HIGH-DIMENSIONAL VARYING-COEFFICIENT MODELS FOR GENOMIC STUDIES

NG HOI MIN

MPhil

The Hong Kong Polytechnic University

2021

The Hong Kong Polytechnic University

Department of Applied Mathematics

High-dimensional varying-coefficient models for genomic studies

NG Hoi Min

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Philosophy

May 2021

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

NG Hoi Min

Abstract

This thesis is concerned with novel statistical approaches for integrative genomic analysis, particularly methods that incorporate interactions between features from different data types into regression models. Recent technological advancements have enabled the collection of different molecular data types on the same patient. Extensive research efforts have been made towards integrating such comprehensive data set to study the biological mechanisms involved in disease development. While existing integrative approaches mainly focus on accounting for the difference in prognostic power across data types, there are relatively few works engaged in incorporating the interaction effects between different types of biological features on disease prognosis. As many chronic diseases are known to be affected by certain molecular features interacting with clinical factors, it is crucial to identify the relevant risk factors along with their interaction effects. In genomic studies, the major challenges of interaction analysis lie in the high dimensionality of the data and heterogeneity across data types. In order to decipher the association between the molecular interplays and a disease outcome, we propose to use the varying-coefficient models to characterize the interaction effects between the genomic features and a set of effect modifiers. We adopt a class of single-index varying-coefficient models to accommodate the potential interaction effects, and we propose a penalized spline-based estimation method for selecting important features with constant or varying effects. In an ongoing study, we consider a varying-coefficient additive hazards model and propose a kernel-based method to estimate the constant and varying effects.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Wong Kin Yau, for his patient guidance and enthusiastic encouragement. His keen outlook on research has been a great source of inspiration and motivated me to pursue postgraduate studies. Also, his professional expertise in applied statistics has sharpened my thinking and brought my work to a higher level. This work could not have been accomplished without his generous support.

I would also like to acknowledge my co-supervisor, Prof. Zhao Xingqiu, for offering administrative support throughout my postgraduate studies. A special thanks to Drs. Jiang Binyan and Liu Chunling, for sharing their innovative ideas in modern statistical researches.

Lastly, I wish to thank all the staff in the Department of Applied Mathematics for providing me with a great deal of support and assistance since my undergraduate years at the Hong Kong Polytechnic University.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Biostatistics is an important branch of statistical science with its primary focus on the applications of statistical approaches to a wide range of biological sciences areas. Among many interesting applications, genomics research is an area where statistical approaches have become popular tools in revealing the genetic basis underlying complex diseases such as cancer. The major goals of cancer genomics include identifying risk factors associated with the progression of cancer and predicting disease outcomes, such as time to tumor progression or death since initial diagnosis or treatment. In conventional cancer studies, clinical factors such as age, gender, and tumor stage are routinely studied and used as prognostic factors. While tumor progression is a dynamic biological process that involves gene mutation and alteration, clinical factors can only explain a proportion of the observed variation in tumor progression. To develop a deeper understanding of the underlying disease mechanisms in cancer and improve outcomes prediction, there has been a growing interest in studying the association between genomic features with cancer development.

Over the past few decades, the field of cancer genomics has been advancing due to

the developments in high-throughput technologies, such as microarrays, single nucleotide polymorphism arrays, proteomics, and RNA sequencing. Such progress in genomic profiling has facilitated the generation of large-scale omics data, which refer to the molecular data measured on the same individuals. For example, in The Cancer Genome Atlas (TCGA), clinical and omics data, including copy number alteration, DNA methylation, mutation, and the expressions of mRNA, microRNA, and protein, were collected from more than 11000 cancer patients across 33 tumor types. Also, in the Molecular Taxonomy of Breast Cancer International Consortium (Curtis et al., 2012), clinical data along with copy number alteration, mutation, and mRNA expression data were collected from about 2000 breast cancer patients. The emergence of these massive data has featured the development of statistical approaches for genomic analysis and provided many key insights into the biological mechanisms underlying cancer progression.

