i \log\left\{\sum_{k=1}^{K} f(Y_i, \eta_{\mathbf{k}}) \pi_{\mathbf{k}}(\mathbf{X_i}; \beta_{\mathbf{n}})\right\} + (1 - \Delta_i) \log\left\{\sum_{k=1}^{K} S(Y_i, \eta_{\mathbf{k}}) \pi_{\mathbf{k}}(\mathbf{X_i}; \beta_{\mathbf{n}})\right\}\right]$$

If $p_n^4/n \to 0$ *as* $n \to \infty$, *then* $||\tilde{\theta}_n - \theta_{n0}|| = O_p(\sqrt{p_n} n^{-1/2})$

Theorem 3.2. *If* $\sqrt{np_n}\lambda_n = O(1)$ *and* $p_n^4/n \to 0$ *as* $n \to \infty$, *then there is a unique maximizer* $\hat{\theta}_n$ *of* $Q_n(\theta_n)$ *such that* $||\hat{\theta}_n - \theta_{n0}|| = O_p\{\sqrt{p_n}(n^{-1/2})\}$.

Finally, we provide the asymptotic distribution of the adaptive lasso estimator. We let

$$\mathbf{b_n} = \{\mathbf{0}, \ldots, \mathbf{0}, \lambda_\mathbf{n}\text{sign}(\beta_\mathbf{n10})/|\tilde{\beta}_\mathbf{n1}|^\gamma, \ldots, \lambda_\mathbf{n}\text{sign}(\beta_\mathbf{nq0})/|\tilde{\beta}_\mathbf{nq}|^\gamma\}^\mathbf{T},$$

$\theta_{n1} = (\eta^T, \beta_{n1}^T, 0^T)^T$ and $\theta_{n10} = (\eta_0^T, \beta_{n10}^T, 0^T)^T$. Let s be the number of parameters for the survival distributions of the K latent subgroups. The first s zeros contained in $\mathbf{b_n}$ are due to the fact that we do not penalize the parameters of the survival distributions.

Theorem 3.3. *If* $n\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$ *and* $p_n^5/n \to 0$ *as* $n \to \infty$, *then under Theorem 1, the adaptive lasso estimator* $\hat{\theta}_n$ *has the following properties:*

(i) $\hat{\beta}_{n2} = 0$ *with probability tending to 1;*

(ii)

$$\sqrt{n}A_n I_n^{-1/2}(\theta_{n10})\{I_n(\theta_{n10})\}[\hat{\theta}_{n1} - \theta_{n10} + \{I_n(\theta_{n10})\}^{-1}\mathbf{b_n}] \longrightarrow_\mathbf{D} \mathbf{N}(\mathbf{0}, \mathbf{G})$$

where A_n *is a* $r \times (s+q)$ *matrix such that* $A_n A_n^T \to G$, *and* G *is a* $r \times r$ *non-negative symmetric matrix.*

One key to the proofs is to obtain a uniform approximation rate for the weights in the expression of $Q_n(\theta_\mathbf{n})$. For this, we use the result established in Theorem 1. The proofs of Theorems 2 and 3 then follow the standard arguments in variable selection for parametric models, including the existence of the local maximum in a neighborhood of the true parameters and verification of the fact that the oracle estimator attains this local maximum, but with careful verification of certain approximation rates in terms of p_n. The details of the proof are given in the Appendix. The theoretical properties for the post-selection estimator, that is, the maximum likelihood estimator of selected important variables after refitting the model without the adaptive lasso penalty, could be easily obtained. Under Theorem 3, the probability that adaptive lasso estimator of unimportant variables does not equal to zero tends to 0. Therefore, the post-selection estimator has the same asymptotic distribution as the adaptive lasso estimator of important variables, which is stated in Theorem 3.

3.4 Simulation Studies

We conduct the simulation study that assumes two latent subgroups exist. We consider 10, 30 and 50 covariates in the regression model and only a few of covariates have nonzero effects. The covariates $X = (X_1, X_2, \ldots, X_p)$, where $p = 10, 30, 50$, are generated from standard normal distribution with moderate correlations. Time to event data for each latent subgroup follow a different Weibull distribution with scale parameter λ and shape parameter κ. The censoring time is generated from an exponential distribution, where the mean is calibrated by a prespecified censoring rate of 10%.

The true values of the scale parameters (i.e., λ_1 and λ_2) of the Weibull distributions for two latent subgroups are set to be 1 and 4.5 respectively, and the true values of the shape parameters (i.e., κ_1 and κ_2) are 1 and 3 for two latent subgroups respectively. Around 40% of the individuals belong to latent subgroup 1 with a 2-year survival probability of 13.5%. Sixty percent of the individuals are in latent subgroup 2 and have a 2-year survival probability of 91.5%. The subgroup-specific survival curves are illustrated in Figure 3.1. The true β associated with latent subgroup membership is calibrated such that the true proportions of the two subgroups are 40% and 60%, respectively. Three scenarios in the simulation study are described below. The sensitivity analysis that evaluates the proposed model when the link function of the multinomial distribution for the latent subgroup membership is nonlinear is included in the Appendix.

Scenario 1. 10 covariates are independently generated from a standard normal distribution, and first three of them are important covariates. Subgroup 1 is regarded as the reference group and $\beta_{\mathbf{1}}$ is set to $\mathbf{0}$.

$$\beta = \begin{pmatrix} \beta_{\mathbf{1}} \\ \beta_{\mathbf{2}} \end{pmatrix} = \begin{pmatrix} 0, & 0, & 0, & 0, & 0, & \ldots, 0 \\ 0.4, & 0.2, & -0.6, & -0.3, & 0, & \ldots, 0 \end{pmatrix}$$

Scenario 2. 30 covariates are generated from standard normal distribution, and first eight of them have nonzero effects. Correlations between X_1 and X_2, X_3 and X_4, X_7 and X_{10} are set to be $0.2, 0.3$ and 0.2, respectively.

$$\beta = \begin{pmatrix} \beta_{\mathbf{1}} \\ \beta_{\mathbf{2}} \end{pmatrix} = \begin{pmatrix} 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & 0, & \ldots, 0 \\ 0.4, & 0.2, & -0.6, & -0.3, & 0.5, & -0.5, & 0.7, & -0.7, & 0.5, & 0, & \ldots, 0 \end{pmatrix}$$

Scenario 3. 50 covariates are generated from standard normal distribution, and first eight of them have nonzero effects. Correlations between X_1 and X_2, X_3 and X_4, X_7 and X_{10} are set to be $0.2, 0.3$ and 0.2, respectively. True values of regression coefficients for important variables are set to be the same as in scenario 2.

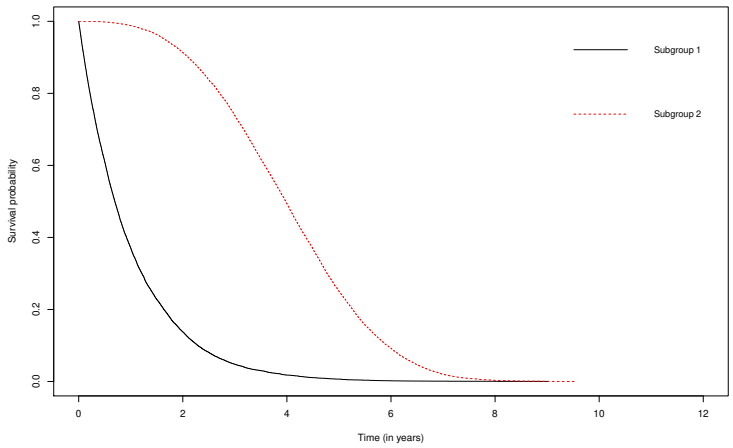


Figure 3.1: The true survival curves in the simulation study.

To implement the EM algorithm to obtain the maximum likelihood estimators of the β's and the survival distribution parameters κ_k, λ_k for $k = 1, 2, \ldots, K$, the stopping criteria for EM is $|l_{obs}(\theta^{(k+1)}) - l_{obs}(\theta^{(k)})| < 10^{-4}$, where $\theta = (\kappa_1, \lambda_1, \ldots, \kappa_k, \lambda_k, \beta^T)^T$. For γ in the adaptive lasso penalty, we use $\gamma = 1$ for all simulation studies. For each simulated dataset, we first identify the number of latent subgroups by applying our method for estimation and calculating BIC. Once the number of latent subgroups is determined, we use the EM algorithm to obtain the maximum likelihood estimates and then implement the adaptive lasso procedure to perform variable selection. We consider the grid $2^{-16}, 2^{-15}, \ldots, 2^{15}, 2^{16}$ for the tuning parameter λ, and report the results that yield the smallest value of average deviance residual. After variable selection, the EM algorithm is reapplied to the models with only selected covariates. We repeat the simulation 1000 times and consider sample sizes of $n = 300, 1000$ and 3000.

We first calculate BIC for models assuming no latent subgroups, two, and three latent subgroups, in datasets that truly consisted of two latent subgroups. BIC suggests that approximately 100% of the datasets with the sample size of 1000 consist of two latent groups. When the sample size increases to 3000, all the datasets are correctly detected consisting of two latent subgroups by the BIC criterion.

Table 3.1 summarizes the prediction accuracy, along with standard errors, for models without and after variable selection for all three scenarios, and also reports the average number of correct and incorrect zero coefficients and corresponding standard errors. The prediction accuracy is calculated by applying the decision rule obtained from the training set to a validation set with sample size $10,000$. Compared to the optimal accuracy rate, our method performs well for all three scenarios, especially for models with only selected important covariates. The prediction accuracy is approaching to the optimal accuracy rate as the sample size increases. When the number of covariates increases, our method also works well in terms of prediction accuracy and variable selection results. The optimal accuracy rate is $1-$Bayes error rate, where the Bayes error rate is calculated via the formula $1 - E\left(\max_k P(B = k|X)\right)$, is the lowest possible test error rate. For scenario 1, when the sample size is 300, important variables are correctly selected in approximately 80% of the datasets, and unimportant variables are selected in approximately 15% of the datasets. As the sample size increases to 1000, important variables are identified in over 99% of the datasets. Meanwhile, the ability to shrink zero coefficients to zero is also improved: the rate of incorrectly selecting unimportant variables is below 5%. For scenario 2 when the number of covariates increases to 30, important variables can be correctly identified in around 80% of the datasets when the sample size is 300, while the unimportant variables are selected in around 32% of the datasets. The ability to identify important variables and shrink zero coefficients to zero is improved when the sample size grows to 1000: important covariates can be picked out in approximately 95% of the datasets, and our method selects unimportant variables in only 5% of the datasets. For scenario 3 with 50 covariates, when the sample size is 300, around 80% of the datasets can correctly distinguish important variables and the rate of incorrectly selecting unimportant variables is 35%. As the sample size increases to 1000, the rates of identifying important variables and selecting unimportant variables are 80% and 11%, respectively. When the sample size further grows to 3000, over 99% of the datasets can correctly recognize important variables and the rate of incorrectly selecting unimportant variables decreases to around 1%. Table 3.2 reports the accuracy of nonzero coefficient post-selection estimates, their standard errors and coverage probabilities for nominal 95% confidence intervals for scenario 1. Due to the limited space, we report results for scenario 2 and 3 in the Appendix. To obtain the standard errors for the maximum likelihood estimates, we use the Louis formula (Louis, 1982) because the latent group membership is treated as missing data in our method. For these three scenarios, we observe similar results: the post-selection estimates are slightly biased on small samples and the bias can be reduced by increasing the

sample size; the 95% confidence intervals for the post-selection estimators based on the estimated coefficients and standard errors have accurate coverage for the true parameters.

Table 3.1: Results from the simulation study with 2 latent groups

	Accuracy (SE)		Comparison	
	without	after	Group 2 vs. 1	
N	variable selection	variable selection	Corr. (SE)	Incorr. (SE)
Scenario 1: 10 independent covariates, 3 of them are important. The optimal accuracy rate is 0.648.				
300	0.617 (0.019)	0.619 (0.030)	5.70 (1.754)	0.63 (0.876)
1000	0.636 (0.009)	0.640 (0.010)	6.53 (0.946)	0.08 (0.278)
3000	0.643 (0.006)	0.645 (0.015)	6.90 (0.326)	0.00 (0.045)
Scenario 2: 30 covariates with moderate correlations, 8 of them are important. The optimal accuracy rate is 0.732.				
300	0.675 (0.020)	0.680 (0.029)	13.47 (6.574)	0.86 (1.146)
1000	0.716 (0.007)	0.724 (0.008)	19.71 (3.245)	0.24 (0.459)
3000	0.728 (0.005)	0.731 (0.005)	21.41 (1.256)	0.03 (0.159)
Scenario 3: 50 covariates with moderate correlations, 8 of them are important. The optimal accuracy rate is 0.732.				
300	0.653 (0.019)	0.666 (0.030)	27.39 (10.290)	0.84 (1.178)
1000	0.705 (0.008)	0.720 (0.010)	36.45 (6.969)	0.22 (0.430)
3000	0.724 (0.005)	0.731 (0.005)	40.86 (2.783)	0.03 (0.179)

Note. Each column corresponds to prediction accuracy and standard errors, average number of correct (Corr.) and incorrect (Incorr.) zero coefficients and standard errors from 1000 simulated datasets.

Table 3.2: Maximum likelihood Estimates after variable selection, their standard errors, and coverage probabilities for nominal 95% confidence intervals from the simulation study with 2 latent groups

N	Parameter	Bias	SE	SEE	CP
300	k_1	0.035	0.113	0.109	0.954
	λ_1	0.015	0.274	0.254	0.852
	k_2	0.096	0.422	0.354	0.915
	λ_2	-0.013	0.210	0.190	0.902
	β_0	-0.003	0.350	0.318	0.901
	β_1	-0.033	0.186	0.110	0.605
	β_2	-0.002	0.277	0.172	0.866
	β_3	0.005	0.202	0.145	0.807
1000	k_1	0.013	0.062	0.061	0.946
	λ_1	-0.000	0.166	0.160	0.890
	k_2	0.007	0.203	0.190	0.931
	λ_2	-0.015	0.109	0.109	0.940
	β_0	0.010	0.195	0.187	0.916
	β_1	-0.008	0.101	0.080	0.883
	β_2	-0.017	0.101	0.097	0.949
	β_3	-0.006	0.088	0.089	0.962
3000	k_1	0.004	0.036	0.036	0.942
	λ_1	0.002	0.099	0.099	0.925
	k_2	0.006	0.110	0.111	0.952
	λ_2	-0.003	0.065	0.064	0.949
	β_0	0.002	0.111	0.112	0.939
	β_1	0.000	0.049	0.050	0.954
	β_2	-0.007	0.056	0.055	0.948
	β_3	-0.005	0.051	0.051	0.947

Note: SE, standard error; SEE, mean of standard error estimator; CP, coverage probability for nominal 95% confidence interval.

3.5 Real Data Application

3.5.1 Application to IBCSG Data

We apply the proposed methodology to data from a breast cancer clinical trial to study the potential heterogeneity of patients in terms of their survival outcomes and investigate important variables that are associated with such heterogeneity. The data was collected from a large clinical trial, IBCSG Trial VI (Colleoni et al., 2002), in premenopausal women with node-positive breast cancer to study both the duration of adjuvant chemotherapy and the reintroduction of delayed chemotherapy. Patients were randomized in a two by two factorial design to receive the following: (A) cyclophosphamide, methotrexate, and fluorouracil

(CMF) for six consecutive cycles (CMF*6); (B) CMF*6 plus three single cycles of reintroduction CMF; (C) CMF*3; and (D) CMF*3 plus three single cycles of reintroduction CMF. The patients' quality of life (QOL) was also measured at baseline and was hypothesized to contain prognostic information and reflect breast cancer progression. Four aspects of QOL, including physical well-being, mood, appetite and perceived coping, were assessed by a self-assessment QOL questionnaire. In addition to treatment effects and patients' QOL, disease-free survival (median follow-up of 7.47 years, rescaled to $[0, 1]$), event status, age at baseline, estrogen receptor (ER) status (1=positive, 0=negative) and the number of positive nodes of the tumor (i.e., node group, 1=number of positive nodes > 4, 0=else) are also considered in the data. After excluding missing values, data are available for 962 patients. The median follow-up for disease free survival (DFS) is 7.47 years and the event rate is around 45%. We rescale the DFS to $[0, 1]$ and standardize continuous variables such as age and the four measures of QOL for computation.

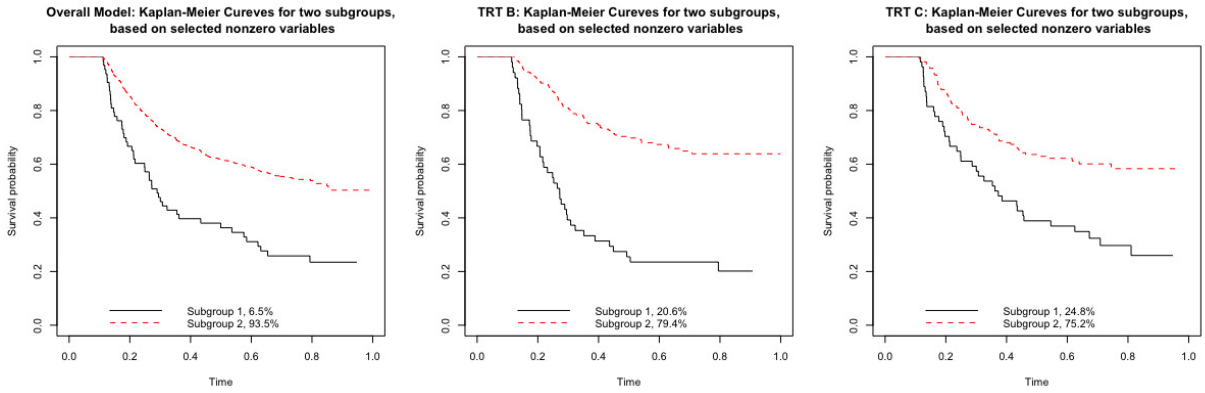
The results of variable selection indicate that treatment does not have a significant effect on the latent subgroup membership assignment. To further explore the heterogeneity of patients under different therapeutic procedures, we apply our method to the datasets of patients under each treatment. According to BIC, two latent subgroups are detected among patients with treatment B and patients with treatment C, and no latent subgroup is identified among patients with treatment A and treatment D. More specifically, for treatment A, the BICs for models assuming no latent subgroup, two and three latent subgroups are calculated as 257.5, 311.7 and 382.4 respectively. For treatment B, the BICs for these three models are 258.2, 221.0 and 347.2 respectively. For treatment C, BICs for these three models become 225.0, 203.9 and 260.0 respectively. For treatment D, the corresponding BIC values for these three models are 250.0, 278.4 and 305.1. Table 3.3 reports the estimated regression coefficients for patients under treatment B and patients under treatment C. Assuming that the survival outcomes for patients follow a Weibull distribution, among patients under treatment B, the survival distribution estimates yield a shape parameter of 3.06 and a scale parameter of 0.24 for latent group 1, and a shape parameter of 1.88 and a scale parameter of 1.44 for latent group 2, based on the model without variable selection. For patients under treatment C, the survival distribution estimates yield a shape parameter of 2.57 and a scale parameter of 0.28 for latent group 1, and a shape parameter of 2.21 and a scale parameter of 1.74 for latent group 2, based on the model without variable selection. After variable selection, two estimated Weibull distributions yield $\hat{\kappa}_1 = 2.88, \hat{\lambda}_1 = 0.27, \hat{\kappa}_2 = 1.87$ and $\hat{\lambda}_2 = 1.84$ for patients under treatment B, and $\hat{\kappa}_1 = 2.57, \hat{\lambda}_1 = 0.28, \hat{\kappa}_2 = 2.10$ and $\hat{\lambda}_2 = 1.75$ for patients under treatment

C. The middle and right panels of Figure 3.2 demonstrate survival profiles for two latent groups from patients under treatment B and treatment C. After we obtain the predicted latent group membership, A logrank test is performed to evaluate the difference between survival profiles for two latent groups from patients under treatment B and treatment C. P-values from logrank test are smaller than 0.0001, which implies that, among patients under treatment B and treatment C, two latent groups are significantly different in terms of their survival profiles.

Based on the results of variable selection, we find that for patients under treatment B, the latent subgroup membership assignment is associated with age, the number of positive nodes, ER status, physical well-being and mood. When looking at patients under treatment C, only age and the number of positive nodes are predictive of the latent subgroup membership assignment, which agrees with the findings in the overall model. With these findings, we conclude that some latent subgroups of patients under treatment B and treatment C respond to the treatment differently due to some important covariates such as age and the number of positive nodes. Therefore, it is of interest to further study how the treatment works differently for some subgroups that are determined by important covariates. We create four subgroups of patients based on dichotomized age and the number of positive nodes. More explicitly, we dichotomize age by a threshold of 40 years, which is learned from previous findings about the IBCSG trial (International Breast Cancer Study Group, 1996). A Cox proportional hazards model with treatment as the only covariate is then applied to each subgroups of patients to evaluate the treatment effect. In the subgroups of patients aged less than 40 years and with more than 4 positive nodes, treatment C has a significant effect on the survival outcomes (p-value=0.031). Compared with treatment A, the hazard ratio of treatment C is 2.48 with 95% confidence interval $(1.085, 5.666)$, which implies that the hazard for patients treated with CMF*3 is higher than for patients treated with CMF*6. Kaplan-Meier curves for patients under each treatment in different subgroups are demonstrated in Figure 3.3.

3.5.2 Application to Assay of Free Light Chain Data

Our proposed method is also applied to the assay of free light chain data, which involves a study of the relationship between the assay of free light chain (FLC) and mortality. The serum FLC assay is important in the diagnosis, prognosis, and disease measurement of plasma cell disorders, such as monoclonal gammopathy of undertermined significance (MGUS), multiple myeloma, and amyloidosis. Dispenzieri et al. (2012) found that elevated FLC levels were indeed associated with higher death rates in the general population.



Note: "Time" is disease free survival and rescaled to $[0, 1]$.

Figure 3.2: Kaplan-Meier Curves for latent subgroups and for subgroups determined by age and the number of positive nodes

The data of assay of serum free light chain consist of an age and sex stratified random sample of residents of Olmsted County aged 50 or older. Our analysis involves 6521 individuals after excluding those with missing values. Events were observed in 1959 individuals, with an event rate of 30% and a median follow-up time of 11.8 years. We first determine that two latent subgroups exist in the data, which is suggested by a BIC of value 5340.8 based on the model assuming two latent subgroups in the data. BIC of models assuming no latent subgroup and three latent subgroups in the data are 7273.2 and 5719.3, respectively. Next, we apply the EM algorithm to estimate the survival distributions, assuming that the underlying survival follows a Weibull distribution. The estimated shape parameters of the Weibull distribution for two latent groups are 1.09 and 1.39, respectively. The corresponding scale parameter estimates are 0.54 and 4.50. The coefficients estimates for the baseline covariates are reported in Table 3.4. Under this model, we implement our variable selection procedure to select important covariates that are associated with latent subgroup membership assignment. After variable selection, we refit the model using the selected nonzero covariates. The estimated Weibull distributions for two latent subgroups are very close to those without variable selection: the survival distribution for latent group 1 has a shape parameter estimate of 1.08 and scale parameter estimate of 0.54; the survival distribution for latent group 2 has a shape parameter estimate of 1.38 and scale parameter estimate of 4.46. The results of the adaptive lasso estimator and maximum likelihood estimator after variable selection are summarized in Table 3.4.