1.1.1 Conventional regression models

Among various statistical approaches, regression analysis is perhaps the most widely used technique with its primary focus to infer the relationships between the predictors and the outcomes of interest. In genomic analyses, regression models are frequently used for identifying relevant biological features that influence disease outcomes and predicting disease outcomes in new observations. Regression models, including the generalized linear models and the survival models, are commonly used to predict certain types of outcome variables, such as continuous, categorical, and survival outcomes.

The linear model is by far the most fundamental form of regression analysis, which investigates the linear relationship between a set of predictors and a continuous outcome variable. Let Y denote the outcome variable and $\boldsymbol{X}$ denote a vector of predictors. The linear model takes the following form:

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients indicating the (constant) effects of $\boldsymbol{X}$ on Y and ε is a random error with mean zero. This model assumes that the mean of Y is a linear combination of $\boldsymbol{X}$. As an extension to the linear model, a class of generalized linear models, introduced by Nelder and Wedderburn (1972), allows the mean of Y to depend on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$ through a non-linear link function, and the probability distribution of Y can be any member of the exponential family. The generalized linear model is particularly useful in real applications due to its flexibility to map the predictors with many types of outcome variables. For example, the logistic regression model is a standard tool for fitting binary outcomes, such as disease diagnosis and classification of risk levels. Also, the Poisson regression is often used for modeling count data, such as the number of times an event occurs.

Apart from the continuous and categorical outcomes, survival outcomes such as time to death or time to relapse are usually available for disease subjects in genomic studies. Survival analysis, or more generally, time-to-event analysis studies the time lapsed to a particular event. In genomic studies, survival analysis is commonly used to assess the influence of risk factors on survival and build a prognostic model for survival prediction in new observations. A common feature specific to survival analysis is right-censoring, which refers to the phenomenon that only some subjects will have experienced the event by the end of the study. Right-censoring usually occurs when the subject is still alive at the end of the study or the subject is lost to follow-up during the period of study. In this case, we cannot follow up on the exact event time for some subjects. Statistical modeling for the right-censored survival data has been extensively studied to assess the association between risk factors and survival outcomes (Klein and Moeschberger, 2003; Kalbfleisch and Prentice, 2011).

Semi-parametric models have been effectively used when the association of survival time with various risk factors is the main interest. We consider two popular semi-parametric models, namely the Cox proportional hazards model (Cox, 1972) and the

additive hazards model (Aalen, 1989; Lin and Ying, 1994). The Cox model is probably the most widely used model describing the association between risk factors and survival time. The hazard function that measures the risk of failure takes the following form:

$$\lambda(t \mid \boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}),$$

where λ_0 is an unspecified baseline hazard function. This model relies on a key assumption that each predictor has a multiplicative effect on the hazard function. As an alternative to the proportional hazards model, the additive hazards model assumes that each predictor has an additive effect on the hazard function. The additive hazards model was first introduced by Aalen (1989), which concerns time-dependent covariates and coefficients. Lin and Ying (1994) studied the model with time-independent coefficients and the hazard function has the following form:

$$\lambda(t \mid \boldsymbol{X}) = \lambda_0(t) + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}.$$

In this thesis, we focus on the semi-parametric survival models with time-independent coefficients as we are interested in the varying coefficients that depend on a set of predictors other than time. We will provide further discussion about the choice of survival model in Chapter 3.

In most problems, least-squares or maximum likelihood estimation can be applied for parameter estimation. Under a linear model, the distribution of Y is assumed to be normal and the score function is linear in $\boldsymbol{\beta}$, so we can readily obtained the estimator $\widehat{\boldsymbol{\beta}}$ by minimizing the sum of squared errors. Under a generalized linear model, the distribution of Y can be any member of the exponential family, and the corresponding score function is usually non-linear in $\boldsymbol{\beta}$. In this case, the maximizer of the loss function cannot be solved explicitly. Accordingly, iterative computational techniques such as the Newton-Raphson algorithm can be adopted to solve for $\boldsymbol{\beta}$.