The latent subgroups membership is associated with the main effects of age, sex, log of the kappa portion of serum free light chain, log of lambda portion of serum free light chain and log of serum creatinine, and

Table 3.3: Parameter estimates for the IBCSG trial data

	Overall models		Treatment-specific models			
			Treatment B		Treatment C	
Covariates	MLE w/o var. sel. (p-value)	MLE after var. sel. (p-value)	MLE w/o var. sel. (p-value)	MLE after var. sel. (p-value)	MLE w/o var. sel. (p-value)	MLE after var. sel. (p-value)
(intercept)	1.21	1.21	1.09	1.05	0.90	0.87
	(<0.0001)	(<0.0001)	(0.0017)	(0.0019)	(0.0176)	(0.0009)
age	0.25	0.25	0.16	0.16	0.57	0.55
	(0.0034)	(0.0030)	(0.3316)	(0.3309)	(0.0015)	(0.0017)
node	-1.06	-1.05	-1.74	-1.77	-1.26	-1.16
	(<0.0001)	(<0.0001)	(<0.0001)	(0.0033)	(0.0007)	(0.0011)
ER status	0.13	0	0.32	0.32	0.03	0
	(0.4869)	(-)	(0.3811)	(0.3740)	(0.9277)	(-)
physical	0.08	0	-0.17	-0.08	0.22	0
	(0.4456)	(-)	(0.4591)	(0.7095)	(0.2959)	(-)
mood	-0.24	0	-0.02	-0.42	-0.36	0
	(0.0375)	(-)	(0.0129)	(0.0451)	(0.1271)	(-)
appetite	0.04	0	0.26	0	0.02	0
	(0.6430)	(-)	(0.2178)	(-)	(0.8964)	(-)
cope	0.21	0	0.27	0	0.32	0
	(0.0302)	(-)	(0.2008)	(-)	(0.0953)	(-)
trtB	-0.03	0	-	-	-	-
	(0.8884)	(-)	(-)	(-)	(-)	(-)
trtC	-0.17	0	-	-	-	-
	(0.4925)	(-)	(-)	(-)	(-)	(-)
trtD	-0.08	0	-	-	-	-
	(0.7354)	(-)	(-)	(-)	(-)	(-)

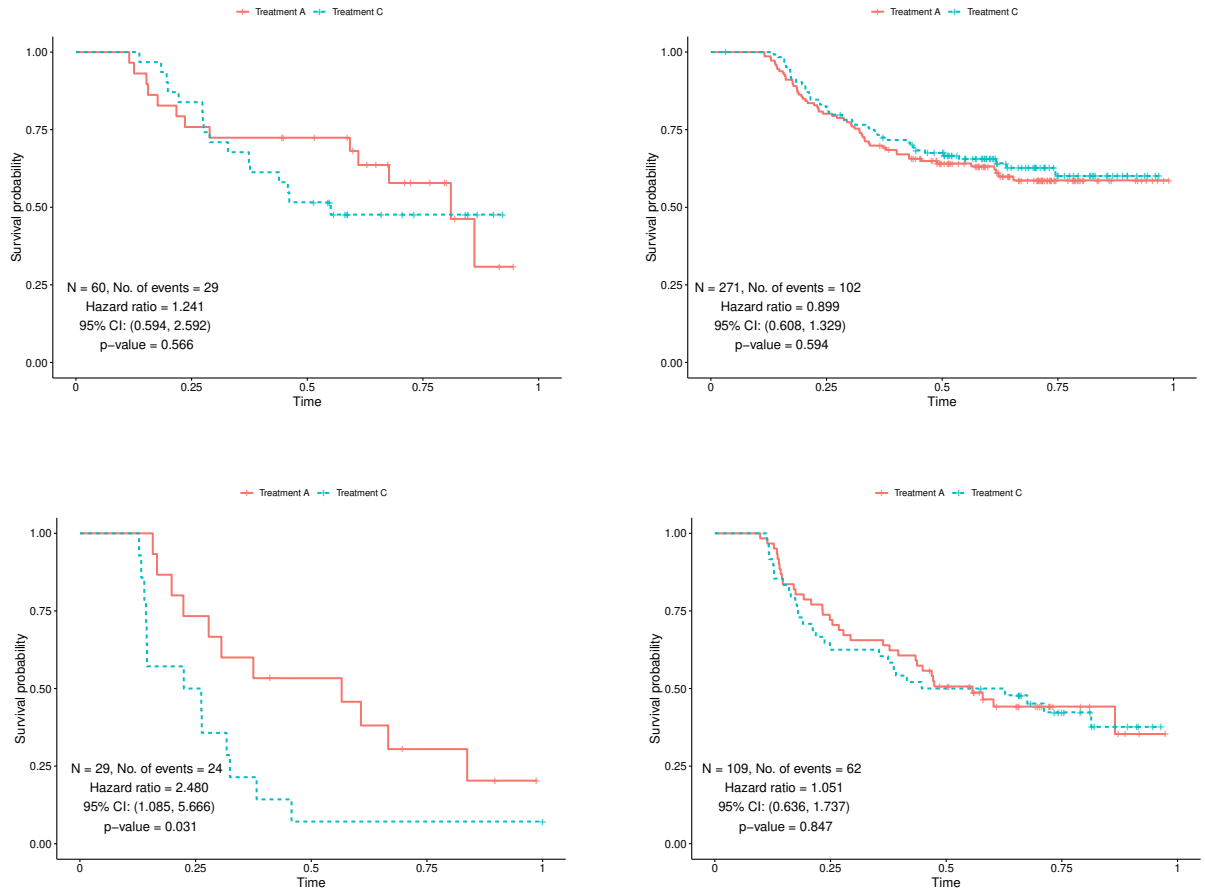
Note. "node": the number of positive nodes in the tumor; "physical": physical well-being; "cope": perceived coping.

interaction effects of age and sex, age and log of lambda portion of serum free light chain, sex and log of lambda portion of serum free light chain, sex and log of serum creatinine, and log of kappa portion of serum free light chain and log of serum creatinine. The latent subgroup membership for each individual is predicted afterwards, based on maximum likelihood estimates without variable selection and after variable selection. The model without variable selection yields that 1693 out of 6521 individuals belong to latent group 1, which makes up about 26% of the total individuals in the data. After variable selection, 1701 individuals ($\approx 26.1\%$) belong to latent group 1. Kaplan-Meier curves for two latent subgroups are utilized to illustrate the survival profiles, which are shown in Figure 3.4.

Table 3.4: Parameter estimates for the assay of free light chain data

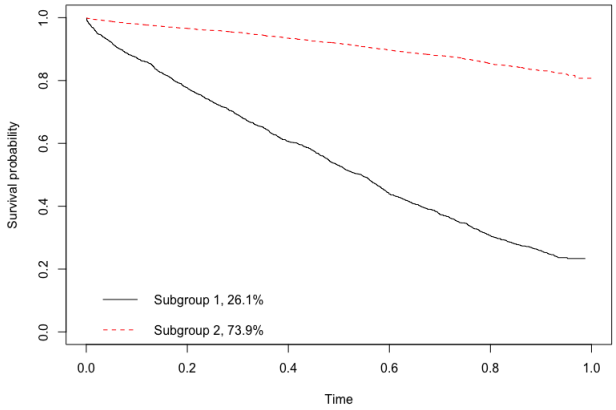
Covariates	MLE without variable selection (p-value)	MLE after variable selection (p-value)
(intercept)	2.16 (0.4188)	2.60 (<0.0001)
age	-3.16 (<0.0001)	-3.37 (<0.0001)
sex	1.54 (0.0001)	1.19 (0.0011)
kappa	-0.83 (0.6086)	-0.75 (0.0009)
lambda	-1.98 (0.4480)	-1.50 (0.0001)
creatinine	2.62 (0.0083)	0.49 (0.4900)
age_sex	-0.26 (0.3669)	-0.04 (0.8880)
kappa_lambda_sum	0.79 (0.8487)	0 (-)
kappa_lambda_ratio	-0.40 (0.2036)	0 (-)
kappa_age	0.36 (0.3745)	0 (-)
kappa_sex	-0.50 (0.3369)	0 (-)
lambda_age	0.98 (0.0240)	1.25 (0.0001)
lambda_sex	-0.60 (0.3301)	-0.64 (0.1173)
creatinine_age	-0.96 (0.1890)	0 (-)
creatinine_sex	1.05 (0.1534)	0.61 (0.4336)
kappa_creatinine	-0.30 (0.7889)	-2.18 (0.0019)
lambda_creatinine	-3.43 (0.0090)	0 (-)

Note: "Age" is in years, standardized by subtracting mean and dividing standard deviation. "kappa"' is the log of kappa portion of serum free light chain. "lambda" is the log of lambda portion of serum free light chain. "creatinine"' is the log of serum creatinine. "age_sex"' is the interaction of standardized age and sex. "kappa_lambda_sum"' is the log of sum of kappa and lambda free light chain. "kappa_lambda_ratio"' is the ratio of kappa and lambda free light chain. The rest are the interaction terms.



Note: Top-left panel: subgroup of patients with age less than 40 years old and the number of positive nodes less than 4. Top-right panel: subgroup of patients with age more than 40 years old and the number of positive nodes less than 4. Bottom-left panel: subgroup of patients with age less than 40 years old and the number of positive nodes more than 4. Bottom-right panel: subgroup of patients with age more than 40 years old and the number of positive nodes more than 4. Treatment A is the reference for hazard ratio and corresponding confidence interval estimates. P-value corresponds to the logrank test of treatment effect for each subgroup. "Time" is disease free survival and rescaled to $[0, 1]$.

Figure 3.3: Kaplan-Meier Curves for patients under each treatment in different subgroups



Note: "Time" is the time from enrollment until death and rescaled to $[0, 1]$.

Figure 3.4: Kaplan-Meier Curves for two subgroups for the assay of free light chain data, based on MLE estimates after variable selection

3.6 Conclusion

In this chapter, we propose a novel algorithm to detect the latent subgroups for individuals with different survival profiles and identify important covariates that are associated with the latent subgroup membership assignment. We have shown that our proposed estimator is consistent and enjoys the oracle properties when the number of covariates diverges with the sample size. Our proposed method can simultaneously estimate the unknown survival distributions and the coefficients that are predictive of latent subgroup membership assignment. The data with a large number of covariates can be handled well through a penalized objective function. This proposed methodology would potentially work as an exploratory step in clinical trial settings before implementing a subgroup analysis to study the treatment effect. The selected important covariates may help to explicitly determine the subgroup and discover how patients in different subgroups respond differently to treatments. Specific treatments could be developed for a target group of patients subsequently. Furthermore, using this proposed algorithm, we could directly classify patients into high-risk and low-risk groups based on their survival profiles. Since the identified classes have distinct survival distributions, each class is clinically meaningful, corresponding to patients with either long or short survival trajectories. Thus, the obtained classes can be useful to differentiate subgroup of patients at least in the following direction. First, the obtained latent classes can be used for patient recruitment in conducting future clinical trials. For example, we can recruit more patients from the high-risk group to empower trials. Another potential application is as illustrated in the real data application, our method can be used to identify subgroups of patients who may more benefit from one treatment as compared to the rest and to explore the baseline characteristics of these subgroups of patients, or their intersections, based on selected important covariates.

In our proposed algorithm, the survival distributions for latent subgroups are assumed to follow Weibull distributions with unknown parameters. It is easy to extend out methodology to other parametric distributions, such as the exponential distribution and the lognormal distribution. Our parametric framework could also be weakened by assuming that the baseline hazard function is semi-parametric and we estimate the baseline cumulative hazard function using the Breslow's estimator.

The distribution of the latent subgroup membership, given baseline covariates, is assumed to be a multinomial distribution. This assumption could be relaxed by considering a tree-based partition on the data, which could be used to identify latent subgroups and select important covariates in a nonparametric framework.

The tree-based method for latent subgroup identification may be helpful for handling the data with many covariates. Lastly, the way we select the best number of latent subgroups is to essentially apply the proposed method to the data by assuming a different number of latent subgroups in the data. A nonparametric approach for determining the best number of latent subgroups could be established as well. We could implement the tree-based partition procedure based on several choices of the number of latent subgroups assumed in the data, compute BIC values for each choice and select the number of latent subgroups associated with the smallest value of BIC. Semi-parametric approaches and their theoretical justification will be detailedly discussed in Chapter 4.

CHAPTER 4: MIXTURE SURVIVAL TREES FOR CANCER RISK CLASSIFICATION

4.1 Introduction

In oncology studies, identifying subpopulations who are high-risk or vulnerable for cancer relapse or death is crucial for drug development, due to extensive heterogeneity among cancer patients and the great cost for conducting oncology studies. Accurate risk classification, which evaluates an individual patient's survival based on his/her clinical status, genetic markers, and environmental exposure, is also essential for developing targeted cancer therapies in the era of precision medicine.

Parametric or semiparametric survival models are commonly used to evaluate individual risks. For example, one can fit a Cox proportional hazards model (Cox, 1972) or other transformation models (Zeng et al., 2016) to obtain a risk score as a linear function of individual covariates. One example of such a risk score is derived from the Framingham Heart Study. The analysis (Kannel and McGee, 1979; Kannel et al., 1979; Wilson et al., 1998) revealed that coronary heart disease risk is associated with age, diabetes, blood pressure, cholesterol level and smoking status. Another example is related to advanced oral cancer, in which (Tseng et al., 2020) classified patients into high- and low-risk groups based on comprehensive clinicopathologic and genetic data by using a Cox proportional hazards model and pre-specified thresholds for risk stratification.

Nonparametric and machine learning methods have also been developed to evaluate a patient's risk. Decision trees, due to their simplicity and interpretability, have become a popular approach to tackle time-to-event data in the literature. Specifically, survival trees (Ciampi et al., 1981; Marubini et al., 1983; Gordon and Olshen, 1985) were developed to extend existing tree-based methods for continuous outcomes to handle survival data, where splitting rules were constructed to optimize the within-node homogeneity and the between-node heterogeneity. More recently, various splitting criteria have been greatly discussed, such as the likelihood ratio test (Ciampi et al., 1987), exponential log-likelihood loss (Davis and Anderson, 1989), the full likelihood deviance (LeBlanc and Crowley, 1992), the integrated absolute difference between two

children nodes survival functions (Moradian et al., 2017), and the integrated concordance measure to evaluate the difference in hazards of two child nodes (Sun et al., 2019). To classify patients into different risk groups, survival functions (Ibrahim and Kudus, 2009; Zhou and McArdle, 2015) or hazard functions (Vergara et al., 2018) at each terminal node were obtained and compared with a pre-specified threshold value to determine the high- or low-risk group of each patient.

There are several limitations with both semiparametric and machine learning approaches for survival risk classification. First, parametric or semiparametric models aim to study the association between the covariates and the outcome, and therefore are not actually developed for risk classification. Model misspecification, for example, due to the monotonicity of the risk scores from a Cox model, can lead to serious misclassification of risk groups. Second, although machine learning methods such as survival trees are more robust to model misspecification, they are designed for survival prediction but not for risk classification directly. In addition, node splitting during the recursive partition is often based on comparing survival functions from nested nodes in order to yield many distinct survival functions at terminal nodes, thus making the choice of decision for risk classification difficult. Instead, these approaches have to rely on some crude summary statistics for the survival functions, such as median survival or survival probabilities at given time points, to classify patients, which likely miss the entire picture of individual survival profiles. Finally, all these methods rely on choosing threshold values for classification, which can be subjective and may not be clinically meaningful.

Instead, a more direct approach for risk classification is to treat risk group labels for patients as missing data so that the observed data can be used to infer these labels. For example, (Liao and Liu, 2019) pointed out that, for some particular cancers such as melanoma, the patient population can be approximately decomposed into two or three latent groups with unique survival profiles in each latent subgroup. Therefore, one could consider a mixture of multiple survival distributions with unknown parameters to characterize survival profiles for each latent group of patients. The resulting estimates for group membership, which are a parametric function of the covariates, directly provide risk classification for each individual. Finite mixture models have been widely used to study heterogeneity in survival data (Larson and Dinse, 1985; Farewell, 1982). Shen and He (2015) constructed a structured logistic-normal mixture model to identify a subgroup with an enhanced treatment effect. They further performed a confirmatory statistical test before model building to examine the existence of subgroups. A Quasi-Newton EM algorithm (Bussy et al., 2019) explored patient risk groups based on discrete survival data. However, all these methods are parametric that assume a restrictive relationship

between the covariates and risk classes, and thus suffer from model misspecification. Computation is also challenging when a large number of covariates are involved.

In this chapter, we propose a mixture survival tree method for risk classification. Specifically, we assume that the patients can be classified into a pre-specified number of risk groups, in which each group has distinct survival profiles that are modelled using a general family of Weibull distributions. We model each group membership nonparametrically in terms of patient's covariates. For estimation, we adopt tree-based methods to estimate the group membership during an EM algorithm. At each iteration, the observed log-likelihood function is used as the splitting criterion to optimize the within-node homogeneity and the between-node heterogeneity in terms of patients' survival behaviors. Since only a binary split is used at the iteration, computation is fast and we show that the likelihood function increases over iterations. More importantly, our proposed method provides a direct risk classification for future patients since the risk group membership can be derived explicitly using the estimated trees.

The rest of this chapter is organized as follows. The rest of this paper is organized as follows. Section 4.2 states the motivation of the proposed tree-based method for survival risk group discovery. Section 4.3 provides details about the proposed tree-based method. Section 4.4 demonstrates the performance of the proposed method via extensive simulation studies. The real data application is illustrated in Section 4.5.

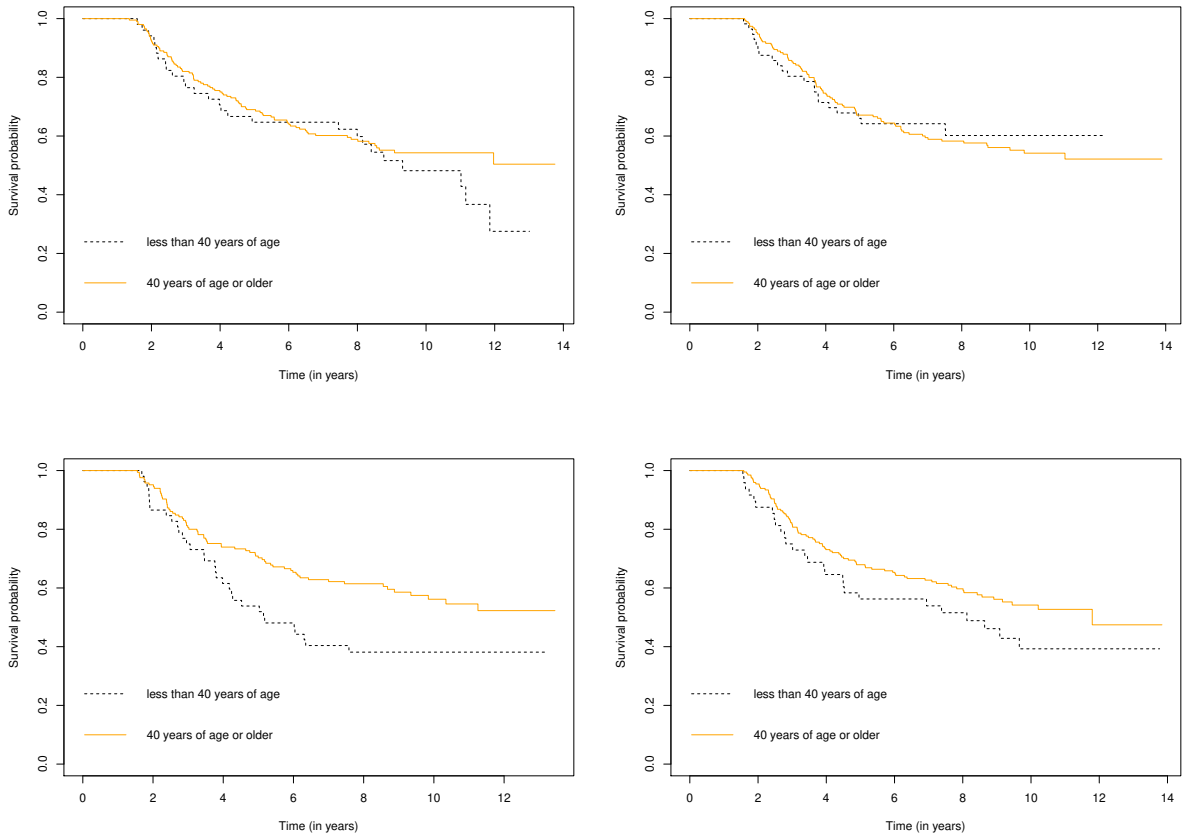
4.2 The IBCSG Breast Cancer Trial

The development of the proposed mixture survival tree was motivated by a large clinical trial, IBCSG Trial VI, conducted by the International Breast Cancer Study Group (IBCSG). IBCSG Trial VI studied both the duration of adjuvant chemotherapy and the reintroduction of delayed chemotherapy in premenopausal women with node-positive breast cancer (Colleoni et al., 2002; Chi and Ibrahim, 2006). Patients were randomized in a two by two factorial design to receive the following: (A) cyclophosphamide, methotrexate, and fluorouracil (CMF) for six consecutive cycles (CMF*6); (B) CMF*6 plus three single cycles of reintroduction CMF; (C) CMF for three consecutive cycles (CMF*3); and (D) CMF*3 plus three single cycles of reintroduction CMF. The patients' quality of life (QOL) was also measured at baseline and was hypothesized to contain prognostic information and reflect breast cancer progression. Four aspects of QOL, including physical well-being, mood, appetite and perceived coping, were assessed by a self-assessment QOL questionnaire. In

addition to treatment effects and patients' QOL at baseline, disease-free survival (median follow-up of 7.47 years, rescaled to $[0, 1]$ in analysis), event status, age at baseline, estrogen receptor (ER) status (1=positive, 0=negative) and the number of positive nodes of the tumor (i.e., node group, 1=number of positive nodes > 4, 0=else) are also considered. After excluding missing values, data are available for 962 patients. The event rate is around 45%. More details about the trial are described by International Breast Cancer Study Group (1996).

To explore the heterogeneity of patients, Kaplan-Meier curves of disease-free survival according to treatment for patients less than 40 years of age and for patients 40 years of age or older are used to demonstrate the difference in patients' survival profiles. The cut-off value of age at 40 years is learned from Colleoni et al. (2002) which presented an increased risk of relapse for patients who were less than 40 years old on treatment C versus treatment A. Figure 4.5 shows that patients who were less than 40 years of age respond to different treatments differently. For example, for patients with age less than 40 years, survival curves for those receiving treatments B and C level off after 6 years but such phenomenon is not observed for patients receiving treatments A and D. Moreover, within the treatment group, the shapes of survival curves are different between patients who were less than 40 years of age and patients 40 years of age or older. For treatments A and B, survival curves for patients under different age groups cross over, while for treatments C and D, patients younger than 40 years of age have a higher survival risk compared with patients 40 years of age or older. These findings imply that latent heterogeneity in patients with respect to treatment exists. Since other baseline patient characteristics are also contained in the data set, to explore how the baseline covariates determine the latent heterogeneity in patients, we propose a tree-based method to discover latent heterogeneity for patients with different survival profiles and directly classify patients into different survival risk groups.

Figure 4.5: Kaplan-Meier curves of disease-free survival according to treatment for patients less than 40 years of age and for paitents 40 years of age or older



Note: The top left panel illustrates the Kaplan-Meier curves for patients receiving treatment A. Top right panel are for patients receiving treatment B. Bottom Left panel are for patients receiving treatment C. Bottom right panel are for patients receiving treatment D.

4.3 Methodology

4.3.1 Mixture Survival Model

We assume that the entire population consists of K different survival risk groups. Each group of patients will follow a specific survival profile. More specifically, we assume that the kth group has survival distribution $S(t, \boldsymbol{\eta}_k)$, which has a parametric form with unknown parameters $\boldsymbol{\eta}_k$, for $k = 1, \ldots, K$. In this paper, we assume that the survival outcome for each latent risk group follows a Weibull distribution, which is a commonly used distribution in survival analysis due to its flexibility and reliability (Liao and Liu,

2019). The Weibull distribution for the kth latent risk group has the form $S(t, \boldsymbol{\eta}_k) = \exp\left\{-(\frac{t}{\lambda_k})^{\kappa_k}\right\}$, where $\boldsymbol{\eta}_k = (\kappa_k, \lambda_k)^T$, κ_k is the shape parameter and λ_k is the scale parameter.

Let T denote the time to event and $\mathbf{X}$ denotes all the baseline covariates, which could be high-dimensional. To classify each patient into one of the survival groups using the baseline covariates $\mathbf{X}$, we introduce a latent group membership B and assume that

$$P(T > t|B = k, \mathbf{X}) = S(t, \boldsymbol{\eta}_k)$$

and

$$P(B = k|\mathbf{X}) = \frac{\exp\{h_k(\mathbf{X})\}}{\sum_{k=1}^{K} \exp\{h_k(\mathbf{X})\}} = g_k(\mathbf{X}),$$

for $k = 1, \ldots, K$, where $h_k(\mathbf{X})$ is a nonparametric function and $h_1(\mathbf{X}) = 0$ (we set subgroup 1 as the reference group). Therefore, the latent group membership determines which group the patient should belong to. This membership depends on the baseline covariates through a nonparametric distribution. Clearly, the proposed model implies that the marginal survival distribution for T takes a mixture form:

$$P(T > t|\mathbf{X}) = \sum_{k=1}^{K} S(t, \boldsymbol{\eta}_k) g_k(\mathbf{X}),$$

and the probability of a patient belonging to risk group k given his or her baseline characteristics, $g_k(\mathbf{X})$, is the risk score that we use to assign latent survival risk group membership. To conduct a future trial for any new patient with baseline covariates $\mathbf{X} = \mathbf{x}$, we then classify this patient into risk group k with maximal value $g_k(\mathbf{X})$, i.e., the most likely group membership.

4.3.2 Tree-based Algorithm for Model Fitting

We propose a tree-based algorithm to estimate the group classification function, $g_k(\mathbf{x})$, $k = 1, ..., K$. Unlike traditional classification trees where the class labels are available, in our proposed tree-based method, the label of survival risk group, B, is unknown in the data. Hence, we will treat it as missing data and the EM algorithm will be used to address the latency of risk group membership. The basic idea of our proposed algorithm is that the covariate space is recursively partitioned to optimize the observed data log-likelihood function and the same survival risk group membership is assigned to patients in the same subregion of the

covariate space. The recursive partitioning is stopped when only a few patients are included in the child nodes.

More specifically, suppose that there are K subgroups contained in the data. We have right-censored observations from n i.i.d. patients, denoted by

$$\{Y_i = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), \mathbf{X_i}, i = 1, \ldots, n\},$$

where C_i is the censoring time. Assuming that the censoring time is independent of T_i given $\mathbf{X}_i$, the observed data log-likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\eta}, h_2, \ldots, h_K; \mathbf{X}, \mathbf{Y}, \boldsymbol{\Delta}) &= \sum_{i=1}^{n} \left[\Delta_i \log \left(\sum_{k=1}^{K} f(Y_i, \boldsymbol{\eta}_k) g_k(\mathbf{X_i}) \right) \right. \\ &\quad \left. + (1 - \Delta_i) \log \left(\sum_{k=1}^{K} S(Y_i, \boldsymbol{\eta}_k) g_k(\mathbf{X_i}) \right) \right] \end{aligned}$$

and the complete data log-likelihood function is given by

$$\begin{aligned} l_c(\boldsymbol{\eta}, h_2, \ldots, h_K; \mathbf{X}, \mathbf{Y}, \boldsymbol{\Delta}, B_i) &= \sum_{i=1}^{n} \sum_{k=1}^{K} I(B_i = k) \left[\Delta_i \log \left\{ f(Y_i; \boldsymbol{\eta}_k) g_k(\mathbf{X_i}) \right\} \right. \\ &\quad \left. + (1 - \Delta_i) \log \left\{ S(Y_i; \boldsymbol{\eta}_k) g_k(\mathbf{X_i}) \right\} \right], \end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)^T$, $f(t, \boldsymbol{\eta}_k) = -S'(t, \boldsymbol{\eta}_k)$ and $g_k(\mathbf{X_i}) = \frac{\exp\{h_k(\mathbf{X_i})\}}{\sum_{k=1}^{K} \exp\{h_k(\mathbf{X_i})\}}$.

To implement the numerical algorithm for studying the nonparametric function $g_k(\mathbf{X})$ and to grow a decision tree, we first set the starting values for $\boldsymbol{\eta}_k$ and $h_k(\mathbf{X})$, for $k = 1, \ldots, K$. $\boldsymbol{\eta}_k$'s are chosen to be close to 0 but have different values to make sure the survival distribution for each subgroup identifiable. $h_k(\mathbf{X})$ relies on the coefficients of a weighted multinomial regression model where more details will be covered later in this section. The starting values for the coefficients of the weighted multinomial regression model are chosen to be 0. For each splitting, based on the data $\mathcal{B}_{t+1}$ in the current node, we apply the EM algorithm. In the E-step, the expected log-likelihood based on all parameters conditional on the observed data is calculated, which is equivalent to calculating the posterior probability of $B_i = k$ given the observed data,

$$q_{ik}^{(t+1)} = \left(\frac{f(Y_i, \boldsymbol{\eta}_k) g_k^{(t)}(\mathbf{X_i})}{\sum_{k=1}^{K} f(Y_i, \boldsymbol{\eta}_k) g_k^{(t)}(\mathbf{X_i})} \right)^{\Delta_i} \left(\frac{S(Y_i, \boldsymbol{\eta}_k) g_k^{(t)}(\mathbf{X_i})}{\sum_{k=1}^{K} S(Y_i, \boldsymbol{\eta}_k) g_k^{(t)}(\mathbf{X_i})} \right)^{1-\Delta_i},$$

where

$$g_k^{(t)}(\mathbf{X_i}) = \frac{\exp\{h_k^{(t)}(\mathbf{X_i})\}}{\sum_{k=1}^{K}\exp\{h_k^{(t)}(\mathbf{X_i})\}},$$

for $i \in \mathcal{B}_{t+1}$ and $k = 1, \ldots, K$ by using $\{\hat{\boldsymbol{\eta}}^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\}$ obtained from parent node. In the M-step, for each feature X_j, $j = 1, \ldots, p$, and for each potential split x, we optimize the objective function

$$\begin{aligned} l(\boldsymbol{\eta}, h_2, \ldots, h_K; X_j, x)|_{\boldsymbol{\eta}=\boldsymbol{\eta}^{(t)}} &= \sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}^{(t+1)} \left[\Delta_i \log\left(f(Y_i, \boldsymbol{\eta}_k)\right)\right. \\ &\quad \left. +(1-\Delta_i)\log\left(S(Y_i; \boldsymbol{\eta}_k)\right) + \log\left(g_k(X_{ij})\right)\right] \end{aligned}$$

which is equivalent to fitting a weighted multinomial regression model

$$L(\boldsymbol{\theta}_k; X_j, x_j) = \sum_{i\in\mathcal{B}_{t+1}}\sum_{k=1}^{K} q_{ik}^{(t+1)} \log\left\{\frac{\exp\{\theta_{0k} + \theta_{1k} I(X_{ij} < x_j)\}}{\sum_{k=1}^{K}\exp\{\theta_{0k} + \theta_{1k} I(X_{ij} < x_j)\}}\right\} \tag{4.5}$$

based on the current data $\mathcal{B}_{t+1}$. To fit the weighted multinomial regression model (4.5), we use the one-step Newton-Raphson update. More specifically, $\theta_{0k}^{(t+1)}$ and $\theta_{1k}^{(t+1)}$ are updated by

$$\boldsymbol{\theta}_k^{(t+1)} = (\theta_{0k}^{(t+1)}, \theta_{1k}^{(t+1)})^T = \boldsymbol{\theta}_k^{(t)} - \left(\frac{\partial^2 L(\boldsymbol{\theta}; X_j, x)}{\partial\boldsymbol{\theta}_k \partial\boldsymbol{\theta}_k^T}|_{\boldsymbol{\theta}_k=\boldsymbol{\theta}_k^{(t)}}\right)^{-1} \frac{\partial L(\boldsymbol{\theta}; X_j, x)}{\partial\boldsymbol{\theta}_k}|_{\boldsymbol{\theta}_k=\boldsymbol{\theta}_k^{(t)}}$$

for $k = 1, \ldots, K$.

Next step is to search for the splitting variable and corresponding splitting value among the grid points that optimize the observed data log-likelihood. An exhaustive search is performed on the features $X_j, j = 1, \ldots, p$ and their corresponding potential splitting values x. Here, x can be the 20th, 30th, ..., 80th percentile of X_j to find the splitter $\{X_{j'}, x_{j'}\}$ that maximizes the objective function (4.5). We then calculate the corresponding $\{\theta_{0k}^{'}, \theta_{1k}^{'}\}$ and update $h_k^{(t+1)}(\mathbf{X_i})$ at each iteration by multiplying a linear combination of an indicator function using $\{\theta_{0k}^{'}, \theta_{1k}^{'}\}$ and the split $\{X_{j'}, x_{j'}\}$,

$$h_k^{(t+1)}(\mathbf{X_i}) = h_k^{(t)}(\mathbf{X_i}) \times \left\{\theta_{0k}^{'} + \theta_{1k}^{'} I(X_{ij'} < x_{j'})\right\}.$$

The nonparametric function $h_k(\cdot),\ k = 1, \ldots, K$, demonstrates the how the covariate space is partitioned over iterations. To estimate the survival distribution for each risk group, we consider the whole dataset $\mathcal{B}$. That

is, for each level of the tree, we combine all the data points in each of the child nodes. The Newton-Raphson algorithm is then used to update the unknown survival parameter $\boldsymbol{\eta}_k$, $k = 1, \ldots, K$, by computing the first and second derivatives of the objective function

$$l(\eta, h_2, \ldots, h_K; \mathbf{X_i}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t)} \left[\Delta_i \log\left(f(Y_i, \boldsymbol{\eta}_k)\right) + (1 - \Delta_i) \log\left(S(Y_i; \boldsymbol{\eta}_k)\right) + \log\left(g_k(\mathbf{X_i})\right)\right]$$

with respect to $\boldsymbol{\eta}$ to obtain $\hat{\boldsymbol{\eta}}^{(t+1)}$. More specifically, at each level, we only iterate once to update survival parameters $\boldsymbol{\eta}_k$, $k = 1, \ldots, K$, based on the whole dataset. The binary splitting is stopped when each leaf contains no more than $\sqrt{n}$ patients and the pruning procedure is performed backwards to obtain the optimal-size subtree.

After fully growing a tree, we obtain the optimal-size subtree by calculating BIC for a series of subtrees. The BIC criterion can be computed by

$$BIC = -2 \log(\widehat{Lik}) + \log(n)(l + 2K),$$

where $\log(\widehat{Lik})$ is the observed data log-likelihood of the current tree, n is the number of observations in $\mathcal{B}$, and l is the number of leaves of the current tree. Pruning is stopped when the tree corresponding to the smallest BIC is found.

A similar procedure is used to select the number of latent risk groups contained in the data by assuming that a multiple number of latent groups exists in the data. The tree that yields the smallest BIC is then chosen as best. We mention here that BIC penalizes the number of leaves for tree pruning, while the penalty for the number of levels is used for the BIC calculation to select the best number of latent groups. Such a choice of penalty for the BIC calculation is evaluated in the simulation studies in Section 4.4.

4.3.3 Monotone Likelihood Property of the Algorithm

In this section, we show that the observed data log-likelihood increases monotonically over iterations based on the algorithm stated in Section 4.3.2.

Without loss of generality, we pick an arbitrary node, $\mathcal{N}^{(t)}$, in tth iteration and the dataset that corresponds to this node is denoted by $\mathcal{B}^{(t)}$. Assume that in the $(t+1)$th iteration, two datasets, $\mathcal{B}_1^{(t+1)}$ and $\mathcal{B}_2^{(t+1)}$, are obtained based on binary splitting from $\mathcal{B}^{(t)}$, where $\mathcal{B}^{(t)} = \mathcal{B}_1^{(t+1)} \cup \mathcal{B}_2^{(t+1)}$. With a slight abuse of notation, we write the observed data as $\mathbf{Y_{obs}} = \{Y_i, \Delta_i, \mathbf{X_i}, i = 1, 2, \ldots, n\}$, the complete data as $\mathbf{Y_c} = \{Y_i, \Delta_i, \mathbf{X_i}, B_i, i = 1, 2, \ldots, n\}$, where B_i is the latent group membership for patient i, and the missing data as $\mathbf{Y_{mis}} = \{B_i, i = 1, 2, \ldots, n\}$. We are led to the following theorem.

Theorem 4.4. *For any node in the tth iteration and its corresponding child nodes in the $(t+1)$th iteration, the value of the observed data log-likelihood increases between two successive iterations. That is,*

$$\begin{aligned} l(\boldsymbol{\eta}^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{obs} \in \mathcal{B}^{(t)}) \quad &\leq \quad l(\boldsymbol{\eta}^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{obs} \in \mathcal{B}_1^{(t+1)}) \\ &+ \quad l(\boldsymbol{\eta}^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{obs} \in \mathcal{B}_2^{(t+1)}), \end{aligned}$$

where $l(\boldsymbol{\eta}^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{obs} \in \mathcal{B}^{(t)})$ denotes the observed data log-likelihood in the tth iteration and $l(\boldsymbol{\eta}^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{obs} \in \mathcal{B}_1^{(t+1)})$, $l(\boldsymbol{\eta}^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{obs} \in \mathcal{B}_2^{(t+1)})$ denote the observed data log-likelihood in the two corresponding child nodes.

The proof of Theorem 1 is given in the Appendix.

4.4 Simulation Studies

4.4.1 Simulation Setting

We performed extensive simulation studies to evaluate the performance of our proposed tree-based method. We consider two latent risk groups in the first simulation study and assume that the data contain three latent risk groups in the second simulation study. In each simulation study, three different scenarios are evaluated and the results are compared with the parametric mixture model, Cox PH model with risk stratification and survival tree. We apply the parametric mixture model by following Bussy et al. (2019). A mixture form, $f(t|X = x) = \sum_{k=1}^{K} \pi_k(x) f_k(t; \alpha_k)$, is considered to model the conditional probability of the survival time T given the baseline covariates. The weights, $\pi_k(x)$, for each mixing component depend on the patient baseline characteristics and have a logistic regression function $\pi_k(x) = \frac{\exp\{x^T \beta_k\}}{\sum_{k=1}^{K} \exp\{x^T \beta_k\}}$. The EM algorithm is used to handle the latency of the survival risk group label. The regression coefficients of

the covariates are estimated by the Newton-Raphson algorithm and then are used to classify patients into different survival risk groups. Note that this approach can be applied to any parametric distribution. The Cox PH model with risk stratification is an application of Tseng et al. (2020), where a Cox PH model including all covariates is applied to the training set and the threshold for survival risk group classification is set to be the median survival times of the training set (tertiles of survival times in observed samples are used for three latent groups scenarios). This threshold will be used to dichotomize the survival data in the test set and the survival risk groups are obtained accordingly. The survival tree is implemented using the R package, *rpart*, where a survival tree that includes all covariates is fitted on the training set and the median survival times of the training set (tertiles of survival times are used for three latent groups scenarios) are set to be the threshold for survival risk group classification. The survival risk groups are obtained by dichotomizing the survival data in the test set using the threshold. We will not incorporate any variable selection procedures to compare the different methods.

For the simulation of two latent subgroups, ten baseline covariates, $X_1, X_2, \ldots, X_{10}$, are considered in the model and only a few of them have nonzero effects. $X_1, X_2, \ldots, X_{10}$ are independently generated from a standard normal distribution. Time-to-event data for each latent subgroup are generated from a Weibull distribution. The true values of the scale parameters (i.e., λ_1 and λ_2) of the Weibull distributions are set to be 1 and 4.5, respectively, and the true values of the shape parameters (i.e., κ_1 and κ_2) are 1 and 3, respectively. The 2-year survival probabilities are 13.5% for the high-risk group and 91.5% for the low-risk group. The censoring time is generated from an exponential distribution, where the mean is calibrated by a pre-specified censoring rate of 30%. The subgroup-specific survival curves are illustrated in the left plot of Figure 4.6. Different scenarios are considered to mimic various cases in actual applications. Specifically, we consider different functional forms for $h(\cdot)$. For all scenarios, we regard the high-risk group as the reference so $h_1(\mathbf{X})$ is set to be 0. For scenario I.1, $h_2(\mathbf{X})$ is set to be a linear function,

$$h_2(\mathbf{X}) = 0.4 + 0.2X_1 - 0.6X_2 - 0.3X_3.$$

In this scenario, the parametric mixture model is expected to have better performance since it is the true model. For scenario I.2, we mimic the recursive partition on the covariate space to assign the risk group

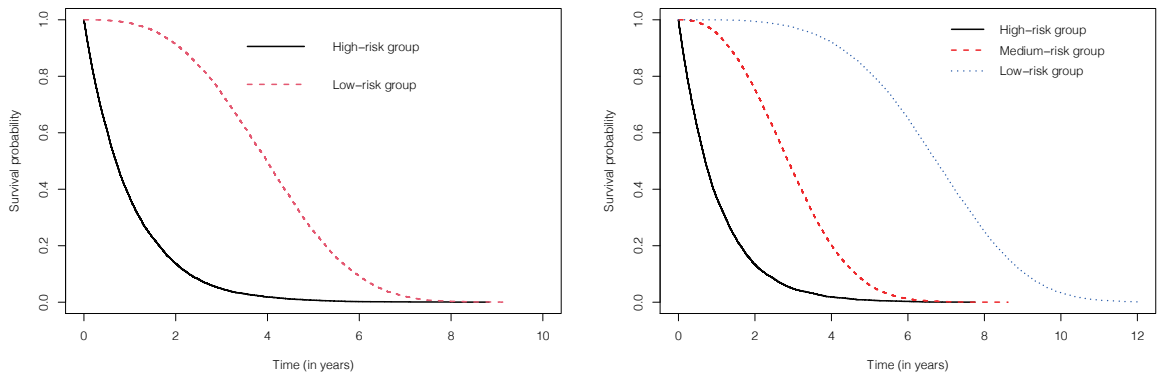
memberships by considering an indicator and interaction terms. Hence, $h_2(\mathbf{X})$ is set to be

$$\begin{aligned} h_2(\mathbf{X}) &= 0.4 + 0.2I(X_1 < 0.4, X_3 > 0.2) - 0.6X_1^2 I(X_2 > 0.1) \\ &+ 0.2I(X_3 > 0.4 \text{ or } \mathrm{X}_2 < -0.2) + 0.2\mathrm{I}(\mathrm{X}_2 < 0.4). \end{aligned}$$

We expect our proposed tree-based method will outperform in this scenario because model misspecification may be an issue in the parametric mixture model and the Cox PH model. For scenario I.3, we consider a more complicated decision boundary for survival risk group membership assignment based on the unit circle

$$h_2(\mathbf{X}) = X_1^2 + X_2^2 - 1.$$

These settings are designed to illustrate the flexibility of our proposed tree-based method.



Note: The left plot illustrates the true survival curves for two latent groups. The right plot shows the true survival curves for three latent groups.

Figure 4.6: The true survival curves for two simulation studies

We consider similar procedures for generating the baseline covariates and survival data for the simulation study with three latent groups. Ten covariates are independently generated from a truncated standard normal distribution from -2 to 2. Note that the truncated normal distribution is used for stabilizing numerical computations only. The censoring rate is set at 30%. Weibull distributions for high-, medium- and low-risk groups have true values 1, 3.3 and 7.4 for the scale parameter and 1, 2.5 and 4 for the shape parameter, respectively. The 2-year survival probabilities are 13.5%, 75.5% and 95.5% for the high-, medium- and

low-risk groups, respectively. The subgroup-specific survival curves are illustrated in the right panel of Figure 4.6. Three similar scenarios are considered in the simulation study for three latent groups. Nonparametric functions $h_2(\mathbf{X})$ and $h_3(\mathbf{X})$, for the medium- and low-risk groups respectively, are linear combinations of the baseline covariates under scenario II.1,

$$\begin{aligned} h_2(\mathbf{X}) &= 0.9 + X_1 - 0.8 * X_2 - X_3, \\ h_3(\mathbf{X}) &= 0.6 - 0.6 * X_1 - X_2 + 0.7 * X_3. \end{aligned}$$

For scenario II.2, we mimic the tree partition by setting

$$\begin{aligned} h_2(\mathbf{X}) &= -0.5 + 0.6 * I(X_1 < -1 \text{ or } \mathrm{X}_2 > 1) - 0.8 * \mathrm{I}(\mathrm{X}_1 > 0 \text{ and } \mathrm{X}_3 > 0.2), \\ h_3(\mathbf{X}) &= 0.5 + 0.4 * I(X_1 < -1 \text{ or } \mathrm{X}_2 > 1) + 0.4 * \mathrm{I}(\mathrm{X}_1 > 0 \text{ and } \mathrm{X}_3 > 0.2). \end{aligned}$$

The decision boundaries for scenario II.3 are based on the unit circle, where we set

$$\begin{aligned} h_2(\mathbf{X}) &= (X_1^2 + X_2^2 - 1) * (I(X_1 > 0.5) + 1), \\ h_3(\mathbf{X}) &= (X_1^2 + X_2^2 - 1) * (I(X_1 \leq 0.5) + 1). \end{aligned}$$

4.4.2 Simulation Results

The first step of our tree-based algorithm is to determine the number of latent groups that exist in the data. To do this, we assume different numbers of latent groups in the data and grow a tree for each choice of number of latent groups. The BIC with penalty on the number of levels of the tree is calculated for each tree and then used to determine the best number of latent groups. We report the simulation results based on 1000 replicates and consider sample sizes of 300, 500, 1000 and 2000. For the scenarios of two latent groups, the candidate choices of the number of latent groups are 1, 2 and 3. The correct number of latent groups is selected in over 70% of the simulated datasets when the sample size equals 300. As the sample size increases to 1000, the number of latent subgroups is correctly selected for almost 90% of the replicates. For the scenarios of three latent groups, the candidate choices of the number of latent groups range from 1 to 4. BIC can pick the right number of latent groups from around 70% of the simulated datasets when the sample

size is 300. The proportion increases to 80% as the sample size goes to 1000. With a sample size of 2000, the right number of latent groups is chosen for over 94% of the simulated datasets.

Tables 4.5 and 4.6 summarize the median prediction accuracy along with median absolute deviation (MAD) for simulation studies with two and three latent groups. The results from our proposed method are obtained based on trees pruned by BIC. For numerical statibility, the Weibull distribution estimates are restricted to $(0, 7]$ for two latent groups scenarios and to $(0, 15]$ for three latent groups scenarios, when the Newton-Raphson algorithm is implemented. We also report the tree structures in terms of average depth and average number of leaves. The prediction accuracy is calculated by applying the decision rule obtained from the training set to an independently generated validation set with sample size $10,000$. The optimal accuracy rate is calculated as $1 -$ Bayes error rate, where the Bayes error rate is calculated by the formula $1 - E\left(\max_k P(B = k|X)\right)$. In general, the prediction accuracy increases as the sample size increases and yields a decreasing trend in the median absolute deviation. The prediction accuracy also approaches the optimal accuracy rate for large sample sizes. In examing the tree structures, the tree-based method yields relatively simple trees even for some complicated scenarios, which is favorable for visualizing and interpreting how the baseline covariates determine the survival risk group classification. For scenarios I.1 and II.1, we assume a linear relationship in the nonparametric function $h(\cdot)$, that is, the parametric mixture model is correctly specified. Therefore, the prediction accuracies from the parametric mixture model are better than the results from the tree-based method. The results from Cox PH model with risk stratification are slightly better than results from survival tree, which may be also due to the linear relationship in the setup of $h(\cdot)$. However, for the other scenarios where the parametric mixture model is not the true model, the tree-based method outperforms the parametric mixture model. Compared with Cox PH model with risk stratification and survival tree, our approach also obtains better results. Survival tree, in general, performs better than Cox PH model with risk stratification, which is reasonable because survival tree can better capture the nonparametric relationship in $h(\cdot)$.

Table 4.5: Results from the simulation study for two latent groups scenarios

	Prediction accuracy				Tree structure	
N	Our approach Med. accu. (MAD)	Mixture Model Med. accu. (MAD)	Cox PH model Med. accu. (MAD)	Survival Tree Med. accu. (MAD)	Ave. depth	Ave. no. leaves
	Scenario I.1. The optimal accuracy rate is 0.649.					
300	0.590 (0.012)	0.614 (0.013)	0.592 (0.012)	0.537 (0.032)	2.79	8.03
500	0.600 (0.014)	0.627 (0.008)	0.606 (0.009)	0.588 (0.023)	3.30	10.36
1000	0.613 (0.009)	0.637 (0.006)	0.612 (0.006)	0.611 (0.011)	3.95	15.12
2000	0.623 (0.007)	0.642 (0.004)	0.618 (0.005)	0.615 (0.009)	4.44	20.90
	Scenario I.2. The optimal accuracy rate is 0.668.					
300	0.602 (0.012)	0.593 (0.017)	0.535 (0.015)	0.527 (0.043)	2.55	6.97
500	0.606 (0.011)	0.602 (0.011)	0.544 (0.013)	0.567 (0.040)	3.00	8.88
1000	0.620 (0.015)	0.607 (0.006)	0.553 (0.010)	0.604 (0.009)	3.76	13.68
2000	0.642 (0.012)	0.610 (0.004)	0.562 (0.007)	0.606 (0.006)	4.39	20.14
	Scenario I.3. The optimal accuracy rate is 0.721.					
300	0.623 (0.019)	0.561 (0.031)	0.502 (0.005)	0.586 (0.018)	3.51	10.61
500	0.664 (0.038)	0.589 (0.021)	0.506 (0.005)	0.620 (0.012)	4.28	15.35
1000	0.702 (0.010)	0.614 (0.008)	0.507 (0.005)	0.622 (0.006)	5.01	23.31
2000	0.709 (0.006)	0.619 (0.004)	0.507 (0.004)	0.627 (0.004)	5.32	29.71

Note. Mixture model: Parametric mixture model. Cox PH model: Cox PH model with risk stratification. Med. accu.: median accuracy. MAD: median absolute deviation. Ave. Depth: average depth. Ave. no. leaves: average number of leaves.

Table 4.6: Results from the simulation study for three latent groups scenarios

	Prediction accuracy				Tree structure	
N	Our approach Med. accu. (MAD)	Mixture Model Med. accu. (MAD)	Cox PH model Med. accu. (MAD)	Survival tree Med. accu. (MAD)	Ave. depth	Ave. no. leaves
	Scenario II.1. The optimal accuracy rate is 0.663.					
300	0.407 (0.044)	0.434 (0.098)	0.479 (0.010)	0.388 (0.088)	2.13	5.17
500	0.421 (0.034)	0.489 (0.041)	0.485 (0.007)	0.413 (0.103)	4.11	15.58
1000	0.434 (0.045)	0.550 (0.010)	0.486 (0.007)	0.469 (0.044)	4.70	23.03
2000	0.467 (0.035)	0.592 (0.006)	0.488 (0.006)	0.522 (0.025)	5.10	30.38
	Scenario II.2. The optimal accuracy rate is 0.551.					
300	0.484 (0.035)	0.328 (0.026)	0.177 (0.009)	0.177 (0.059)	3.81	12.61
500	0.502 (0.026)	0.340 (0.028)	0.177 (0.004)	0.191 (0.038)	4.02	15.01
1000	0.515 (0.020)	0.368 (0.020)	0.179 (0.002)	0.258 (0.018)	4.45	20.40
2000	0.523 (0.016)	0.428 (0.014)	0.187 (0.003)	0.324 (0.008)	4.92	28.10
	Scenario II.3. The optimal accuracy rate is 0.611.					
300	0.390 (0.035)	0.365 (0.071)	0.358 (0.004)	0.400 (0.021)	3.96	13.01
500	0.427 (0.037)	0.404 (0.050)	0.359 (0.004)	0.407 (0.022)	4.33	16.18
1000	0.473 (0.026)	0.445 (0.016)	0.359 (0.003)	0.407 (0.020)	5.15	25.20
2000	0.498 (0.015)	0.449 (0.016)	0.359 (0.004)	0.408 (0.013)	5.96	38.47

Note. Mixture model: Parametric mixture model. Cox PH model: Cox PH model with risk stratification. Med. accu.: median accuracy. MAD: median absolute deviation. Ave. Depth: average depth. Ave. no. leaves: average number of leaves.

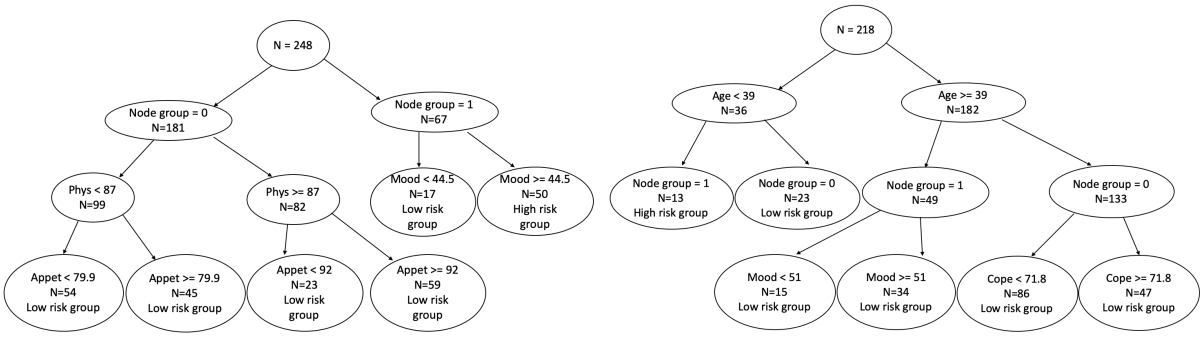
4.5 Real Data Application

We apply the proposed methodology to the IBCSG data discussed in Section 4.2 to study potential heterogeneity of patients in terms of their survival outcomes and to investigate important variables that are associated with such heterogeneity.

To explore the heterogeneity of patients under different therapeutic procedures, we apply our method to the datasets of patients under each treatment. According to BIC, two latent subgroups are detected among patients with treatment B and patients with treatment C, and no latent subgroups are identified among patients with treatment A and treatment D. More specifically, for treatment A, the BICs for models assuming no latent subgroup, two and three latent subgroups are calculated as 257.5, 272.1 and 1156.2 respectively. For treatment B, the BICs for these three models are 258.2, 190.4 and 583.9. For treatment C, the BICs for these three models are 225.0, 214.0 and 1008.2. For treatment D, the BICs for these three models are 250.0, 272.8 and 1472.2.

Our previous analysis using mixture model indicated that treatment does not have a significant effect on the latent subgroup membership assignment. To explore the heterogeneity of patients under different therapeutic procedures, we analyze data from patients under treatments B and C by assuming that two latent groups exist in the data. The results are also compared with the parametric mixture model and the Cox PH model with risk stratification by following the same procedure stated in Section 4.4. Assuming that the survival outcomes for patients follow a Weibull distribution, for patients under treatment B, the survival distribution estimates yield a shape parameter of 1.85 and a scale parameter of 0.34 for the high-risk group, and a shape parameter of 1.37 and a scale parameter of 2.47 for the low-risk group. For patients under treatment C, the survival distribution estimates yield a shape parameter of 1.79 and a scale parameter of 0.35 for the high-risk group, and a shape parameter of 1.21 and a scale parameter of 2.34 for the low-risk group. Kaplan-Meier curves for patients under treatment B and treatment C are illustrated in upper panels of Figure 4.8. The logrank test is also employed to compare the survival distributions of the two latent groups. Note that the logrank test, serving as an ad hoc measurement in this real data application, is our attempt to compare the survival distributions of the two latent groups and we will not treat the logrank test as a formal test. The results from the logrank test indicate that, for patients under treatment B and treatment C, two latent groups have significantly different survival distributions (p-values are both smaller than 0.0001). Figure 4.7 shows

the fitted trees for patients under treatment B and treatment C. Based on Figure 4.7, we see that for patients under treatment B, the latent risk group membership assignment is associated with the number of positive nodes, physical well-being, appetite, and mood. When looking at patients under treatment C, age, the number of positive nodes, mood and perceived coping are important factors in the latent risk group membership assignment.



Note. Left panel: decision tree for patients receiving treatment B. Right panel: decision tree curves for patients receiving treatment C.

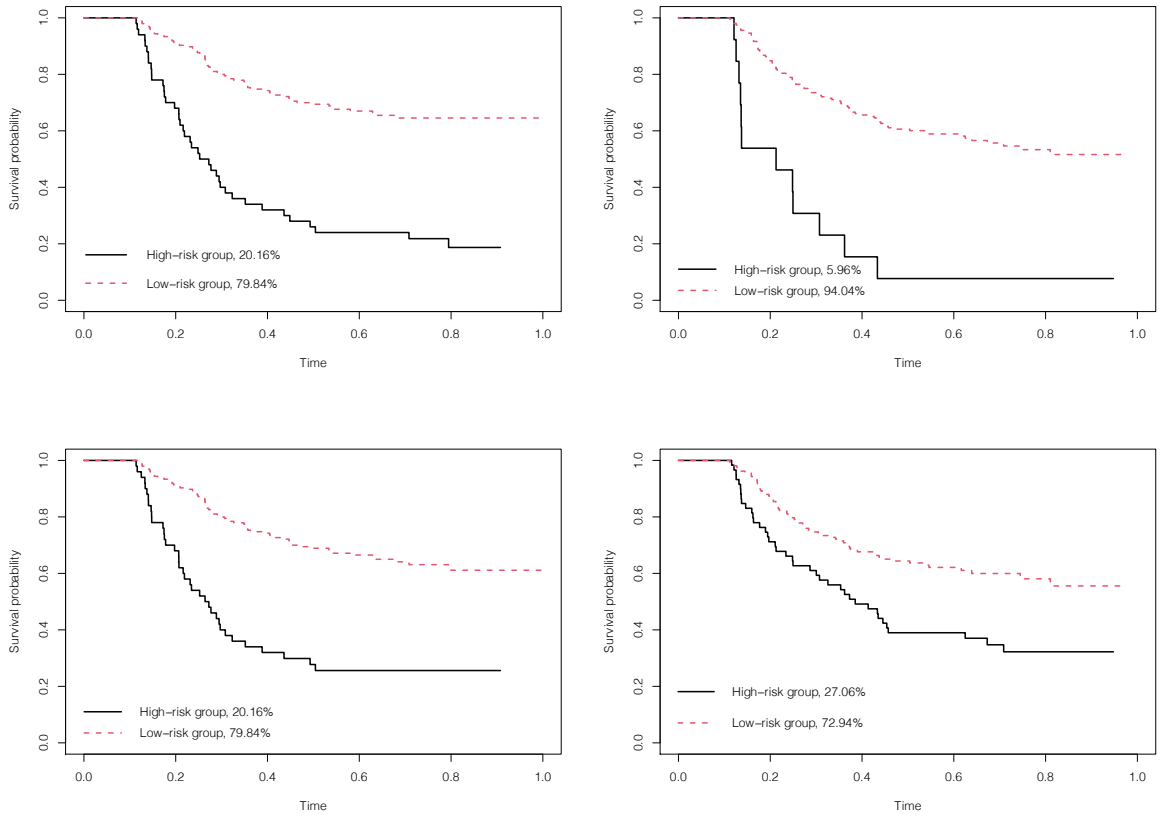
Figure 4.7: Decision trees for patients receiving treatment B and treatment C

To compare with the parametric mixture model, the Cox PH model with risk stratification and the survival tree approach, we apply these methods to the same dataset and report the agreement of survival risk group classification. For the Cox PH model with risk stratification and the survival tree approach, we use the first quartile of the survival times obtained from the observed sample as the threshold to classify patients into high- and low-risk groups since we cannot calculate the median survival times from the observed sample and the classification using the first quartile of the survival times yields the highest agreement in terms of predicted subgroup membership when comparing with our approach. As for treatment B, around 94.4% of the patients have the same predicted risk group membership from the parametric mixture model and our tree-based method. 48.8% of the patients are assigned to the same survival risk group when comparing the Cox PH model with risk stratification and the tree-based method. All the patients have the same predicted survival risk group membership when comparing the survival tree approach and our approach. For patients receiving treatment C, there are 68.8% of the patients having the same predicted survival risk group membership when applying the parametric mixture model and our tree-based method. Comparing the Cox PH model with stratification and our approach, 31.2% of the patients are classified into the same risk group. For the

comparison with the survival tree approach, 75.7% of the patients are classified into the same risk group. For the model with treatment B, the number of positive nodes and mood have significant effects on the latent risk group membership assignment based on results from both the parametric mixture model and the Cox PH model with risk stratification. These two covariates are also important based on the mixture survival tree according to the Figure 4.7. In the mixture survival tree, the first split is on the number of positive nodes greater than 4, separating a group of 181 patients who have less than 4 positive nodes in the tumor and the rest of the 67 patients who have more than 4 positive nodes in the tumor. A second split is made based on the value of mood at 44.5, which means that for those with more than 4 positive nodes in the tumor, 17 of them with mood smaller than 44.5 consist of one group and the rest of the 50 patients are thought to be similar in terms of their survival profile. For the model with treatment C, we find that age and the number of positive nodes are significantly associated with the latent group membership assignment in both parametric mixture model and the Cox PH model with risk stratification, which also agrees with the finding in the mixture survival tree as seen from Figure 4.7. The first split in the mixture survival tree is at age 39 years and the splits on the second level are both made based on the number of positive nodes in the tumor. In the survival tree, the trees for patients receiving treatment B and C are very similar. the first split from both trees is made based on the number of positive nodes in the tumor smaller than 4. For patients with more than 4 positive nodes in the tumor, the second split is made based on the value of mood at 44.5 (for patients receiving treatment B) and 49.5 (for patients receiving treatment C). It is of greater interest to examine how these variables classify patients into different risk groups. Although we find similar set of covariates that determines the latent risk group membership assignment, it is difficult to detect interaction effects between covariates based on the parametric mixture model approach and the Cox PH model. These two methods are only capable of linearly describing the relationship between the covariates and risk groups. In contrast, the mixture survival tree provides more direct visualization of how the baseline covariates determine different survival risk groups. For example, there are interaction effects among the number of positive nodes, physical well-being, and appetite when classifying patients with treatment B into different risk groups as well as interaction effects among age, the number of positive nodes and mood when assigning risk group memberships for patients with treatment C.

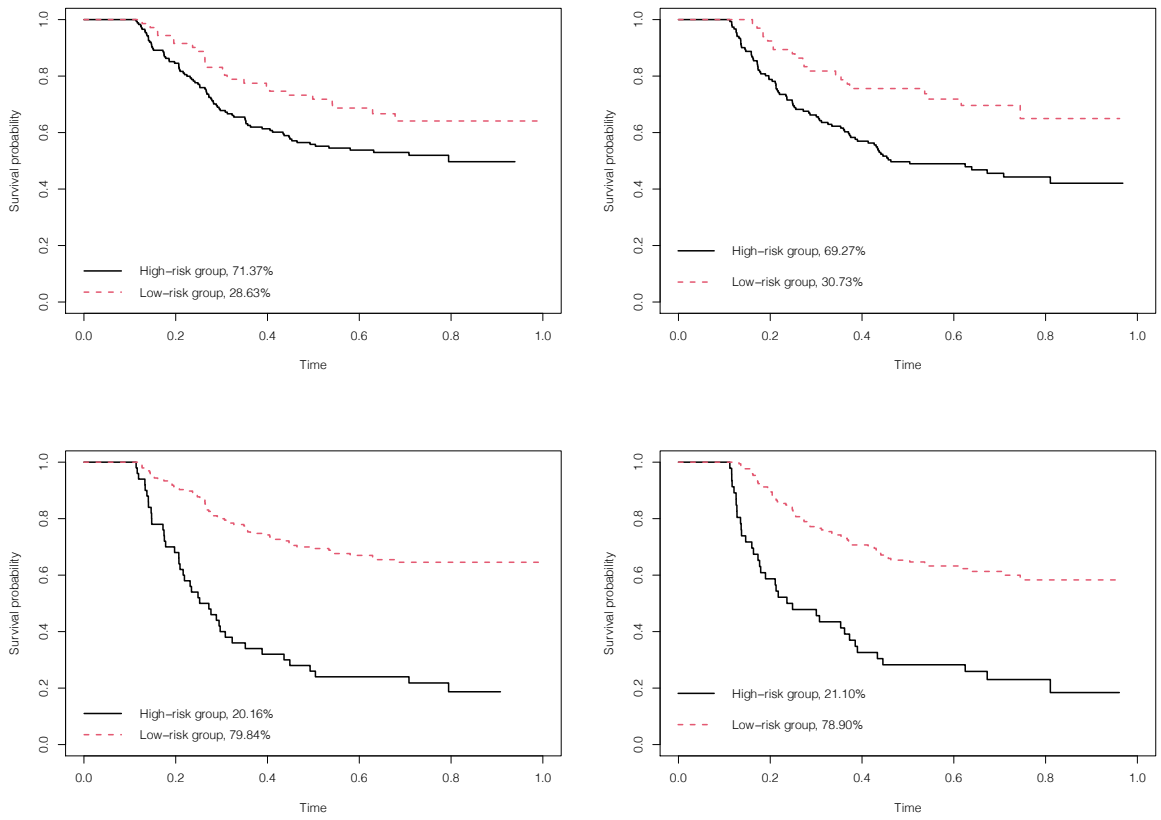
The goal for survival risk classification is to optimize the heterogeneity in survival outcome between risk groups. From this point of view, a logrank test is performed to evaluate the difference between survival

distributions estimated by the parametric mixture model, Cox PH model with risk stratification and the survival tree approach to further compare these methods with the mixture survival tree. The results suggest that for all methods with treatment B and treatment C, the two survival distributions from the high- and low-risk groups are significantly different from each other (p-values are all < 0.05). For patients with treatment B, the logrank test statistic has a value of 50.1 when testing the difference between the survival distributions estimated by the mixture survival tree, and for the parametric mixture model, Cox PH model and the survival tree approach, the corresponding test statistics are 48.8, 4.5 and 50.1, respectively. This implies that the mixture survival tree is better at distinguishing distinct survival profiles. Similar results are also observed in patients with treatment C. The logrank test statistics are 46.9, 15.9, 10.3 and 34.4 for the mixture survival tree, the parametric mixture model, Cox PH model and the survival tree approach, respectively. The Kaplan Meier curves obtained from the parametric mixture model, Cox PH model with risk stratification and the survival tree approach are also plotted in Figures 4.8 and 4.9 to better illustrate the difference of classification results among all methods and the difference between survival distributions for patients in high- and low-risk groups.



Note. Left panels: Kaplan Meier estimates from our approach and parametric mixture model (from top to bottom) for patients receiving treatment B. Right panels: Kaplan Meier estimates from our approach and parametric mixture model (from top to bottom) for patients receiving treatment C.

Figure 4.8: Kaplan Meier curves of high-risk and low-risk groups, predicted by our approach and parametric mixture model, for patients receiving treatment B and treatment



Note. Left panels: Kaplan Meier estimates from Cox PH model with risk stratification and survival tree (from top to bottom) for patients receiving treatment B. Right panels: Kaplan Meier estimates from Cox PH model with risk stratification and survival tree (from top to bottom) for patients receiving treatment C.

Figure 4.9: Kaplan Meier curves of high-risk and low-risk groups, predicted by Cox PH model with risk stratification and survival tree, for patients receiving treatment B and treatment

4.6 Conclusion

In this chapter, we propose a tree-based algorithm to explore latent heterogeneity for patients with different survival profiles and further classify patients into different survival risk groups. With the latent group membership, the EM algorithm is used to model the mixing components in the data. We propose a new splitting criterion within the framework of recursive partitioning. By optimizing the observed data log-likelihood function in each split, we maximize the within-node homogeneity and the between-node heterogeneity. Our proposed tree-based algorithm is capable of simultaneously estimating the unknown survival distributions and predicting the latent subgroup membership. A simple and interpretable tree-like structure is also presented to characterize how the baseline covariates determine the unobserved heterogeneity in patients. The simulation studies show that our proposed method works well in various settings, especially when the latent subgroup membership assignment depends on baseline covariates via a non-linear relationship. Using our approach, we could directly classify patients into different risk (e.g., high vs low risk) groups based on their survival profiles. Since the identified classes have distinct survival distributions, each class is clinically meaningful, corresponding to patients with either long or short survival trajectories. Thus, the obtained classes can be useful to differentiate subgroups of patients at least in the following direction. First, the obtained latent classes can be used for patient recruitment in conducting future clinical trials. For example, we can recruit more patients from the high-risk group to empower trials. Another potential application is as illustrated in the real data application, our method can be used to identify a subgroup of patients who may benefit more from one treatment as compared to the rest and to explore the baseline characteristics of this subgroup of patients based on selected important covariates.

One possible extension of the proposed work is to study the treatment response heterogeneity in patients within the framework of our proposed tree-based method. A semi-parametric approach can be established to simultaneously estimate the baseline survival and treatment effects for patients by assuming that the latent heterogeneity in patients is associated with heterogeneous survival and treatment responses.

CHAPTER 5: RANDOM FOREST FOR SUBGROUP ANALYSIS WITH HETEROGENEOUS TREATMENT RESPONSES

5.1 Introduction

The primary goal of subgroup analysis in clinical trials is to explore and estimate the heterogeneous treatment effect across subgroups, which might be determined by the baseline patient characteristics and/or their specific survival experience. From this point of view, subgroup analysis is fundamental to the interpretation of clinical trial results and beneficial to the development of new medicines. Significant treatment effects, sometimes, in the whole population of clinical trials cannot be observed. Whereas substantial heterogeneity in treatment effects would be discovered in a small group of patients, indicating that the treatment is beneficial to a subpopulation. As known, the heterogeneity in treatment effects could be detected by recursively partitioning the population into several subpopulations based on baseline patients characteristics and their survival profiles. However, some statistical concerns also arise due to the low power of statistical tests in clinical trials. In addition, the individualized treatment effect estimate may be of less interest, especially when the sample size is not large enough. Alternatively, a stratified treatment effect is more useful. The average treatment effect in each strata/subgroup can indicate positive or negative treatment effects, or whether or not the treatment effect greater than clinical threshold. In this regard, stratified treatment effect estimate, especially when a small number of latent subgroups is discovered and each latent subgroup has a distinct treatment effect, is favorable. Additionally, such treatment effect estimate is advantageous to detect higher power because sufficient sample is easy to obtain.

By the motivation of stratified treatment effects, we propose a random forest approach to discover latent subgroups of patients with specific survival experience and treatment responses. The unobserved subgroup labels for patients are treated as missing data and EM algorithm is embedded to optimize the observed data loglikelihood function in the binary splitting of individual trees. Ensemble methods are then applied to

combine individual trees in the forest. Variable importance measurement is developed to understand and visualize each feature's contribution to the model fitting.

The rest of this chapter is organized as follows, Section 5.2 describes the model set-up and algorithm implementation of our proposed method. Section 5.3 summarizes the finite sample performance via extensive simulation studies. Section 5.4 analyzes a phase III clinical trial of patients with hematological malignancies using the proposed method.

5.2 Methodology

5.2.1 Mixture Survival Model

We assume that there are K different latent subgroups existing in the entire population, where each latent subgroup corresponds to a distinct treatment effect. More specifically, each subgroup of patients follow a specific survival profile and the difference of the survival profile among different subgroups is due to the fact that patients from different subgroups respond to the treatment differently.

Suppose that we have the right-censored observations from n i.i.d. patients, denoted by

$$\{Y_i = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), A_i, \mathbf{X_i}, i = 1, \ldots, n\},$$

where T_i denotes the time to event, C_i is the censoring time, A_i is the binary treatment assignment and the baseline covariates are written as $\mathbf{X_i}$ for subject i. Assuming that the censoring time is independent of T_i given $\mathbf{X_i}$. We assume that the survival distribution of the kth group follows a Cox model,

$$\lambda_k(t; A, \mathbf{X}) = \lambda_0(t) \exp(\beta_k A + \boldsymbol{\gamma}\mathbf{X}),$$

where λ_0 is the common baseline hazard rate for all subgroups, β_k is the treatment effect for the kth subgroup, and $\boldsymbol{\gamma}$ is the coefficient of baseline characteristics.

To classify each patient into one of the subgroups using the baseline covariates $\mathbf{X}$, we introduce a latent group membership B and assume

$$P(T > t|B = k, A, \mathbf{X}) = S(t, \beta_k, \boldsymbol{\gamma}; A, \mathbf{X}),$$

where $S(t, \beta_k, \boldsymbol{\gamma}; A, \mathbf{X}) = \exp\left\{-\int_0^t \lambda_0(u)du \exp(\beta_k A + \boldsymbol{\gamma}\mathbf{X})\right\}$ and

$$P(B = k|\mathbf{X}) = \frac{\exp\{h_k(\mathbf{X})\}}{\sum_{k=1}^{K} \exp\{h_k(\mathbf{X})\}} = g_k(\mathbf{X})$$

for $k = 1, \ldots, K$, where $h_k(\mathbf{X})$ is a nonparametric function and $h_1(\mathbf{X}) = 0$ (set subgroup 1 as the reference group). Therefore, the latent group membership determines which group the patient should belong to. This membership depends on the baseline covariates through a nonparametric distribution and we could study this nonparametric distribution by recursively partitioning the covariate space. Clearly, the proposed model implies that the marginal survival distribution for T takes a mixture form:

$$P(T > t|A, \mathbf{X}) = \sum_{k=1}^{K} S(t, \beta_k, \boldsymbol{\gamma}; A, \mathbf{X}) g_k(\mathbf{X}).$$

To conduct a future trial, for any new patient with baseline covariates $\mathbf{X} = \boldsymbol{x}$, we then classify this patient into group k with maximal value $g_k(\mathbf{X})$, that is, the most likely group membership.

5.2.2 Tree-based Algorithm for Model Fitting

We propose a tree-based algorithm to estimate the group classification function, $g_k(\mathbf{X})$, $k = 1, \ldots, K$. Since the label of subgroups, B, is unknown in the data, we treat it as missing data and incorporate the EM algorithm to handle the latency of subgroup membership. The idea of our proposed tree-based algorithm is that the covariate space is recursively partitioned to optimize the observed data log-likelihood function and the same subgroup membership is assigned to patients in the same subregion of the covariate space. The splitting is stopped when only a few patients remain in the child nodes or the number of patients from either treatment or control arm is smaller than some thresholds.

More explicitly, suppose that there are K latent subgroups in the data, we obtain the observed data log-likelihood function

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_2, \ldots, h_K; \mathbf{A}, \mathbf{X}, \mathbf{Y}, \boldsymbol{\Delta}) = \sum_{i=1}^{n} \Bigg[& \Delta_i \log \left(\sum_{k=1}^{K} f(Y_i, \beta_k, \boldsymbol{\gamma}; A_i, \mathbf{X_i}) g_k(\mathbf{X_i}) \right) \\ & + (1 - \Delta_i) \log \left(\sum_{k=1}^{K} S(Y_i, \beta_k, \boldsymbol{\gamma}; A_i, \mathbf{X_i}) g_k(\mathbf{X_i}) \right) \Bigg], \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, $f(t, \beta_k, \boldsymbol{\gamma}; A, \mathbf{X}) = -S^{'}(t, \beta_k, \boldsymbol{\gamma}; A, \mathbf{X})$ and $g_k(\mathbf{X_i}) = \frac{\exp\{h_k(\mathbf{X_i})\}}{\sum_{k=1}^{K}\exp\{h_k(\mathbf{X_i})\}}$, based on the data $\{Y_i = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), A_i, \mathbf{X_i}, i = 1, \ldots, n\}$. The complete data log-likelihood is written as

$$l_c\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_2, \ldots, h_K; \mathbf{A}, \mathbf{X}, \mathbf{Y}, \boldsymbol{\Delta}, \mathbf{B}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} I\left(B_i = k\right)\left[\Delta_i \log\left\{f(Y_i, \beta_k, \boldsymbol{\gamma}; A_i, \mathbf{X_i}) g_k\left(\mathbf{X_i}\right)\right\}\right.$$
$$\left. + (1 - \Delta_i)\log\left\{S(Y_i, \beta_k, \boldsymbol{\gamma}; A_i, \mathbf{X_i}) g_k\left(\mathbf{X_i}\right)\right\}\right]$$

To numerically learn the nonparametric function $g_k(\mathbf{X})$ and to grow a decision tree, the first step is to set the starting values for $\beta_k, k = 1, \ldots, K$, $\boldsymbol{\gamma}$ and $h_k(\mathbf{X})$. The starting values for $\beta_k, k = 1, \ldots, K$ are chosen to be close to 0 but have different values to guarantee the identifiability of the mixture survival model and the starting values for $\boldsymbol{\gamma}$ are set to be 0. $h_k(X)$ relies on the coefficients of a weighted multinomial regression model where more details will be discussed later in this section. The starting values for the coefficients of the weighted multinomial regression model are chosen to be 0. For each splitting, based on the data $\mathcal{B}_{t+1}$ in the current node, we apply the EM algorithm.

In the E-step, the expected log-likelihood based on all parameters conditional on the observed data is calculated, which is equivalent to calculating the posterior probability of $B_i = k$ given the observed data,

$$q_{ik}^{(t+1)} = \left(\frac{e^{\beta_k^{(t)} A_i}\exp\left(-\Lambda_0(t) e^{\beta_k^{(t)} A_i + \boldsymbol{\gamma}^{(t)}\mathbf{X_i}}\right) g_k^{(t)}(\mathbf{X_i})}{\sum_{k=1}^{K} e^{\beta_k^{(t)} A_i}\exp\left(-\Lambda_0(t) e^{\beta_k^{(t)} A_i + \boldsymbol{\gamma}^{(t)}\mathbf{X_i}}\right) g_k^{(t)}(\mathbf{X_i})}\right)^{\Delta_i}$$
$$\left(\frac{\exp\left(-\Lambda_0(t) e^{\beta_k^{(t)} A_i + \boldsymbol{\gamma}^{(t)}\mathbf{X_i}}\right) g_k^{(t)}(\mathbf{X_i})}{\sum_{k=1}^{K}\exp\left(-\Lambda_0(t) e^{\beta_k^{(t)} A_i + \boldsymbol{\gamma}^{(t)}\mathbf{X_i}}\right) g_k^{(t)}(\mathbf{X_i})}\right)^{1-\Delta_i},$$

where

$$g_k^{(t)}\left(\mathbf{X_i}\right) = \frac{\exp\left\{h_k^{(t)}\left(\mathbf{X_i}\right)\right\}}{\sum_{k=1}^{K}\exp\left\{h_k^{(t)}\left(\mathbf{X_i}\right)\right\}}$$

for $i \in \mathcal{B}_{t+1}$ and $k = 1, \ldots, K$ by using $\{\hat{\beta}_k^{(t)}, \hat{\boldsymbol{\gamma}}^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\}$ obtained from parent node. The baseline hazard rate λ_0 for patient i is approximated by

$$\frac{\Delta_i}{\sum_{j=1}^{n}\sum_{k=1}^{K} q_{jk}^{(t)} I(Y_j \geq Y_i)\exp(\beta_k^{(t)} A_i + \boldsymbol{\gamma}^{(t)}\mathbf{X_i})}. \tag{5.6}$$

In the M-step, for each feature $X_j, j = 1, \ldots, p$, and for each potential split x, we optimize the objective function

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_2, \ldots, h_K; X_j, x)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}=\boldsymbol{\gamma}^{(t)}} = \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t+1)} \left[\Delta_i \log\left(f(Y_i, \beta_k, \boldsymbol{\gamma}; A_i, \mathbf{X_i})\right)\right.$$
$$\left. + (1 - \Delta_i) \log\left(S(Y_i, \beta_k, \boldsymbol{\gamma}; A_i, \mathbf{X_i})\right) + \log\left(g_k(X_{ij})\right)\right], \quad (5.7)$$

which is equivalent to fitting a weighted multinomial regression model

$$L\left(\theta; X_j, x_j\right) = \sum_{i \in \mathcal{B}_{t+1}} \sum_{k=1}^{K} q_{ik}^{(t+1)} \log\left\{\frac{\exp\left\{\theta_{0k} + \theta_{1k} I\left(X_{ij} < x_j\right)\right\}}{\sum_{k=1}^{K} \exp\left\{\theta_{0k} + \theta_{1k} I\left(X_{ij} < x_j\right)\right\}}\right\} \quad (5.8)$$

based on the current data $\mathcal{B}_{t+1}$. To fit the weighted multinomial model (5.8), we use the one-step Newton-Raphson update. θ_{0k}^{t+1} and $\theta_{1k}^{(t+1)}$ are calculated using the following formula

$$\theta_k^{(t+1)} = \left(\theta_{0k}^{(t+1)}, \theta_{1k}^{(t+1)}\right)^T = \theta_k^{(t)} - \left(\frac{\partial^2 L\left(\theta; X_j, x\right)}{\partial \theta_k \partial \theta_k^T}\bigg|_{\theta_k = \theta_k^{(t)}}\right)^{-1} \frac{\partial L\left(\theta; X_j, x\right)}{\partial \theta_k}\bigg|_{\theta_k = \theta_k^{(t)}}$$

for $k = 1, \ldots, K$

Next, we search for the optimal splitting variable and corresponding splitting value among the grid points that optimize the observed data log-likelihood. An exhaustive search is performed on the features $X_j, j = 1, \ldots, p$ and their corresponding potential splitting values x. The 20th, 30th, ..., 80th percentile of X_j would be choices for potential splitting values x. We seek the optimal splitter $\{X_{j'}, x_{j'}\}$ that maximizes the objective function (5.8). Then, the corresponding $\{\theta'_{0k}, \theta'_{1k}\}$ are calculated, and $h_k^{(t+1)}(\mathbf{X_i})$ is updated by

$$h_k^{(t+1)}(\mathbf{X_i}) = \theta'_{0k} + \theta'_{1k} I(X_{ij'} < x_{j'}), \text{ for } i \in \mathcal{B}_{t+1}$$

Our proposed tree-based algorithm also estimates the survival profiles for each latent subgroups, that is, the algorithm provides treatment effect estimate, $\beta_k, k = 1, \ldots, K$, for each latent subgroup and the estimates of baseline covariates effects, $\boldsymbol{\gamma}$. To obtain these estimates, we consider the whole dataset $\mathcal{B}$. After we grow each level of the tree, we combine all the data points from the child nodes and apply the Newton-Raphson algorithm to update $\beta_k, k = 1, \ldots, K$, and $\boldsymbol{\gamma}$ by computing the first and second derivatives of the partial

loglikelihood function

$$\sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}\Delta_i \left[-\log\left(\sum_{s=1}^{n}\sum_{k=1}^{K} q_{sk} I\left(Y_s \geq Y_i\right) \exp\left(\beta_k A_s + \boldsymbol{\gamma}\mathbf{X_s}\right) \right) + \beta_k A_i + \boldsymbol{\gamma}\mathbf{X_i} \right]$$

with respect to $\beta_k, k = 1, \ldots, K$, and $\boldsymbol{\gamma}$. This partia loglikelihood function is obtained by plugging the baseline hazard estimate (5.6) into the objective function (5.7).

The binary splitting is stopped when each terminal node contains no more than $\sqrt{n}$ patients or the number of events from either treatment or control arm smaller than some thresholds (we use 5 in the simulation studies and the real data analysis).

5.2.3 Random Forest Algorithm for Combining Individual Decision Trees

The proposed tree-based method can be further extended to random forest by applying ensemble method (Breiman, 1996, 2001). The idea in random forest is to combine many tree predictors and average them to reduce the variance. The random selection of features in the tree splitting process helps to further reduce the variance by reducing the correlation between trees. Each tree in a forest is built based on a bootstrap sample of the training data. The splitting feature at each node is determined from a subset of covariates that are randomly selected from the covariate list to grow each tree. The trees in the forest are fully grown and the pruning procedure is usually not implemented. So the sizes of trees are sufficiently large, the sizes of terminal nodes are small and the trees tend to achieve a low bias. To obtain the prediction results on a future data, each tree of the forest is applied to the data and each subject in the data obtain a predicted class label based on majority vote results from trees in the forest.

Suppose that $\mathbb{T} = \{\mathcal{T}_d\}_{d=1}^{D}$ is a collection of D individual trees constructed based on bootstrap samples of the original training data. $\mathcal{T}_d$ for $d = 1, \ldots, D$ is grown by following the procedure in Section 5.2.2. At each node, m $(m < p)$ features are randomly selected as candidate splitting features and the optimal one among m candidate features is selected to optimize the objective function (5.7). The binary splitting is stopped until the stopping criterion is met. For any future subject i with baseline covariates $\mathbf{X_i} = \mathbf{x}$, we obtain the predicted subgroup membership $\hat{B}_{id}$ by applying the decision rule obtained from tree $\mathcal{T}_d$ for $d = 1, \ldots, D$. All $\mathcal{T}_d$, for $d = 1, \ldots, D$ then vote the most popular class $\hat{B}_i$ as the prediction for subject i.

5.3 Simulation Studies

5.3.1 Simulation Setting

Extensive simulation studies are performed to evaluate the performance of our proposed random forest model. We consider that two latent subgroups exist in the data. Binary treatment is considered in the simulation studies. The treatment has a beneficial effect in one subgroup and is harmful in the other subgroup, compared with the control group.

Ten baseline covariates, $X_1, X_2, \ldots, X_{10}$, are considered and only a few of them have nonzero effects. Ten baseline covariates are independently generated from a standard normal distribution. The treatment assignment is generated from Bernoulli distribution with the probability of 0.5 and takes value of -1 or 1. The baseline hazard rate is set to be a constant λ with the value of 2. Censoring times are generated from an exponential distribution with censoring rate around 30%. The true effect of treatment in two latent subgroups are set to be $\beta_1 = -1.5$ and $\beta_2 = 1.5$. Kaplan Meier curves in Figure 5.10 demonstrate the different treatment responses in different subgroups. Different scenarios are considered to mimic various cases in actual applications. We apply the proposed algorithm to study different functional forms for $h(\cdot)$. For both scenarios, we regard group 1 as the reference group so $h_1(\mathbf{X})$ is set to be 0. For scenario I.1, $h_2(\mathbf{X})$ is a linear combination of a few baseline covariates,

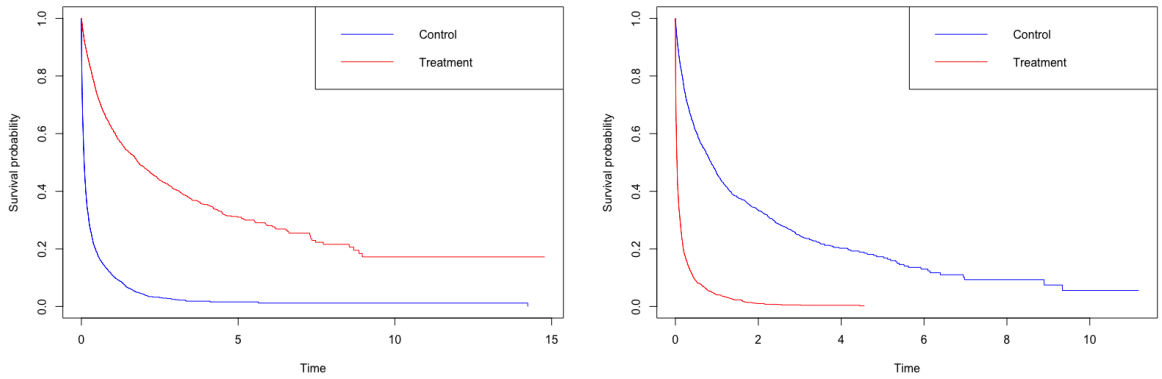
$$h_2(\mathbf{X}) = -0.6 + 0.4X_1 - 0.2X_2 + 0.5X_3.$$

For scenario I.2, we mimic a tree structure to assign latent subgroup membership, so $h_2(\mathbf{X})$ is an indicator function

$$h_2(\mathbf{X}) = -1 + 2I\{(X_1 < 0, X_2 < -0.5) \text{ or } (X_1 \geq 0, X_3 \geq 0.5)\}.$$

5.3.2 Simulation Results

We report the simulation results based on 100 replicates and consider sample size of 100 and 400. The number of trees in each forest is from 50 to 500 and the number of random features considered at each splitting is 3 and 5. Table 5.7 summarizes the mean classification accuracy and the standard errors for the simulation study. We calculate the classification accuracy by predicting the fitted random forest model on an



Note: The left plot illustrates the true survival curves for two treatment arms in the group that treatment is beneficial. The right plot shows the true survival curves for two treatment arms in the group that treatment is harmful.

Figure 5.10: The true survival curves for two treatment arms in simulation studies of two latent subgroups

independently generated validation set with sample size $10,000$. The optimal accuracy rate is calculated as 1 - Bayes error rate, where the Bayes error rate is calculated by the formula $1 - E\left(\max_k P(B = k|X)\right)$. Note that this accuracy assumes known group membership so the actual truth can be smaller than this number.

As table 5.7 indicates, the average classification accuracy shows large improvement when increasing sample size from 100 to 400, and the standard errors also demonstrate a decreasing trend as the sample size increases. A slowly increasing trend in average classification accuracy is observed when the number of trees in each forest increases from 50 to 500. Little difference in classification accuracy is observed when the number of random features considered at each splitting increase from 3 to 5.

Table 5.7: Average classification accuracy and standard error from the simulation study

Scenario I. The optimal accuracy rate is 0.666.			
N	Ntree	Mtry=3	Mtry=5
100	50	0.600 (0.034)	0.593 (0.034)
	100	0.601 (0.034)	0.596 (0.034)
	200	0.603 (0.034)	0.597 (0.035)
	500	0.604 (0.035)	0.597 (0.036)
400	50	0.620 (0.017)	0.617 (0.018)
	100	0.623 (0.016)	0.620 (0.012)
	200	0.624 (0.017)	0.622 (0.018)
	500	0.625 (0.017)	0.622 (0.018)
Scenario II. The optimal accuracy rate is 0.730.			
N	Ntree	Mtry=3	Mtry=5
100	50	0.591 (0.039)	0.588 (0.037)
	100	0.595 (0.038)	0.591 (0.037)
	200	0.597 (0.039)	0.592 (0.037)
	500	0.597 (0.029)	0.593 (0.038)
400	50	0.630 (0.021)	0.635 (0.023)
	100	0.636 (0.021)	0.639 (0.022)
	200	0.636 (0.020)	0.638 (0.014)
	500	0.639 (0.020)	0.641 (0.022)

Note. Ntree: number of tree in each forest. Mtry: number of features randomly selected at each splitting.

5.4 Real Data Application

We apply the proposed method to a Phase III randomized clinical trial data in (Lipkovich et al., 2017). This trial recruited patients with hematological malignancies and randomly assigned patients to treatment or control groups. Patients in treatment group received an experimental therapy plus best supporting care and only best supporting care was provided to patients in control group. A total of 599 patients were enrolled, 303 of them were in the treatment group and the other 296 patients were in the control group. 14 baseline covariates were recorded, including demographic characteristics, clinical variables associated with baseline disease severity, and cytogenetic markers. More explicitly, the covariates list contains both nominal covariates such as patient sex, race, patient's prior therapy outcome, and nine cytogenetic markers (presence or absence),

and ordinal covariates such as cytogenetic category and prognostic score for myelodysplastic syndromes risk assessment (IPSS-R score). Eight patients had missing/unknown IPSS-R score, so they were excluded for further analysis. For the ordinal variables, cytogenetic category had five different levels: very good, good, intermediate, poor and very poor. IPSS-R score had four levels in Lipkovich et al. (2017): low, intermediate, high and very high. Only one patient had low IPSS-R score, so we combine low and intermediate IPSS-R score to one level. The median follow-up time was 7.13 months and the censoring rate was 17%. The primary endpoint in this trial was the overall survival. The hazard ratio was used to evaluate treatment effect, an estimate of 0.85 with p-value of 0.07 for the hazard ratio was obtained from a Cox PH model including treatment only. The Kaplan Meier curves for patients receiving two treatments are plotted in Figure 5.4 This finding implies that the treatment does not demonstrate a significant effect in the overall population and a subgroup analysis to explore some subpopulations with beneficial treatment effects may be useful.

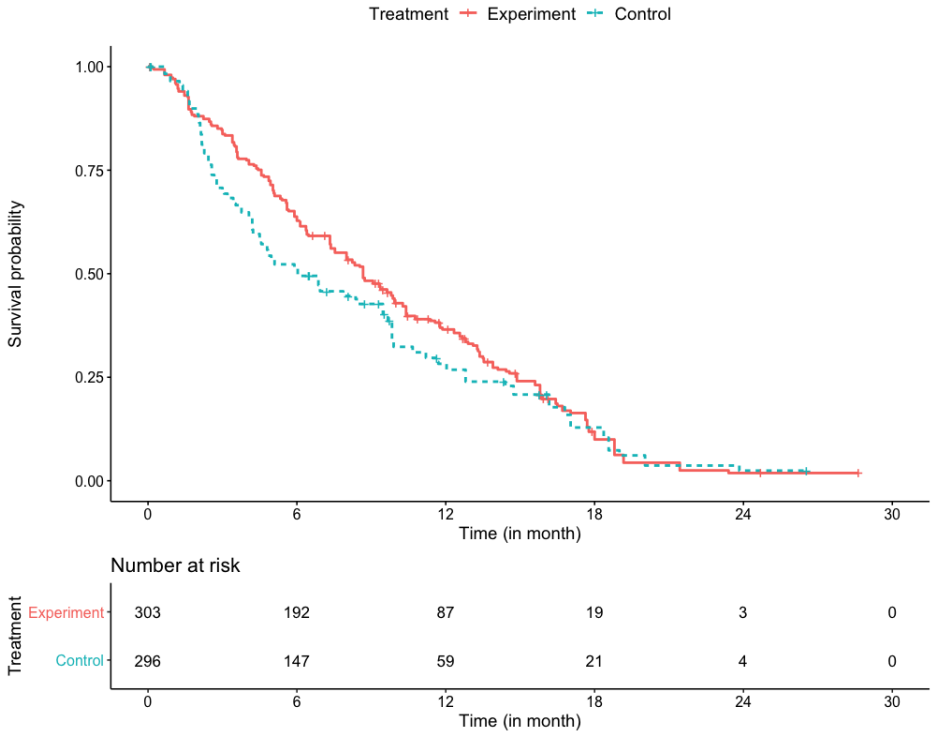


Figure 5.11: Kaplan-Meier curves for each treatment

The first step to analyze this data is to determine the number of latent subgroups. We consider BIC in the group selection and use average observed data loglikelihood to calculate BIC value. The average observed data loglikelihood is obtained from the forest by averaging over the values of observed data loglikelihood of

all individual trees in the forest. The BIC is calculated by

$$BIC = -2 \times \frac{1}{D}\sum_{d=1}^{D}\log(\widehat{Lik}_d) + \log(n) \times \left(\frac{1}{D}\sum_{d=1}^{D} l_d + K + p\right)$$

where $\log(\widehat{Lik}_d)$ is the observed data loglikelihood of tree $\mathcal{T}_d$, n is the number of observations in the data, l_d is the number of terminal nodes of tree $\mathcal{T}_d$, K and p are the number of subgroups and covariates respectively. To use BIC criterion in the group selection, we assume that a multiple number of latent groups exists in the data and construct a forest for each choice. The forest yields the smallest BIC is chosen as best and the corresponding number of latent subgroups is selected. We will use this average observed data loglikelihood to evaluate the importance of each feature variable. More details regarding the variable importance calculation will be discussed later in this section.

The random forest is constructed based on 200 trees and 3 features are randomly selected at each node for splitting. The first step is to determine the number of latent subgroups contained in the data. As stated in Section 5.2, we assume that there are $0 \sim 4$ latent groups existing in the dataset, and for each choice, we build a forest and calculate the corresponding BIC value. BIC suggests that there are 2 latent groups in the dataset. We then explore the two latent subgroups in more details. As illustrated in Figure 5.4, for the two latent subgroups that our proposed model identified, a significant treatment effect is detected in one group, with the hazard ratio estimate of 0.43 and 95% CI $(0.33, 0.54)$, which implies that the experimental treatment is beneficial to patients in this subgroup. In the other subgroup, the treatment effect is also significant, but on the other direction: the hazard ratio estimate is 1.37 and 95% CI is $(1.05, 1.79)$, which suggests that the experimental treatment is harmful to patients in this subgroup, compared with the control treatment.

Another attractive feature of random forest is the variable importance. To understand which feature variable is important in the latent subgroup discovery, we develop an algorithm to reveal the importance of each variable. In the context of subgroup analysis, the variable importance will provide clues that which variable contributes most to the differences in treatment responses. Following the idea of likelihood ratio test, we consider the tree collection $\mathbb{T} = \{\mathcal{T}_d\}_{d=1}^{D}$, which is the random forest including all p baseline variables. We obtain the average value of observed data loglikelihood $\frac{1}{D}\sum_{d=1}^{D}\log(\widehat{Lik}_d)$ from $\mathbb{T}$. Then, we construct forests $\mathbb{T}^{-j} = \{\mathcal{T}_d^{-j}\}_{d=1}^{D}$ for $j = 1, \ldots, p$, where each forest is built by excluding variable j. The corresponding average values of observed data loglikelihood $\frac{1}{D}\sum_{d=1}^{D}\log(\widehat{Lik}_d^{-j})$ for $j = i, \ldots, p$, are

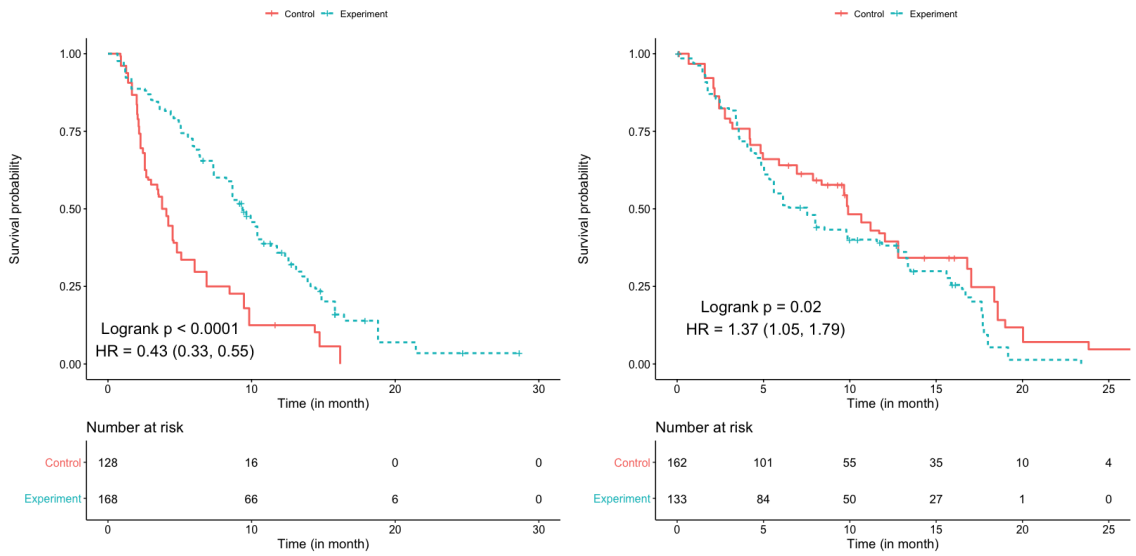


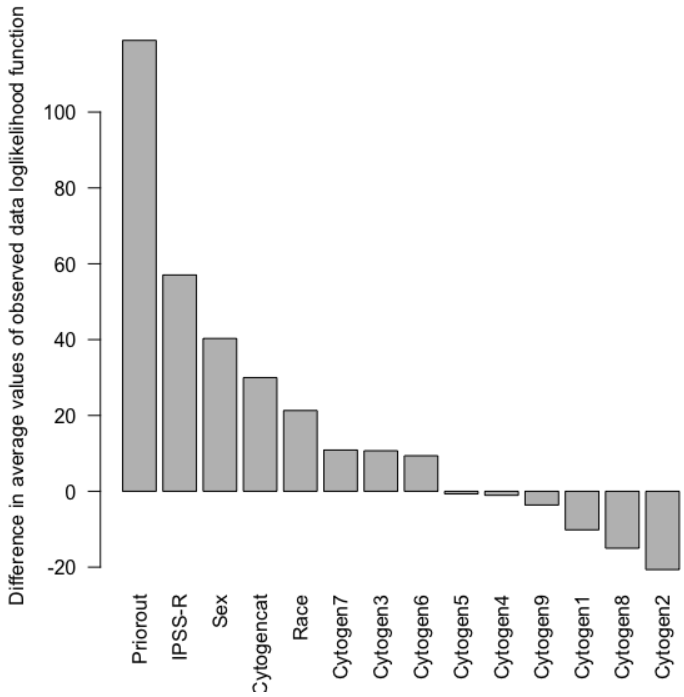
Figure 5.12: Kaplan Meier curves by treatment for patients in group 1 (left panel), in group 2 (right panel)

computed. We define the difference $\frac{1}{D}\sum_{d=1}^{D}\log(\widehat{Lik}_d) - \frac{1}{D}\sum_{d=1}^{D}\log(\widehat{Lik}_d^{-j})$ as the variable importance of variable j for $j = 1, \ldots, p$.

Figure 5.4 shows the variable importance of each variable. According to this figure, patient's prior therapy outcome, IPSS-R score, patient's sex, cytogenetic category and patient's race are important modifiers for the treatment effect. Among these important variables, patient's prior therapy outcome contributes the most to the model fit so we also plot the distribution of this variable in the two identified subgroups to further explore patients with which kind of characteristic experience a beneficial treatment effect. From the top left panel of Figure 5.4, patients with failure or progress prior therapy outcome are more likely to experience a beneficial treatment effect. In addition, patients with a very high IPSS-R score tend to experience a beneficial treatment effect as well, as demonstrated in the top right panel of Figure 5.4. For cytogenetic category, those with intermediate or poor cytogenetic category are likely to experience a positive treatment effect, as shown in bottom left panel of Figure 5.4. We also observe a small increase in the average value of observed data loglikelihood function when excluding some variables such as biomarkers Cytogen5, Cytogen4, Cytogen9, Cytogen1, Cytogen8, Cytogen2. Such increase may be because individual trees in the forest could be different.

The important variables such as patient's prior therapy outcome, IPSS-R score and cytogenetic category are also identified using other subgroup analysis methods including IT procedure and SIDES method

(Lipkovich et al., 2017) and 5-step stratified testing and amalgamation routine (5-STAR) (Mehrotra and Marceau West, 2020). Lipkovich et al. (2017) provide subgroup discovery results by using SIDES method (Lipkovich et al., 2011) and IT method (Su et al., 2009). One subgroups with beneficial treatment effects are found to be defined by cytogenetic category and IPSS-R score by these two methods. Moreover, Mehrotra and Marceau West (2020) utilize conditional inference tree algorithm to analyze this dataset and five risk strata are formed. A few of these risk strata demonstrate significant treatment effect relative to the control treatment, while the other strata suggest negligible treatment effect. Some splits in the conditional inference tree are based on cytogenetic category, IPSS-R score and patient's prior therapy outcome.



Note. Y axis is the difference of average values of observed data loglikelihood function between the model including all variables and the model excluding one variable. "Priorout" is patient's prior therapy outcome. "IPSS-R" is IPSS-R score. "Cytogencat" is cytogenetic category. "Cytogen1", . . . , "Cytogen9" are nine cytogenetic markers.

Figure 5.13: Variable importance from the proposed random forest model

To further understand the relationship between the baseline covariates and predicted latent subgroup and facilitate the latent subgroup prediction for future patients, we consider a logistic regression model including predicted latent subgroup membership as the response variable and important baseline covariates

identified by the random forest model as predictors. Table 5.8 reports the estimation results for the logistic regression model parameters. All predictors are significantly associated with the latent subgroup membership assignment. We then randomly split the data into a training set (N=400) and a testing set (N=191). The prediction accuracy on the test set is 0.87. Figure 5.4 illustrates the ROC curve based on the logistic regression model fit on the testing set. The corresponding AUC is 0.947. These results imply that the logistic regression model fits the data well and the baseline covariates are predictive of the latent subgroup membership.

Table 5.8: Estimation results for the logistic regression parameters

Covariate	Estimate	Std. Error	z value	p-value
(Intercept)	0.4111	0.3124	1.3162	0.1881
Sex	2.2442	0.3700	6.0651	< 0.0001
Cytogencat	0.5743	0.1807	3.1785	0.0015
IPSS	-5.8133	0.5905	-9.8452	< 0.0001
Cytogen3	1.3157	0.4277	3.0760	0.0021
Cytogen6	-2.0323	0.5203	-3.9062	0.0001
Cytogen7	-4.3313	0.6423	-6.7437	< 0.0001
Race, asian	3.9323	1.2585	3.1246	0.0018
Race, black	-0.9258	0.8637	-1.0718	0.2838
Priorout, failure	-6.8745	0.8206	-8.3776	< 0.0001
Priorout, relapse	2.4182	0.3943	6.1332	< 0.0001

Note. The whites is the reference group for race variable. Progress is the reference group for patient's prior therapy outcome.

We also apply the parametric mixture model described in Section 3.2 to this dataset, where the patient-level treatment assignment is blinded. The variable selection procedure of the proposed parametric mixture model is not implemented to the analysis. BIC criterion suggests that two latent groups in this dataset and we plot the Kaplan Meier curves of two groups in the left panel of Figure 5.16. The parametric mixture model is also used to explore latent subgroups in either treatment or control arms. Among patients receiving experimental treatment, no latent subgroups is discovered. The survival curve for patient receiving experimental treatment is shown in the middle panel of Figure 5.16 For the control arm, two latent subgroups are identified by comparing BIC values of forests assuming the presence of $0 \sim 3$ latent groups. According to the right plot of Figure 5.16, one subgroup of patients in the control arm have short-term survival, the median survival time is around 4 months. The other subgroup of patients in the control arm demonstrate a longer survival time with a median survival time around 10 months. This finding also agrees with our observation of

results from the proposed random forest model in Figure 5.4. The survival curves for patients in the treatment group (the green curves) in the middle and right plots in Figure 5.4 have similar shape, comparing with curves of patients in the control group (the red curves). In the subgroup that treatment is beneficial, patients receiving control treatment die fast at the beginning. While in the other subgroup, patients receiving control treatment achieve much longer survival. This implies that the difference of treatment effects in two subgroups identified by random forest model may be because patients respond to the control treatment very differently in two subgroups.

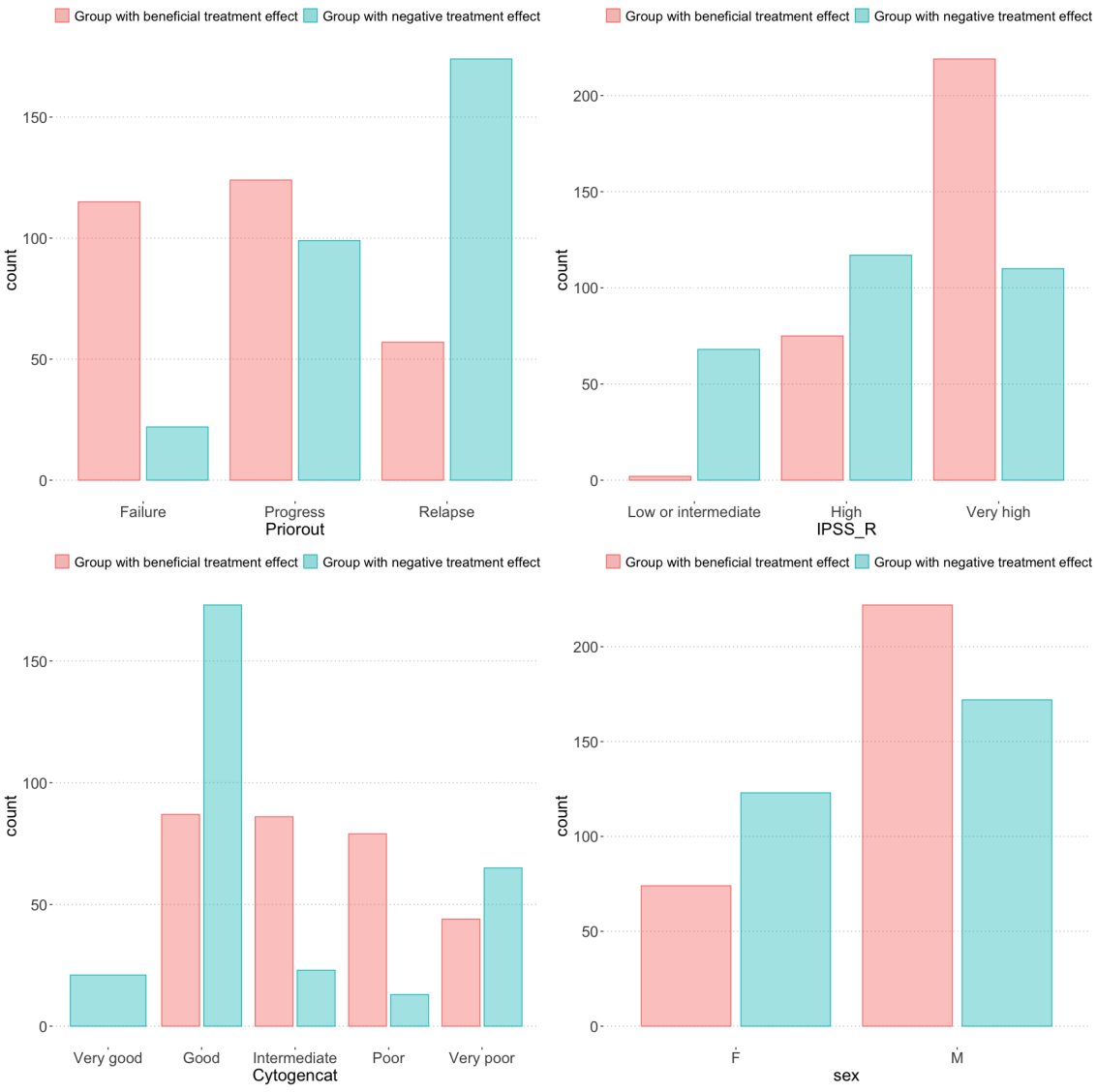


Figure 5.14: Barplots of the distributions of patient's prior therapy outcome (top left panel), IPSS-R score (top right panel), cytogenetic category (bottom left panel) and sex (bottom right panel)

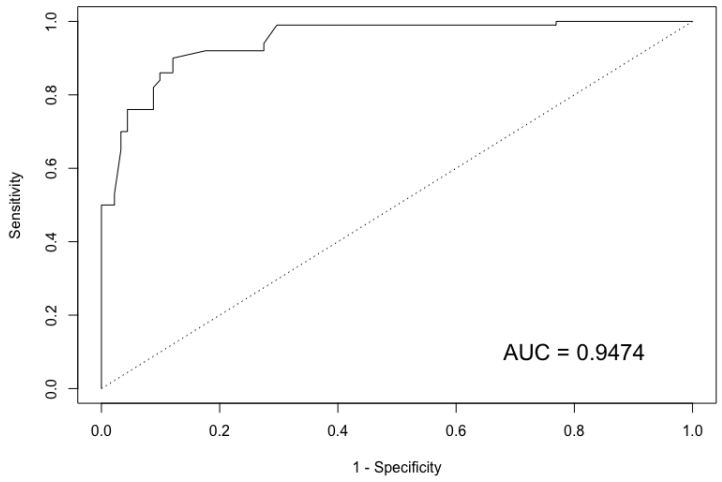


Figure 5.15: ROC curve based on prediction results of logistic regression model from a randomly sampled test set

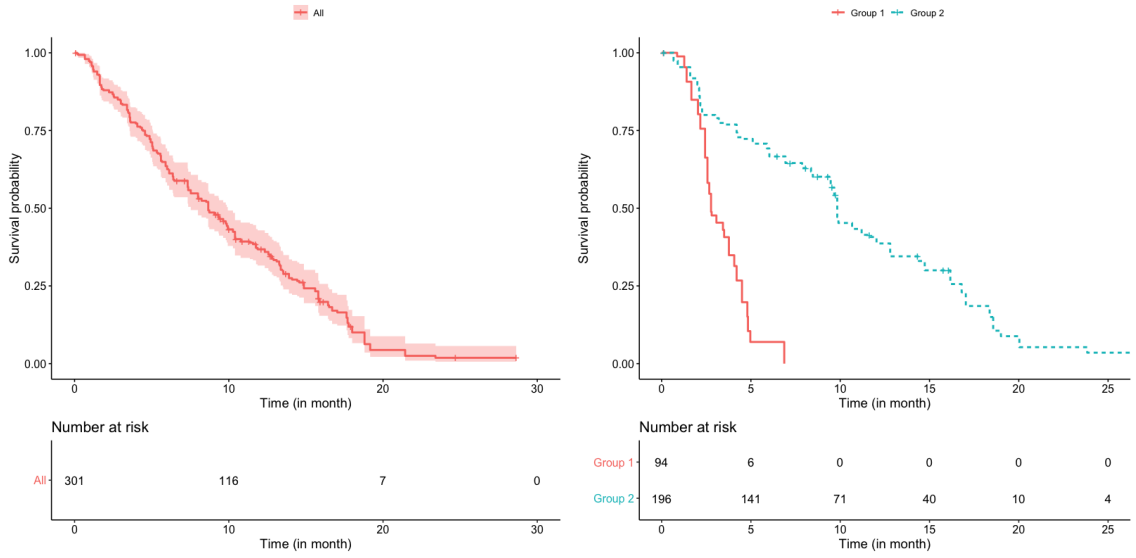


Figure 5.16: Kaplan Meier curves from parametric mixture model without treatment information, in treatment group (left panel), in control group (right panel)

5.5 Discussion

In this chapter, we develop a random forest model to explore latent subgroups with heterogeneous treatment effects using survival outcomes. We assume that the survival profiles of patients are related to baseline covariates and treatment effects. The differences among subgroups are owing to the different treatment responses of patients from different subgroups. We learn the relationship between latent subgroup membership and the baseline covariates by recursively partitioning the covariates space in each individual decision tree. The unobserved subgroup membership is treated as missingness and the EM algorithm is incorporated to handle the missing data. We apply the EM algorithm via the method of weights (Ibrahim, 1990; Lipsitz and Ibrahim, 1996, 1998), where we calculate the conditional distribution of the missing data given the observed data and the current estimate of the parameters in E-step. The computation depends on the complete data and the incomplete data likelihood is not required. As a result of applying the EM algorithm, the binary splitting at each node in individual decision tree is to optimize the observed data loglikelihood function.

An application to the data from a phase III clinical trial in patients with hematological malignancies is provided. Our proposed method directly classifies patients into two large subgroups and discovers significant treatment effects. Moreover, a variable importance measurement is derived to visualize the important features that determines latent subgroup membership assignment and modify different treatment responses among subgroups. We also provide an example demonstrating the latent subgroup prediction for future patients by using a logistic regression model, which facilitates the understanding and interpretation of the relationship between latent subgroup membership and baseline covariates, and enables a quick and simple implementation for future patients latent subgroup classification.

Unlike other subgroup analysis methods, the heterogeneous treatment effects are modeled in terms of interactions of treatment and covariates using parametric modeling, which may be insufficient to explore the heterogeneity in complicated forms. Our proposed tree-based nonparametric approach improves the flexibility in detecting the heterogeneity in patients' treatment responses. In addition, some other nonparametric subgroup analysis methods also employ recursive partition as a natural way to explore heterogeneous treatment effects, while they consider the subgroup analysis in a post hoc manner. Many latent subgroups are formed after the construction of model. The number of latent subgroup are determined afterwards by using

some statistical tests to examine the heterogeneity between small subgroups and some thresholds are adopted to combine subgroups. The choice of such threshold may be subjective and needs additional investigation. On the other hand, our proposed method aims to reveal latent subgroups in the data with larger sample size and no further combination is needed. So one may be more confident when conduct tests to evaluate treatment effects. Furthermore, based on the identified subgroups and the variable importance results, one may extract more information regarding how patients response to treatment differently. Such information potentially helps to design a future trial or propose new research hypotheses.

CHAPTER 6: EXTENSIONS AND FUTURE RESEARCH

6.1 Inferring Latent Heterogeneity Using Many Feature Variables with Survival Outcome

In the proposed method in Chapter 3, we assume that the survival distributions from different latent groups are distinct, which implies that the mixture model can be identifiable. The selection on number of latent subgroups using BIC can also help to distinguish non-identifiable cases, where BIC will suggest that no latent subgroups exists in the data. It has been well understood that the model selection method with a fixed number of covariates using BIC criterion (Schwarz et al., 1978) can identify the true model consistently (Shao, 1997; Shi and Tsai, 2002). For the situation with a diverging number of covariates, the asymptotic behavior of a slightly modified version of BIC criterion has been greatly discussed as well and the consistency in linear regression model selection with a diverging number of covariates for penalized estimators has been studied (Wang et al., 2009). When the number of covariates diverges in a mixture model setting, although the BIC criterion works well in our empirical studies, we are not aware of the theoretical results of BIC criterion in this situation. We will pursue this interesting topic in our future work.

Moreover, our proposed variable selection procedure does not cover the ultra-high dimensional case in which the dimensionality p_n is much larger than the sample size n. In this case, some problems need to be solved. For example, the estimate of the number of latent groups based on BIC may not be consistent (Drton and Plummer, 2017), and the variable selection procedure will be challenged (Fan et al., 2009). Further work on applying our proposed method to ultra-high dimensional data needs to be done. This work focuses on identifying latent groups who have a distinct survival experience. The same idea can be extended to study latent groups who may respond to treatments differently. The latter will be particularly characterized by different treatment effects which can be constant or time-varying. Variable selection will also be important to determine a small list of feature variables for medical decisions. We will pursue these extensions in future work.

6.2 Mixture Survival Trees for Cancer Risk Classification

Some extensions from the proposed tree-based method can also be considered. In our proposed algorithm, the survival distributions for the latent subgroups are modeled via Weibull distributions with unknown parameters. It is easy to extend our algorithm to other parametric distributions such as the exponential distribution and log-normal distribution. In addition, the parametric survival distribution assumption can be further weakened, and a Cox PH model could also be considered. In addition to the survival distribution assumptions, identifying important covariates that are associated with the survival risk group classification is also an important issue, especially when a large number of covariates are used in the dataset. Although the tree structure offers a straightforward explanation regarding the covariate effects, a summary statistic reflecting the variable's importance is still desirable. A potential direction of our future work is to incorporate ensemble methods, such as bagging, into the current tree-based algorithm. Variable importance calculation in a random forest can be easily applied to our algorithm and the predictive performance may be improved. When there are a large number of covariates in the data, the proposed tree-based algorithm may be unstable due to the sample size, and one strategy is to follow a two-step approach from Liao et al. (2020) to first select important initial variables using machine learning techniques and then apply the selected variables to our proposed algorithm.

6.3 Random forest for Subgroup Analysis with Heterogeneous Treatment Responses

The proposed random forest model can be extended to the following directions. First, the assumption of common baseline hazard rate for each subgroup can be relaxed by assuming different baseline hazard rates for different subgroups. Some kernel smooth techniques can be incorporate to facilitate the estimation of survival distributions. The incorporation of different baseline hazard rates for different subgroups may enable larger flexibility in dealing with unobserved heterogeneity in patients. Some diagnostic procedures may be considered to examine the assumptions of proportional hazard model before analyzing the data. Most diagnostic procedures are developed based on residuals such as Schoenfeld residuals (Schoenfeld, 1980, 1982), Cox-Snell residuals (Cox and Snell, 1968; Kay, 1977) and martingale residuals (Lagakos, 1981; Barlow and Prentice, 1988; Therneau et al., 1990).

Second, we consider that the effects of baseline covariates are the same across all subgroups in the proposed model, which implies that the differences among subgroups only depend on different treatment effects. The common baseline covariates coefficients assumption may be extended, resulting in a more general model that each subgroup of patients has distinct baseline covariates coefficients. This means that the tree splitting depends not only on different treatment effects but also on differences in the effects of baseline covariates. An additional step may be needed to single out the partitions that are based on different treatment effects.

Third, we consider binary treatment assignment in the proposed method. This can be extended to multiple levels of treatment assignment, for example, in factorial experiments. Furthermore, the treatment measured on a continuous scale, such as different doses of a new drug, can be also introduced in this proposed method to better accommodate real world examples.

Fourth, we handle right censored data in the proposed method. The random forest model can also be extended to deal with other types of censoring such as left censoring and interval censoring data. In addition to the censored data, other type of outcomes can be incorporated within the framework of our proposed random forest model.

Lastly, the primary goal of the proposed method is to explore the unobserved heterogeneity in patients' treatment responses. The output of the proposed method is a direct classification of patients into different subgroups, along with an estimate of average treatment effect for patients in each subgroup. As stated in the real data application in Chapter 5, the overall treatment effect in the data may be not significant before we discover latent subgroups. One potential direction of the future research is to develop formal statistical tests to evaluate the overall treatment effect after adjusting for latent subgroups identified in the data on an independent validation dataset. Furthermore, the number of events required to identify subgroups in future trials could be determined by power and sample size calculation. The statistical testing procedure could be also used to explore how we combine the identified subgroups for a clinically meaningful interpretation.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 3

Proof of Theorem 1

Set $||\mathbf{u}|| = C$, where C is a large enough constant. We need to show that for any given ϵ there exists a large constant C such that, for large n,

$$P\left(\sup_{||\mathbf{u}||=C} l_{n,obs}(\theta_{n0} + \sqrt{p_n}n^{-1/2}\mathbf{u}) \leq l_{n,obs}(\theta_{n0})\right) \geq 1 - \epsilon. \tag{6.9}$$

This implies that with probability at least $1-\epsilon$ there is a local maximum $\tilde{\theta}_n$ in the ball $\{\theta_{n0} + \sqrt{p_n}n^{-1/2}\mathbf{u} : ||\mathbf{u}|| \leq C\}$ such that $||\tilde{\theta}_n - \theta_{n0}|| = O_p(\sqrt{p_n}n^{-1/2})$.

Furthermore, we have

$$\begin{aligned}
D_n(\mathbf{u}) &= l_{n,obs}(\theta_{n0} + \sqrt{p_n}n^{-1/2}\mathbf{u}) - l_{n,obs}(\theta_{n0}) \\
&= \sqrt{p_n}n^{-1/2}\nabla^T l_{n,obs}(\theta_{n0})\mathbf{u} + \frac{1}{2}\frac{p_n}{n}\mathbf{u}^T\nabla^2 l_{n,obs}(\theta_{n0})\mathbf{u} \\
&\quad + \frac{1}{6}\left(\frac{p_n}{n}\right)^{3/2}\nabla^T\left\{\mathbf{u}^T\nabla^2 l_{n,obs}(\theta_n^*)\mathbf{u}\right\}\mathbf{u} \\
&\widehat{=} I_1 + I_2 + I_3, \text{ where the vector } \theta_n^* \text{ lies between } \theta_{n0} \text{ and } \theta_{n0} + \sqrt{p_n}n^{-1/2}\mathbf{u}.
\end{aligned}$$

By condition (C2) and Markov's inequality,

$$\begin{aligned}
|I_1| &= |\sqrt{p_n}n^{-1/2}\nabla^T l_{n,obs}(\theta_{n0})\mathbf{u}| \\
&\leq \sqrt{p_n}n^{-1/2}||\nabla^T l_{n,obs}(\theta_{n0})|| \cdot ||\mathbf{u}|| \\
&= \sqrt{p_n}n^{-1/2}O_p((np_n)^{1/2})||\mathbf{u}|| \\
&= O_p(p_n)||\mathbf{u}||.
\end{aligned} \tag{6.10}$$

For I_2, we first show that $||\frac{1}{n}\nabla^2 l_{n,obs}(\theta_{n0}) + I_{n0}(\theta_{n0})|| = o_p(\frac{1}{p_n})$.

By Chebyshev's inequality and condition (C2), for any given $\epsilon > 0$,

$$P\left(||\frac{1}{n}\nabla^2 l_{n,obs}(\theta_{n0}) + I_{n0}(\theta_{n0})|| \geq \frac{\epsilon}{p_n}\right) \leq \frac{p_n^2}{n^2\epsilon^2}E\left(\sum_{j,l=1}^{p_n}\left[\frac{\partial^2 l_{n,obs}(\theta_{n0})}{\partial\theta_{nj}\partial\theta_{nl}} - E\frac{\partial^2 l_{n,obs}(\theta_{n0})}{\partial\theta_{nj}\partial\theta_{nl}}\right]\right)^2$$

$$= \frac{p_n^4}{n} = o(1).$$

Then, we have

$$\begin{aligned} I_2 &= \frac{1}{2}\frac{p_n}{n}\mathbf{u}^T\nabla^2 l_{n,obs}(\theta_{n0})\mathbf{u} \\ &= \frac{1}{2}p_n\mathbf{u}^T\left[\frac{1}{n}\nabla^2 l_{n,obs}(\theta_{n0}) + I_{n0}(\theta_{n0})\right]\mathbf{u} - \frac{1}{2}p_n\mathbf{u}^T I_{n0}(\theta_{n0})\mathbf{u} \\ &= \frac{1}{2}||\mathbf{u}||^2 o_p(1) - \frac{1}{2}p_n\mathbf{u}^T I_{n0}(\theta_{n0})\mathbf{u}. \end{aligned} \tag{6.11}$$

By condition (C3) and the Cauchy-schwarz inequality, we have

$$\begin{aligned} |I_3| &= |\frac{1}{6}(\frac{p_n}{n})^{3/2}\nabla^T\{\mathbf{u}^T\nabla^2 l_{n,obs}(\theta_n^*)\mathbf{u}\}\mathbf{u}| \\ &= |\frac{1}{6}(\frac{p_n}{n})^{3/2}\sum_{j,l,m=1}^{p_n}\frac{\partial^3 l_{n,obs}(\theta_n^*)}{\partial\theta_{nj}\partial\theta_{nl}\partial\theta_{nm}}u_j u_l u_m| \\ &\leq \frac{1}{6}(\frac{p_n}{n})^{3/2}\sum_{i=1}^{n}\left(\sum_{j,l,m=1}^{p_n} M_{njlm0}^2(X_i, Y_i, \Delta_i)\right)^{1/2}||\mathbf{u}||^3, \end{aligned}$$

since $p_n^4/n \to 0$ as $n \to \infty$,

$$\begin{aligned} &= n(\frac{p_n}{n})^{3/2}||\mathbf{u}||^2 O_p(p_n^{3/2}) \\ &= o_p(p_n)||\mathbf{u}||^2. \end{aligned} \tag{6.12}$$

By choosing a sufficiently large constant C, (6.10) and (6.12) are dominated by (6.11). Therefore, the inequality (6.9) holds, and this completes the proof.

Proof of Theorem 2

We write the weighted multinomial log-likelihood function as

$$l_n(\theta_n) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tilde{q}_{nik}\left[\Delta_i \log f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i)\log S(Y_i, \boldsymbol{\eta}_k) + \log \pi_k(X_i; \beta_n)\right],$$

where $\tilde{q}_{nik}$ are the optimal weights calculated based on the maximum likelihood estimator $(\tilde{\eta}^T, \tilde{\beta}^T)^T$ is

$$\begin{aligned}\tilde{q}_{nik} &= \frac{(\Delta_i f_k(Y_i;\tilde{\eta}) + (1-\Delta_i)S_k(Y_i;\tilde{\eta}))\pi_k(X_i;\tilde{\beta}_n)}{\sum_{k=1}^K(\Delta_i f_k(Y_i;\tilde{\eta}) + (1-\Delta_i)S_k(Y_i;\tilde{\eta}))\pi_k(X_i;\tilde{\beta}_n)} \\ &= q_{nik}(Y_i, X_i, \Delta_i, f_k(\cdot), S_k(\cdot), \tilde{\eta}, \tilde{\beta}_n)\end{aligned}$$

To facilitate the proofs of Theorems 2 and 3, we first establish the following lemma under conditions (C1) - (C4).

Lemma A.1. *Denote the first-order derivative of $l_n(\theta_n)$ with respect to θ_n by $U_n(\theta_n)$. Then $(np_n)^{-1/2}U_n(\theta_{n0}) = O_p(1)$, where $O_p(1)$ is bounded in probability.*

Proof of Lemma A.1. Let

$$\begin{aligned}l_n(\theta_n) &= l_{n0}(\theta_n) + l_{nd}(\theta_n) \\ &= \sum_{i=1}^n\sum_{k=1}^K q_{nik0}\left(\Delta_i \log f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i)\log S(Y_i, \boldsymbol{\eta}_k) + \log \pi_k(X_i;\beta_n)\right) \\ &+ \sum_{i=1}^n\sum_{k=1}^K(\tilde{q}_{nik} - q_{nik0})\left(\Delta_i \log f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i)\log S(Y_i, \boldsymbol{\eta}_k) + \log \pi_k(X_i;\theta_n)\right).\end{aligned}$$

Then we write $(np_n)^{-1/2}U_n(\theta_{n0}) = (np_n)^{-1/2}\left[U_{n0}(\theta_{n0}) + U_{nd}(\theta_{n0})\right]$.

By condition (C2) and Markov's inequality, we obtain

$$\begin{aligned}P\left((np_n)^{-1/2}\frac{\partial l_{n0}(\theta_{n0})}{\partial \eta_l} \le M\right) &\le \frac{E\left((np_n)^{-1}(\frac{\partial l_{n0}(\theta_{n0})}{\partial \eta_l})^2\right)}{M^2} \longrightarrow 0, \\ P\left((np_n)^{-1/2}\frac{\partial l_{n0}(\theta_{n0})}{\partial \beta_{nkj}} \le M\right) &\le \frac{E\left((np_n)^{-1}(\frac{\partial l_{n0}(\theta_{n0})}{\partial \beta_{nkj}})^2\right)}{M^2} \longrightarrow 0,\end{aligned}$$

as $n \longrightarrow \infty$, for some $M > 0$, $l = 1, \ldots, s$, $j = 1, \ldots, p_n/K$ and $k = 1, \ldots, K$.

Therefore, we have $(np_n)^{-1/2}U_{n0}(\theta_{n0}) = O_p(1)$.

Next, we consider $(np_n)^{-1/2}U_{nd}(\theta_{n0})$, where

$$\frac{\partial l_{nd}(\theta_n)}{\partial \boldsymbol{\eta}} = \sum_{i=1}^n\sum_{k=1}^K(\tilde{q}_{nik} - q_{nik0})\left(\Delta_i \frac{\partial f_k(Y_i;\boldsymbol{\eta})/\partial \boldsymbol{\eta}}{f_k(Y_i;\boldsymbol{\eta})} + (1-\Delta_i)\frac{\partial S_k(Y_i;\boldsymbol{\eta})/\partial \boldsymbol{\eta}}{S_k(Y_i;\boldsymbol{\eta})}\right),$$

$$\frac{\partial l_{nd}(\theta_n)}{\partial \boldsymbol{\beta}_{nk}} = \sum_{i=1}^{n} \frac{(\tilde{q}_{nik} - q_{nik0}) \left(\sum_{j\neq k} e^{\beta_{nj}^T X_{ni}}\right) - \sum_{j\neq k}(\tilde{q}_{nij} - q_{nij0}) e^{\beta_{nk}^T X_{ni}}}{\sum_{k=1}^{K} e^{\beta_{nk}^T X_{ni}}} X_{ni},$$

for $j, k = 1, \ldots, K$.

Using the asymptotic results for $\tilde{\theta}_n$ from Theorem 1 and the mean value theorem, for each $k = 1, \ldots, K$, we have

$$\begin{aligned}
\sup_{i=1,\ldots,n} |\tilde{q}_{nik} - q_{nik0}| &= \sup_{i=1,\ldots,n} |\nabla_{\theta_n}^T q_{nik}(X_{ni}, Y_i, \Delta_i, f_k(\cdot), S_k(\cdot), \theta_n^*)(\tilde{\theta}_n - \theta_{n0})| \\
&\leq \sup_{i=1,\ldots,n} ||\nabla_{\theta_n}^T q_{nik}(X_{ni}, Y_i, \Delta_i, f_k(\cdot), S_k(\cdot), \theta_n^*)|| \cdot ||\tilde{\theta}_n - \theta_{n0}|| \\
&\longrightarrow 0 \text{ almost surely,}
\end{aligned}$$

where the vector θ_n^* lies between $\tilde{\theta}_n$ and θ_{n0}, and

$$\begin{aligned}
\frac{\partial q_{nik}}{\partial \eta} &= \frac{\left\{\left[\Delta_i \frac{\partial f(Y_i, \boldsymbol{\eta}_k)}{\partial \eta} + (1-\Delta_i)\frac{\partial S(Y_i, \boldsymbol{\eta}_k)}{\partial \eta}\right] \pi_k(X_i, \beta_n)\right\}}{\sum_{k=1}^{K}\left[\Delta_i \frac{\partial f(Y_i, \boldsymbol{\eta}_k)}{\partial \eta} + (1-\Delta_i)\frac{\partial S(Y_i, \boldsymbol{\eta}_k)}{\partial \eta}\right] \pi_k(X_i, \beta_n)}\cdot \\
&\quad \frac{\left\{\sum_{k=1}^{K}\left[\Delta_i f(Y_i, \boldsymbol{\eta}_k) - (1-\Delta_i) S(Y_i, \boldsymbol{\eta}_k)\right] \pi_k(X_i, \beta_n)\right\}}{\sum_{k=1}^{K}[\Delta_i \frac{\partial f(Y_i, \boldsymbol{\eta}_k)}{\partial \eta} + (1-\Delta_i)\frac{\partial S(Y_i, \boldsymbol{\eta}_k)}{\partial \eta}]\pi_k(X_i, \beta_n)} \\
&\quad - \frac{\left([\Delta_i f(Y_i, \eta_k) + (1-\Delta_i) S(Y_i, \boldsymbol{\eta}_k)\right] \pi_k(X_i, \beta_n))}{\sum_{k=1}^{K}[\Delta_i \frac{\partial f(Y_i, \boldsymbol{\eta}_k)}{\partial \eta} + (1-\Delta_i)\frac{\partial S(Y_i, \boldsymbol{\eta}_k)}{\partial \eta}]\pi_k(X_i, \beta_n)}\cdot \\
&\quad \frac{\left[\sum_{k=1}^{K}\left(\Delta_i \frac{\partial f(Y_i, \boldsymbol{\eta}_k)}{\partial \eta} - (1-\Delta_i)\frac{\partial S(Y_i, \boldsymbol{\eta}_k)}{\partial \eta}\right) \pi_k(X_i, \beta_n)\right]}{\sum_{k=1}^{K}[\Delta_i \frac{\partial f(Y_i, \boldsymbol{\eta}_k)}{\partial \eta} + (1-\Delta_i)\frac{\partial S(Y_i, \boldsymbol{\eta}_k)}{\partial \eta}]\pi_k(X_i, \beta_n)}
\end{aligned}$$

$$\frac{\partial q_{nik}}{\partial \beta_{nk}} =$$

$$\frac{\sum_{j\neq k}\left[\Delta_i f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i) S(Y_i, \boldsymbol{\eta}_k)\right] e^{\beta_{nk}^T X_{ni}} [\Delta_i f(Y_i; \eta_k) + (1-\Delta_i) S(Y_i; \eta_k)] e^{\beta_{nj}^T X_i} X_i}{\left(\sum_{k=1}^{K}[\Delta_i f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i) S(Y_i, \boldsymbol{\eta}_k)] e^{\beta_{nk}^T X_i}\right)^2},$$

$$\frac{\partial q_{nik}}{\partial \beta_{nj}} =$$

$$\frac{-[\Delta_i f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i) S(Y_i, \boldsymbol{\eta}_k)] e^{\beta_{nk}^T X_{ni}} [\Delta_i f(Y_i; \eta_k) + (1-\Delta_i) S(Y_i; \eta_k)] e^{\beta_{nj}^T X_i} X_i}{\left(\sum_{k=1}^{K}[\Delta_i f(Y_i, \boldsymbol{\eta}_k) + (1-\Delta_i) S(Y_i, \boldsymbol{\eta}_k)] e^{\beta_{nk}^T X_i}\right)^2},$$

for $j, k = 1, \ldots, K$ and $i = 1, \ldots, n$.

Thus, we have $(np_n)^{-1/2} U_{nd}(\theta_{n0}) = o_p(1)$ and it follows that $(np_n)^{-1/2} U_n(\theta_{n0}) = O_p(1)$. □

Proof of Theorem 2. Consider the penalized objective function

$$Q_n(\theta_n) = l_n(\theta_n) - n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|) = \ l_n(\theta_n) - n\lambda_n \sum_{j=1}^{p_n} |\beta_{nj}|/|\tilde{\beta}_{nj}|^{\gamma}.$$

Since $l_n(\theta_n)$ is essentially the likelihood of a weight multinomial regression, $l_n(\theta_n)$ is strictly concave, and the penalty term is strictly convex, it follows that $Q_n(\theta_n)$ is strictly concave when n is large. Thus, there exists a unique maximizer $\hat{\theta}_n$ of $Q_n(\theta_n)$ for large n. Set $||\mathbf{u}|| = C$, where C is a large enough constant. It is sufficient to show that for any given ϵ there exists a large constant C such that, for large n,

$$P\left(\sup_{||\mathbf{u}||=C} Q_n(\theta_{n0} + \sqrt{p_n}(n^{-1/2})\mathbf{u}) \leq Q_n(\theta_{n0})\right) \geq 1 - \epsilon. \tag{6.13}$$

This implies that with probability at least $1-\epsilon$ there is a local maximum $\tilde{\theta}_n$ in the ball $\{\theta_{n0} + \sqrt{p_n}(n^{-1/2})\mathbf{u} : ||\mathbf{u}|| \leq C\}$, for $C > 0$. Furthermore, we have

$$\begin{aligned} D_n(\mathbf{u}) &= Q_n(\theta_{n0} + \sqrt{p_n}(n^{-1/2})\mathbf{u}) - Q_n(\theta_{n0}) \\ &\leq l_n(\theta_{n0} + \alpha_n\mathbf{u}) - l_n(\theta_{n0}) \\ &\quad - n\lambda_n \sum_{j=1}^{q_n} \left(|\beta_{nj0} + \sqrt{p_n}(n^{-1/2})u_j|/|\tilde{\beta}_{nj}|^{\gamma} - |\beta_{nj0}|/|\tilde{\beta}_{nj}|^{\gamma}\right) \\ &\widehat{=} (I) + (II). \end{aligned}$$

By a Taylor's series expansion, we have

$$\begin{aligned} (I) &= \sqrt{p_n}(n^{-1/2})\nabla^T l_n(\theta_{n0})\mathbf{u} + \frac{1}{2}\mathbf{u}^T\nabla^2 l_n(\theta_{n0})\mathbf{u}p_n/n \\ &\quad + \frac{1}{6}\nabla^T\left\{\mathbf{u}^T\nabla^2 l_n(\theta_n^*)\mathbf{u}\right\}\mathbf{u}p_n^{3/2}n^{-3/2} \\ &\widehat{=} I_1 + I_2 + I_3, \end{aligned}$$

where the vector θ_n^* lies between θ_{n0} and $\theta_{n0} + \sqrt{p_n}(n^{-1/2})\mathbf{u}$.

We first consider the first term of (I). It follows from Lemma A.1 that

$$|I_1| \quad = \quad |\sqrt{p_n}(n^{-1/2})\nabla^T l_n(\theta_{n0})\mathbf{u}|$$

$$
\begin{aligned}
&\leq \sqrt{p_n}(n^{-1/2})||U_n(\theta_{n0})|| \cdot ||\mathbf{u}|| \\
&= \sqrt{p_n}(n^{-1/2})O_p(\sqrt{np_n})||\mathbf{u}|| \\
&= O_p(p_n)||u||. \qquad (6.14)
\end{aligned}
$$

Similar to Lemma A.1, by Chebyshev's inequality and condition (C2), for any given $\epsilon > 0$, we obtain

$$||\frac{1}{n}\nabla^2 l_n(\theta_{n0}) + I_n(\theta_{n0})|| = o_p(\frac{1}{p_n}).$$

Then, we have

$$
\begin{aligned}
I_2 &= \frac{p_n}{2n}\mathbf{u}^T\nabla^2 l_n(\theta_{n0})\mathbf{u} \\
&= \frac{1}{2}p_n\mathbf{u}^T\left[\frac{1}{n}\nabla^2 l_n(\theta_{n0}) + I_n(\theta_{n0})\right]\mathbf{u} - \frac{1}{2}p_n\mathbf{u}^T I_n(\theta_{n0})\mathbf{u} \\
&= \frac{1}{2}||\mathbf{u}||^2 o_p(1) - \frac{1}{2}p_n\mathbf{u}^T I_n(\theta_{n0})\mathbf{u}. \qquad (6.15)
\end{aligned}
$$

By condition (C3) and the Cauchy-schwarz inequality, we have

$$
\begin{aligned}
|I_3| &= |\frac{1}{6}\nabla^T\{\mathbf{u}^T\nabla^2 l_n(\theta_n^*)\mathbf{u}\}\mathbf{u}(\frac{p_n}{n})^{-3/2}| \\
&= \frac{1}{6}|\sum_{j,l,m=1}^{p_n}\frac{\partial^3 l_n(\theta_n^*)}{\partial\theta_{nj}\partial\theta_{nl}\partial\theta_{nm}}u_j u_l u_m(\frac{p_n}{n})^{-3/2}| \\
&\leq \frac{1}{6}\sum_{i=1}^{n}\left(\sum_{j,l,m=1}^{p_n}M_{njlm}^2(X_{ni})\right)^{1/2}||\mathbf{u}||^3(\frac{p_n}{n})^{-3/2} \\
&= n||\mathbf{u}||^2(\frac{p_n}{n})^{-3/2}O_p(p_n^{3/2}) \\
&= o_p(p_n)||\mathbf{u}||^2. \qquad (6.16)
\end{aligned}
$$

Next, we consider (II) by a Taylor's series expansion. We have

$$
\begin{aligned}
|(II)| &= n\lambda_n\sum_{j=1}^{q_n}\left(\frac{\left|\beta_{nj0} + \sqrt{p_n}(n^{-1/2})u_j\right|}{\left|\tilde{\beta}_{nj}\right|^\gamma} - \frac{|\beta_{nj0}|}{\left|\tilde{\beta}_{nj}\right|^\gamma}\right) \\
&\leq n\lambda_n \cdot n^{-1/2}\sqrt{p_n}\sum_{j=1}^{q_n}\frac{|u_j|}{\left|\tilde{\beta}_{nj}\right|^\gamma}
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{np_n}\lambda_n \sum_{j=1}^{q_n} |u_j| \left\{ \frac{1}{|\beta_{nj0}|^\gamma} - \frac{\gamma \text{sign}(\beta_{nj0})}{|\beta_{nj0}|^{\gamma+1}} (\tilde{\beta}_{nj} - \beta_{nj0}) + o_p\left(\left| \tilde{\beta}_{nj} - \beta_{nj0} \right| \right) \right\} \\
&= \sqrt{np_n}\lambda_n \sum_{j=1}^{q_n} |u_j| \left\{ \frac{1}{|\beta_{nj0}|^\gamma} + O_p\left(\sqrt{p_n/n} \right) \right\} \\
&= \sqrt{np_n}\lambda_n O_p(1)||\mathbf{u}|| \\
&= O_p(1)||\mathbf{u}||, \text{ since } \sqrt{np_n}\lambda_n = O_p(1). \qquad (6.17)
\end{aligned}
$$

By choosing a large enough constant C, (6.14), (6.16) and (6.17) are dominated by (6.15). Therefore, inequality (6.13) holds, and this completes the proof. □

Proof of Theorem 3

(i). To show part (i) of theorem 2, we first show that if $\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$ and $p_n^5/n \to 0$ as $n \to \infty$, then with probability 1, for any given $(\eta^T, \beta_{n1}^T)^T$ satisfying $||(\eta^T, \beta_{n1}^T)^T - (\eta_0^T, \beta_{n10}^T)^T|| = O_p(\sqrt{p_n/n})$ and any constant C,

$$
Q_n\{(\eta^T \beta_{n1}^T, 0)^T\} = \max_{||\beta_{n2}|| \le C(p_n/n)^{1/2}} Q_n\{(\eta^T, \beta_{n1}^T, \beta_{n2}^T)^T\}.
$$

Set $\epsilon_n = C\sqrt{p_n/n}$, then it is equivalent to show that with probability 1 as $n \to \infty$, for any $(\eta^T, \beta_{n1}^T)^T$ satisfying $||(\eta^T, \beta_{n1}^T)^T - (\eta_0^T, \beta_{n10}^T)^T|| = O_p\left(\sqrt{p_n/n}\right)$, for $j = q_n + 1, \ldots, p_n$, we have

$$
\begin{aligned}
\frac{\partial Q_n(\theta_n)}{\partial \beta_{nj}} &< 0, \text{ for } 0 < \beta_{nj} < \epsilon_n, \qquad (6.18) \\
\frac{\partial Q_n(\theta_n)}{\partial \beta_{nj}} &> 0, \text{ for } -\epsilon_n < \beta_{nj} < 0. \qquad (6.19)
\end{aligned}
$$

By a Taylor's series expansion,

$$
\begin{aligned}
\frac{\partial Q_n(\theta_n)}{\partial \beta_{nj}} &= \frac{\partial}{\partial \beta_{nj}} \{ l_n(\theta_n) - n\lambda_n \sum_{i=1}^{n} |\beta_{nj}| / |\tilde{\beta}_{nj}|^\gamma \} \\
&= \frac{\partial}{\partial \beta_{nj}} l_n(\theta_n) - n\lambda_n \frac{\text{sign}(\beta_{nj})}{|\tilde{\beta}_{nj}|^\gamma} \\
&= \frac{\partial}{\partial \beta_{nj}} l_n(\theta_{n0}) + \sum_{l=1}^{p_n} \frac{\partial^2 l_n(\theta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} (\beta_{nl} - \beta_{nl0})
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad \sum_{l,m=1}^{p_n} \frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nm}}(\beta_{nl}-\beta_{nl0})(\beta_{nm}-\beta_{nm0}) - n\lambda_n \frac{\text{sign}(\beta_{nj})}{|\tilde{\beta}_{nj}|^\gamma} \\
&\widehat{=} \quad I_1 + I_2 + I_3 + I_4, \text{ where } \theta_n^* \text{ lies between } \theta_n \text{ and } \theta_{n0}.
\end{aligned}
$$

According to Lemma A.1, we obtain that $I_1 = \frac{\partial l_n(\theta_{n0})}{\partial\beta_{nj}} = O_p(\sqrt{np_n})$.

We can write I_2 as follows:

$$
\begin{aligned}
I_2 &= \sum_{l=1}^{p_n} \frac{\partial^2 l_n(\theta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}}(\beta_{nl}-\beta_{nl0}) \\
&= \sum_{l=1}^{p_n} \left\{ \frac{\partial^2 l_n(\theta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}} - E\left(\frac{\partial^2 l_n(\theta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}} \right) \right\} (\beta_{nl}-\beta_{nl0}) + \sum_{l=1}^{p_n} E\left(\frac{\partial^2 l_n(\theta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}} \right) (\beta_{nl}-\beta_{nl0}) \\
&\widehat{=} I_{21} + I_{22}.
\end{aligned}
$$

We obtain the following argument by the Cauchy-Schwarz inequality,

$$
|I_{21}| \leq ||\theta_n - \theta_{n0}|| \left\{ \sum_{l=1}^{p_n} \left(\frac{\partial^2 l_n(\theta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}} - E(\frac{\partial^2 l_n(\beta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}}) \right)^2 \right\}^{1/2}.
$$

We have that $||\beta_n - \beta_{n0}|| = O_p(\sqrt{p_n/n})$, and by condition (C2),

$$
\left\{ \sum_{l=1}^{p_n} \left(\frac{\partial^2 l_n(\theta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}} - E(\frac{\partial^2 l_n(\beta_{n0})}{\partial\beta_{nj}\partial\beta_{nl}}) \right)^2 \right\}^{1/2} = O_p(\sqrt{np_n}).
$$

The term I_{22} becomes

$$
|I_{22}| = |n \sum_{l=1}^{p_n} I_n(\theta_{n0})(j,l)(\beta_{nl}-\beta_{nl0})|,
$$

where $I_n(\theta_{n0})(j,l)$ is the (j,l)th element of $I_n(\theta_{n0})$. By the Cauchy-Schwarz inequality and condition (C2), we obtain

$$
\begin{aligned}
|I_{22}| &\leq n||\theta_n - \theta_{n0}|| \left\{ \sum_{l=1}^{p_n} I_n^2(\theta_{n0})(j,l) \right\}^{1/2} \\
&= nO_p(\sqrt{p_n/n})O(1) \\
&= O_p(\sqrt{np_n}).
\end{aligned}
$$

Therefore, the term I_2 equals to $O_p(\sqrt{np_n})$.

Then, we consider the term I_3, which can be written as

$$\begin{aligned} I_3 &= \sum_{l,m=1}^{p_n} \left\{ \frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\beta_{nl}\beta_{nm}} - E\left(\frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nm}} \right) \right\} (\beta_{nl} - \beta_{nl0})(\beta_{nm} - \beta_{nm0}) \\ &+ \sum_{l,m=1}^{p_n} E\left(\frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nm}} \right) (\beta_{nl} - \beta_{nl0})(\beta_{nm} - \beta_{nm0}) \\ &\widehat{=} I_{31} + I_{32}. \end{aligned}$$

By the Cauchy-Schwarz inequality and condition (C3), we have

$$\begin{aligned} |I_{31}| &\leq ||\theta_n - \theta_{n0}||^2 \left\{ \sum_{l,m=1}^{p_n} \left(\frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nm}} - E(\frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nm}}) \right)^2 \right\}^{1/2} \\ &= O_p(\frac{p_n}{n} \cdot \sqrt{np_n^2}) \\ &= o_p(\sqrt{np_n}). \end{aligned}$$

The term I_{32} becomes

$$\begin{aligned} |I_{32}| &\leq ||\theta_n - \theta_{n0}||^2 \left\{ \sum_{l,m=1}^{p_n} E\left(\frac{\partial^3 l_n(\theta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nm}} \right)^2 \right\}^{1/2} \\ &= O_p(\frac{p_n}{n} \cdot np_n) \cdot C_5^{1/2} \\ &= o_p(\sqrt{np_n}). \end{aligned}$$

Combining I_{31} and I_{32}, we obtain that $I_3 = o_p(\sqrt{np_n})$.

Recall that $I_1 = O_p(np_n)$, $I_2 = O_p(\sqrt{np_n})$, it follows that $I_1 + I_2 + I_3 = O_p(\sqrt{np_n})$. Since $\sqrt{n/p_n}\lambda_n \to \infty$ and $\liminf_{n\to\infty} \inf_{\beta\to 0^+} \frac{1}{|\tilde{\beta}|^\gamma} > 0$, it is clear that the sign of $\frac{\partial Q_n(\theta_n)}{\partial\beta_{nj}}$ is determined by the sign of β_{nj} by explicitly writing out

$$\frac{\partial Q_n(\theta_n)}{\partial\beta_{nj}} = n\lambda_n \left\{ O_p\left(\frac{1}{\sqrt{n/p_n}\lambda_n} \right) - \frac{sign(\beta_{nj})}{|\tilde{\beta}_{nj}|^\gamma} \right\}.$$

Therefore, (6.18) and (6.19) follow. From Theorem 1, we know that $Q_n(\theta_n)$ has a root-(n/p_n)-consistent local maximizer $\hat{\theta}_n$. The above proof shows that part (i) holds for $\hat{\theta}_n = (\hat{\eta}^T, \hat{\beta}_{n1}^T, 0)^T$, which implies that

$\hat{\beta}_{n2} = 0$ with probability tending to 1.

(ii). Let θ_{n1} denote $(\eta^T, \beta_{n1}^T, 0^T)^T$ and θ_{n10} denote $(\eta^T, \beta_{n10}^T, 0^T)^T$. Since we have that $P(\hat{\beta}_{n2} = 0) \to 1$ obtained in part (i), we only need to consider the derivative of the asymptotic expansion of $\hat{\theta}_{n1}$ in the probability set $\{\hat{\beta}_{n2} = 0\}$. We use a Taylor's series expansion on $\nabla_{\theta_{n1}} Q_n(\hat{\theta}_n)$ at θ_{n10}, which yields

$$\begin{aligned}
\nabla_{\theta_{n1}} Q_n(\theta_n)|_{\theta_n=\theta_{n10}} &= \nabla_{\theta_{n1}} l_n(\theta_{n10}) + \nabla^2_{\theta_{n1}} l_n(\theta_{n10})(\hat{\theta}_{n1} - \theta_{n10}) \\
&+ \frac{1}{2}(\hat{\theta}_{n1} - \theta_{n10})^T \nabla^3_{\theta_{n1}} l_n(\theta^*_{n1})(\hat{\theta}_{n1} - \theta_{n10}) \\
&- \boldsymbol{b_n}, \text{ where } \theta^*_{n1} \text{ and } \theta^{**}_{n1} \text{ lies between } \theta_{n10} \text{ and } \hat{\theta}_{n1}. \qquad (6.20)
\end{aligned}$$

If we have that

$$I_n(\theta_{n10})(\hat{\theta}_{n1} - \theta_{n10}) + \boldsymbol{b_n} = \frac{1}{n}\nabla_{\theta_{n1}} l_n(\theta_{n10}) + o_p(n^{-\frac{1}{2}}), \qquad (6.21)$$

then we multiply $\sqrt{n} A_n I_n^{-\frac{1}{2}}(\theta_{n10})$ on the both side of above equation, and (6.21) becomes

$$\begin{aligned}
&\sqrt{n} A_n I_n^{\frac{1}{2}}(\theta_{n10})(\hat{\theta}_{n1} - \theta_{n10} + I_n(\theta_{n10})^{-1}\boldsymbol{b_n}) = \\
&\frac{1}{\sqrt{n}} A_n I_n^{-\frac{1}{2}}(\theta_{n10}) \nabla_{\theta_{n1}} l_n(\theta_{n10}) + o_p(A_n I_n^{-\frac{1}{2}}(\theta_{n10})).
\end{aligned}$$

The last term $o_p(A_n I_n^{-\frac{1}{2}}(\theta_{n10}))$ is $o_p(1)$, and is implied by the conditions of Theorem 2.
Let Z_{ni} denote $\frac{1}{\sqrt{n}} A_n I_n^{-\frac{1}{2}}(\theta_{n10}) \nabla_{\theta_{n1}} l_{ni}(\theta_{n10})$ for $i = 1, 2, \ldots, n$. For any ϵ and $i = 1, 2, \ldots, n$, by Holder's inequality, we have

$$\begin{aligned}
\sum_{i=1}^{n} E(||Z_{ni}||^2)\mathbf{1}(||Z_{ni}|| > \epsilon) &= nE(||Z_{n1}||^2)\mathbf{1}(||Z_{n1}|| > \epsilon) \\
&\leq nE(||Z_{n1}||^4)^{1/2}[P(||Z_{n1}| > \epsilon|)]^{1/2}.
\end{aligned}$$

By Markov's inequality, $A_n A_n^T \to G$ and condition (C2), it follows that

$$P(||Z_{n1}|| > \epsilon) \leq \frac{E(||A_n I_n^{-\frac{1}{2}}(\theta_{n10}) \nabla_{\theta_{n1}} l_{n1}(\theta_{n10})||^2)}{n\epsilon} = O(n^{-1})$$

and

$$\begin{aligned} E(||Z_{n1}||^4) &= \frac{1}{n^2} E(||A_n I_n^{-\frac{1}{2}}(\theta_{n10}) \nabla_{\theta_{n1}} l_{n1}(\theta_{n10})||^4) \\ &\leq \frac{1}{n^2} \lambda_{max}(A_n A_n^T) \lambda_{max}(I_n(\theta_{n10})) E(||\nabla_{\theta_{n1}}^T l_{n1}(\theta_{n10})||^2) \\ &= O((p_n/n)^2). \end{aligned}$$

Therefore, $\sum_{i=1}^n E(||Z_{ni}||^2)\mathbf{1}(||Z_{ni}|| > \epsilon) = O(n\frac{p_n}{n}\frac{1}{\sqrt{n}}) = o(1)$.

By $A_n A_n^T \to G$, the covariance of Z_{ni} for $i = 1, 2, \ldots, n$ can be written as

$$\sum_{i=1}^n \text{cov}(Z_{ni}) = n\text{cov}(Z_{n1}) = \text{cov}(A_n I_n^{-\frac{1}{2}}(\theta_{n10}) \nabla_{\theta_{n1}} l_{n1}(\theta_{n10})) \to G.$$

Then, by the Lindeberg-Feller central limit theorem, $\frac{1}{\sqrt{n}} A_n I_n^{-\frac{1}{2}}(\theta_{n10}) \nabla_{\theta_{n1}} l_n(\theta_{n10})$ asymptotically follows a multivariate normal distribution.

Next, since $\hat{\theta}_{n1}$ need to satisfy $\nabla_{\theta_{n1}} Q_n(\hat{\theta}_{n1}) = 0$, by letting the Taylor's series expansion in (6.20) equal to zero, we have

$$\begin{aligned} \frac{1}{n}\left(\{\nabla_{\theta_{n1}}^2 l_n(\theta_{n10})\}(\hat{\theta}_{n1} - \theta_{n10}) - \boldsymbol{b_n}\right) = -\frac{1}{n}\{&\nabla_{\theta_{n1}} l_n(\theta_{n10}) \\ &+ \frac{1}{2}(\hat{\theta}_{n1} - \theta_{n10})^T \nabla_{\theta_{n1}}^3 l_n(\theta_{n1}^*)(\hat{\theta}_{n1} - \theta_{n10})\}. \end{aligned}$$

Let L_n denote $\nabla_{\theta_{n1}}^2 l_n(\theta_{n10})$ and V_n denote $\frac{1}{2}(\hat{\theta}_{n1} - \theta_{n10})^T \nabla_{\theta_{n1}}^3 l_n(\theta_{n1}^*)(\hat{\theta}_{n1} - \theta_{n10})$. We have

$$\begin{aligned} ||\frac{1}{n} V_n||^2 &\leq \frac{1}{n^2} \sum_{i=1}^n n^2 ||\hat{\theta}_{n1} - \theta_{n10}||^4 \sum_{j,l,m=1}^{q_n} M_{njlm}^2(X_{ni}, Y_{ni}, \Delta_{ni}) \\ &= O_p((\frac{p_n}{n})^2 p_n^3) = o_p(\frac{1}{n}), \end{aligned} \tag{6.22}$$

which is implied by the Cauchy-Schwarz inequality and conditions (C3)-(C4). By Chebyshev's inequality and condition (C4), we obtain

$$\lambda_i(\frac{1}{n} L_n + I_n(\theta_{n10}) + \Sigma_{\lambda_n}) = o_p(1/\sqrt{n}),$$

for $i = 1, 2, \ldots, q_n$, where $\lambda_i(M)$ represents the ith eigenvalue of a symmetric matrix M. Since $||\hat{\theta}_{n1} - \theta_{n10}|| = O_p(\sqrt{p_n/n})$, we have

$$\left(\frac{1}{n}L_n + I_n(\theta_{n10})\right)(\hat{\theta}_{n1} - \theta_{n10}) = o_p(1/\sqrt{n}). \tag{6.23}$$

Combining (6.22) and (6.23), it follows that (6.21) holds, and this completes the proof of Theorem 3.

Additional simulation results

Additional results for simulation scenario 2 and 3

Tables A.1 and A.2 report the accuracy of nonzero coefficient post-selection estimates, their standard errors (SE), the mean of standard error estimator (SEE) and the coverage probability for nominal 95% confidence interval (CP) from the simulation study scenarios 2 and 3. To obtain the standard errors for the maximum likelihood estimates, we use the Louis formula (Louis, 1982) because the latent group membership is treated as missing data in our method. The post-selection estimates are biased on small samples and the bias can be greatly reduced by increasing the sample size. The 95% confidence intervals for the post-selection estimators based on the estimated coefficients and standard errors have accurate coverage for the true parameters.

Sensitivity analysis

In this sections, we show the results of simulation studies that evaluate the model performance when the link function in the multinomial distribution of the latent group membership assignment is nonlinear. The assumptions regarding the baseline covariates, the time to event data for each latent subgroup, the censoring time and the true values of regression coefficients are the same as scenario 1 in the simulation study. In this setting, we write the probability that one patient belongs to a specific latent group given baseline covariates as

$$P(B = k|X) = \frac{\exp\{h_k(X)\}}{\sum_{k=1}^{K}\exp\{h_k(X)\}} = \pi_k(X, \beta), \text{ for } k = 1, 2$$

where $h_1(X) = 0$ since subgroup 1 is set to be the reference group, and $h_2(X) = X_1^2 + X_2^2 - 1$.

Table A.3 reports the classification accuracy, along with standard errors, for models without and after variable selection when the model is misspecified. The classification accuracy is calculated by applying the decision rule obtained from the training set to a validation set with sample size $10,000$. For both scenarios that the model is misspecified, the classification accuracy is increasing as the sample size increases. The accuracy is also improved after performing variable selection. Compared to the optimal accuracy rate, our method performs reasonably well.

Table A.1: Maximum likelihood Estimates after variable selection, their standard errors, and coverage probabilities for nominal 95% confidence intervals from simulation scenario 2

N	Parameter	Bias	SE	SEE	CP
300	k_1	0.028	0.111	0.095	0.919
	λ_1	0.028	0.264	0.173	0.757
	k_2	0.061	0.390	0.272	0.867
	λ_2	-0.012	0.251	0.154	0.876
	β_0	0.003	0.702	0.346	0.809
	β_1	0.031	0.441	0.175	0.563
	β_2	-0.189	0.772	0.280	0.774
	β_3	-0.097	0.568	0.234	0.726
	β_4	0.127	0.777	0.253	0.716
	β_5	-0.166	0.561	0.271	0.868
	β_6	0.304	0.806	0.331	0.890
	β_7	-0.278	0.676	0.326	0.911
	β_8	0.212	0.637	0.282	0.902
1000	k_1	0.008	0.057	0.055	0.932
	λ_1	0.002	0.125	0.110	0.881
	k_2	0.014	0.154	0.151	0.948
	λ_2	-0.006	0.092	0.087	0.935
	β_0	0.017	0.173	0.159	0.910
	β_1	-0.016	0.132	0.082	0.754
	β_2	-0.021	0.123	0.113	0.943
	β_3	-0.002	0.114	0.103	0.936
	β_4	0.011	0.125	0.111	0.939
	β_5	-0.022	0.115	0.108	0.944
	β_6	0.034	0.120	0.116	0.943
	β_7	-0.027	0.128	0.116	0.928
	β_8	0.024	0.117	0.108	0.936
3000	k_1	0.003	0.032	0.032	0.947
	λ_1	0.003	0.071	0.067	0.925
	k_2	0.008	0.091	0.088	0.947
	λ_2	0.001	0.056	0.051	0.929
	β_0	-0.001	0.096	0.093	0.944
	β_1	-0.001	0.063	0.056	0.934
	β_2	-0.007	0.064	0.063	0.951
	β_3	-0.003	0.062	0.060	0.944
	β_4	0.003	0.063	0.063	0.941
	β_5	-0.003	0.061	0.060	0.947
	β_6	0.007	0.063	0.064	0.962
	β_7	-0.005	0.063	0.064	0.962
	β_8	0.005	0.063	0.060	0.939

Table A.2: Maximum likelihood Estimates after variable selection, their standard errors, and coverage probabilities for nominal 95% confidence intervals from simulation scenario 3

N	Parameter	Bias	SE	SEE	CP
300	k_1	0.037	0.118	0.098	0.907
	λ_1	-0.051	0.218	0.161	0.767
	k_2	-0.024	0.378	0.258	0.824
	λ_2	-0.063	0.258	0.152	0.836
	β_0	0.247	0.547	0.338	0.773
	β_1	0.041	0.373	0.187	0.568
	β_2	-0.139	0.541	0.256	0.709
	β_3	-0.101	0.446	0.225	0.691
	β_4	0.120	0.494	0.242	0.680
	β_5	-0.172	0.428	0.267	0.835
	β_6	0.262	0.517	0.304	0.856
	β_7	-0.303	0.546	0.317	0.871
	β_8	0.200	0.376	0.269	0.886
1000	k_1	0.008	0.056	0.054	0.926
	λ_1	0.008	0.133	0.109	0.873
	k_2	0.021	0.164	0.151	0.928
	λ_2	-0.000	0.095	0.087	0.919
	β_0	0.004	0.193	0.160	0.893
	β_1	-0.014	0.129	0.085	0.774
	β_2	-0.027	0.138	0.114	0.934
	β_3	-0.015	0.117	0.107	0.942
	β_4	0.024	0.128	0.114	0.938
	β_5	-0.039	0.122	0.110	0.934
	β_6	0.045	0.132	0.118	0.932
	β_7	-0.050	0.136	0.119	0.922
	β_8	0.031	0.119	0.110	0.947
3000	k_1	0.002	0.032	0.032	0.946
	λ_1	0.004	0.072	0.067	0.920
	k_2	0.005	0.091	0.088	0.943
	λ_2	-0.000	0.054	0.051	0.941
	β_0	-0.000	0.095	0.093	0.937
	β_1	-0.000	0.065	0.056	0.936
	β_2	-0.003	0.066	0.063	0.942
	β_3	-0.009	0.062	0.060	0.945
	β_4	0.008	0.063	0.063	0.953
	β_5	-0.002	0.062	0.060	0.938
	β_6	0.006	0.067	0.064	0.942
	β_7	-0.006	0.066	0.064	0.944
	β_8	0.005	0.061	0.060	0.944

Table A.3: Results from the sensitivity analysis

	Accuracy (SE)	
N	without variable selection	after variable selection
The optimal accuracy rate is 0.726.		
300	0.562 (0.046)	0.594 (0.056)
1000	0.596 (0.038)	0.609 (0.043)
3000	0.618 (0.013)	0.620 (0.004)

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 4

Proof of Theorem 1

We write the complete data log-likelihood as

$$\begin{aligned} l_c(\eta, h_2, \ldots, h_K; Y_c) &= \sum_{i=1}^{n}\sum_{k=1}^{K} I(B_i = k)\left[\Delta_i \log\left\{f(Y_i;\eta_k)g_k(X_i)\right\}\right. \\ &\quad \left.+(1-\Delta_i)\log\left\{S(Y_i;\eta_k)g_k(X_i)\right\}\right]. \end{aligned}$$

To show that the observed data log-likelihood increases in successive iterations, we calculate the difference in the observed data log-likelihood between two successive iterations as

$$\begin{aligned} & l(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{obs} \in \mathcal{B}_1^{(t+1)}) + l(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{obs} \in \mathcal{B}_2^{(t+1)}) \\ & -l(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{obs} \in \mathcal{B}^{(t)}) \\ = & \left\{E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_c)|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right]\right. \\ & +E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_c)|Y_{obs} \in \mathcal{B}_2^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\ & \left.-E\left[l_c(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_c)|Y_{obs} \in \mathcal{B}^{(t)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right]\right\} \\ & -\left\{E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{mis}|Y_{obs})|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right]\right. \\ & +E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{mis}|Y_{obs})|Y_{obs} \in \mathcal{B}_2^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\ & \left.-E\left[l_c(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{mis}|Y_{obs})|Y_{obs} \in \mathcal{B}^{(t)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right]\right\} \\ \equiv & (I) - (II). \end{aligned}$$

The quantity (I) is non-negative from the algorithm stated in Section 3 of the main paper since $\{\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}\}$ satisfy

$$\begin{aligned} & E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_c)|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\ & +E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_c)|Y_{obs} \in \mathcal{B}_2^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\ \geq & E\left[l_c(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_c)|Y_{obs} \in \mathcal{B}^{(t)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right]. \end{aligned}$$

For (II), we have

$$
\begin{aligned}
& E\left[l_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{mis}|Y_{obs})|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\
& \quad -E\left[l_c(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{mis}|Y_{obs})|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\
= \quad & E\left[\log\left\{\frac{L_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{mis}|Y_{obs})}{L_c(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{mis}|Y_{obs})}\right\}|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right] \\
\leq \quad & \log\left(E\left[\frac{L_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{mis}|Y_{obs})}{L_c(\eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}; Y_{mis}|Y_{obs})}|Y_{obs} \in \mathcal{B}_1^{(t+1)}, \eta^{(t)}, h_2^{(t)}, \ldots, h_K^{(t)}\right]\right) \\
= \quad & \log\int_{\mathcal{Y}_{mis}(Y_{obs})} L_c(\eta^{(t+1)}, h_2^{(t+1)}, \ldots, h_K^{(t+1)}; Y_{mis}|Y_{obs})dY_{mis} \\
= \quad & 0,
\end{aligned}
$$

where $\mathrm{L_c}(\cdot) = \exp\{\mathrm{l_c}(\cdot)\}$ is the complete data likelihood function.

The inequality above holds according to Jensen's inequality. A similar argument for dataset $\mathcal{B}_2^{(t+1)}$ can be easily obtained. Therefore, (II) is non-positive, which implies that the observed-data likelihood function increases over successive iterations.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alosh, M. and Huque, M. F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*, 28(1):3–23.

Altstein, L. and Li, G. (2013). Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model. *Biometrics*, 69(1):52–61.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, 75(1):65–74.

Bennis, A., Mouysset, S., and Serrurier, M. (2020). Estimation of conditional mixture weibull distribution with right censored data using neural network for time-to-event analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, volume 12084, pages 687–698. Springer.

Breaux, H. J. (1967). On stepwise multiple linear regression. Technical report, Army Ballistic Research Lab Aberdeen Proving Ground MD.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bussy, S., Guilloux, A., Gaïffas, S., and Jannot, A. S. (2019). C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical Methods in Medical Research*, 28(5):1523–1539.

Chen, C. and Beckman, R. A. (2009). Hypothesis testing in a confirmatory phase III trial with a possible subset effect. *Statistics in Biopharmaceutical Research*, 1(4):431–440.

Chi, Y. Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2):432–445.

Cho, H. J. and Hong, S. M. (2008). Median regression tree for analysis of censored survival data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(3):715–726.

Ciampi, A., Bush, R., Gospodarowicz, M., and Till, J. (1981). An approach to classifying prognostic factors related to survival experience for non-hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975. *Cancer*, 47(3):621–627.

Ciampi, A., Chang, C., Hogg, S., and McKinney, S. (1987). Recursive partition: A versatile method for exploratory-data analysis in biostatistics. In *Biostatistics*, pages 23–50. Springer.

Ciampi, A., Hogg, S. A., McKinney, S., and Thiffault, J. (1988). Recpam: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. methods and program features. *Computer Methods and Programs in Biomedicine*, 26(3):239–256.

Ciampi, A., Negassa, A., and Lou, Z. (1995). Tree-structured prediction for censored survival data and the cox model. *Journal of Clinical Epidemiology*, 48(5):675–689.

Ciampi, A., Thiffault, J., Nakache, J.-P., and Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, 4(3):185–204.

Colleoni, M., Litman, H., Castiglione-Gertsch, M., Sauerbrei, W., Gelber, R., Bonetti, M., Coates, A., Schumacher, M., Bastert, G., Rudenstam, C., et al. (2002). Duration of adjuvant chemotherapy for breast cancer: a joint analysis of two randomised trials investigating three versus six courses of cmf. *British Journal of Cancer*, 86(11):1705–1714.

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

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265.

Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, 8(8):947–961.

Dispenzieri, A., Katzmann, J. A., Kyle, R. A., Larson, D. R., Therneau, T. M., Colby, C. L., Clark, R. J., Mead, G. P., Kumar, S., Melton III, L. J., et al. (2012). Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pages 517–523.

Doove, L. L., Dusseldorp, E., Van Deun, K., and Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Advances in Data Analysis and Classification*, 8(4):403–425.

Drton, M. and Plummer, M. (2017). A bayesian information criterion for singular models. *Journal of the Royal Statistical Society*, 79(2):323–380.

Dusseldorp, E., Conversano, C., and Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: Stima. *Journal of Computational and Graphical Statistics*, 19(3):514–530.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine*, 33(2):219–237.

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. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

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

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038.

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

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports*, 69(10):1065–1069.

Hendry, D. F. and Richard, J. F. (1987). Recent developments in the theory of encompassing. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769.

Ibrahim, N. A. and Kudus, A. (2009). Decision tree for prognostic classification of multivariate survival data and competing risks. In *Recent Advances in Technologies*, pages 1–34. INTECH Open Access Publisher.

International Breast Cancer Study Group (1996). Duration and reintroduction of adjuvant chemotherapy for node-positive premenopausal breast cancer patients. *Journal of Clinical Oncology*, 14(6):1885–1894.

Kannel, W. and McGee, D. (1979). Diabetes and glucose tolerance as risk factors for cardiovascular disease: the framingham study. *Diabetes Care*, 2(2):120–126.

Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J., and Castelli, W. P. (1979). An investigation of coronary heart disease in families: the framingham offspring study. *American Journal of Epidemiology*, 110(3):281–290.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3):227–237.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038.

Knight, K., Fu, W., et al. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Kovalchik, S. A., Varadhan, R., and Weiss, C. O. (2013). Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Statistics in Medicine*, 32(28):4906–4923.

Krisam, J. and Kieser, M. (2014). Decision rules for subgroup selection based on a predictive biomarker. *Journal of Biopharmaceutical Statistics*, 24(1):188–202.

Lagakos, S. (1981). The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika*, 68(1):93–98.

Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(3):201–211.

Law, M. H., Figueiredo, M. A., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166.

Law, M. H., Jain, A. K., and Figueiredo, M. (2003). Feature selection in mixture-based clustering. In *Advances in Neural Information Processing Systems*, pages 641–648.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 48(2):411–425.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467.

Liao, J. J., Farooqui, M. Z., Marinello, P., Hartzel, J., Anderson, K., Ma, J., and Gause, C. K. (2020). Using artificial intelligence tools in answering important clinical questions: The KEYNOTE-183 multiple myeloma experience. *Contemporary Clinical Trials*, 99:106179.

Liao, J. J. and Liu, G. F. (2019). A flexible parametric survival model for fitting time to event data in clinical trials. *Pharmaceutical Statistics*, 18(5):555–567.

Lipkovich, I. and Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of Biopharmaceutical Statistics*, 24(1):130–153.

Lipkovich, I., Dmitrienko, A., and B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21):2601–2621.

Lipsitz, S. R. and Ibrahim, J. G. (1996). Using the em-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2(1):5–14.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the cox model. *Biometrics*, 54(3):1002–1013.

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.

Loh, W.-Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34(11):1818–1833.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15(4):661–675.

Marubini, E., Morabito, A., and Valsecchi, M. (1983). Prognostic factors and risk groups: some results given by using an algorithm suitable for censored survival data. *Statistics in Medicine*, 2(2):295–303.

McGiffin, D. C., Galbraith, A. J., McLachlan, G. J., Stower, R. E., Wong, M. L., Stafford, E. G., Gardner, M. A., Pohlner, P. G., and O'Brien, M. F. (1992). Aortic valve infection: risk factors for death and recurrent endocarditis after aortic valve replacement. *The Journal of Thoracic and Cardiovascular Surgery*, 104(2):511–520.

Mehrotra, D. V. and Marceau West, R. (2020). Survival analysis using a 5-step stratified testing and amalgamation routine (5-STAR) in randomized clinical trials. *Statistics in Medicine*, 39(30):4724–4744.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Michiels, S., Potthoff, R. F., and George, S. L. (2011). Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. *Statistics in Medicine*, 30(13):1502–1518.

Millen, B. A., Dmitrienko, A., Ruberg, S., and Shen, L. (2012). A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Drug Information Journal: DIJ/Drug Information Association*, 46(6):647–656.

Molinaro, A. M., Dudoit, S., and Van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177.

Moradian, H., Larocque, D., and Bellavance, F. (2017). L1 splitting rules in survival forests. *Lifetime Data Analysis*, 23(4):671–691.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J.-F. (2005). Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15(3):231–239.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4):535–569.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

Qiu, X. and Wang, Y. (2019). Composite interaction tree for simultaneous learning of optimal individualized treatment rules and subgroups. *Statistics in Medicine*, 38(14):2632–2651.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242.

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, 110(509):303–312.

Shi, P. and Tsai, C.-L. (2002). Regression model selection - a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):237–252.

Song, Y. and Chi, G. Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine*, 26(19):3535–3549.

Steingrimsson, J. A., Diao, L., Molinaro, A. M., and Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in Medicine*, 35(20):3595–3612.

Steingrimsson, J. A., Diao, L., and Strawderman, R. L. (2019). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525):370–383.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2):141–158.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4(1).

Sun, Y., Chiou, S. H., and Wang, M. C. (2019). ROC-guided survival trees and ensembles. *Biometrics*, 76(4):1177–1189.

Tanniou, J., Van Der Tweel, I., Teerenstra, S., and Roes, K. C. (2016). Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Medical Research Methodology*, 16(1):1–15.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

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

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

Tseng, Y. J., Wang, H. Y., Lin, T. W., Lu, J. J., Hsieh, C. H., and Liao, C. T. (2020). Development of a machine learning model for survival risk stratification of patients with advanced oral cancer. *JAMA Network Open*, 3(8):e2011768.

Vergara, P., Tzou, W. S., Tung, R., Brombin, C., Nonis, A., Vaseghi, M., Frankel, D. S., Di Biase, L., Tedrow, U., Mathuria, N., Nakahara, S., Tholakanahalli, V., et al. (2018). Predictive score for identifying survival and recurrence risk profiles in patients undergoing ventricular tachycardia ablation: the I-VT score. *Circulation: Arrhythmia and Electrophysiology*, 11(12):e006730.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Zeng, D., Mao, L., and Lin, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103(2):253–271.

Zhou, Y. and McArdle, J. J. (2015). Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80(3):811–833.

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751.