Since the semi-parametric survival models involve an infinite-dimensional nuisance parameter λ_0, maximum likelihood estimation cannot be used directly to estimate $\boldsymbol{\beta}$. Cox (1972) developed an estimation method for $\boldsymbol{\beta}$ based on the notion of partial-likelihood for the proportional hazards model. Suppose there are n observations that are potentially right-censored. For the i-th subject, let C_i be the censoring time, $\widetilde{Y}_i = \min(Y_i, C_i)$ be the observed time, and $\Delta_i = I(Y_i \leq C_i)$ be the event indicator for $i = 1, \ldots, n$. The partial-likelihood function is defined as:

$$\prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta})}{\sum_{h:\widetilde{Y}_h \geq \widetilde{Y}_i} \exp(\boldsymbol{X}_h^{\mathrm{T}} \boldsymbol{\beta})} \right\}^{\Delta_i}.$$

A distinguishing feature of the partial-likelihood is that it depends only on the order rather than the exact times in which the events occur. Since the partial-likelihood does not involve λ_0, $\boldsymbol{\beta}$ can be estimated by maximizing the partial-likelihood. Cox (1975) described the asymptotic inference theory for the partial-likelihood estimator and suggested that the resulting estimator is consistent and asymptotically normal. Tsiatis et al. (1981) and Andersen and Gill (1982) showed that the partial-likelihood estimator possesses asymptotic properties similar to those of the standard maximum likelihood estimator. Lin and Ying (1994) considered a similar technique to estimate the time-independent coefficients for the additive hazards model.

1.1.2 Penalized regression methods

In genomic studies, we often encounter data set whose number of features is larger than the sample size. Such data is referred to as high-dimensional data and is becoming increasingly available as data collection technologies evolve. For example, over 20000 gene expressions were measured for each TCGA patient but the effective sample size for a tumor type is usually only in the order of hundreds. The so-called curse of dimensionality makes the maximum likelihood inference infeasible. For analyzing high-dimensional data,

various penalized regression methods have been investigated to avoid possible degeneracy and infeasibility of the maximum likelihood estimation.

The past two decades have been a remarkable era for the study of penalized regression methods. Penalized regression, also known as regularization or shrinkage, generally refers to the estimation of regression coefficients under some constraint or penalty. Penalty functions that serve different purposes in penalized regression have been studied and frequently applied to high-dimensional data analysis. At an early stage, penalized regression methods were designed for improving prediction accuracy by balancing the bias-variance trade-off. Hoerl and Kennard (1970a,b) introduced the ridge regression in a linear model, which minimizes the residual sum of squares (RSS) subject to a multiple of $\|\boldsymbol{\beta}\|_2$. Frank and Friedman (1993) discussed a generalization of ridge regression and subset selection through a penalty function of the form $\|\boldsymbol{\beta}\|_q$ for $q > 0$. This so-called bridge regression includes the ridge regression with $q = 2$ as a special case.

To achieve better model interpretability, penalty functions that induce variable selection have been studied, examples include the convex penalties (Tibshirani, 1996; Knight and Fu, 2000) and the non-convex penalties (Fan and Li, 2001; Zhang, 2007). These penalization techniques rely on a sparsity assumption that only a few variables have important implications on the outcome. The sparse penalization techniques attempt to shrink the regression coefficients for the less contributive variables to exactly zero, and thus the corresponding features are not included in the selected model. Furthermore, these sparse penalization techniques can help reduce overfitting if irrelevant features exist.

The least absolute shrinkage and selection operator (lasso), introduced by Tibshirani (1996), is probably the most popular penalized regression method for analyzing high-dimensional data due to its computational efficiency and ability to explore sparsity. Tibshirani (1996) considered a linear regression model and defined the lasso estimator as a minimizer of an objective function that consists of the RSS and an L_1-penalty as

follows:

$$\arg\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1,$$

where $\lambda > 0$ is a tuning parameter controlling the degree of shrinkage applied to the regression coefficients in $\boldsymbol{\beta}$. The L_1-penalization technique has been implemented to a broad range of models, such as other generalized linear models (Tibshirani, 1996) and survival models (Tibshirani, 1997; Ma and Huang, 2005), by replacing the RSS with a (negative) log-likelihood in the objective function. While imposing the L_1-penalty results in sparse solutions, the estimation bias can be large because the same degree of shrinkage is applied to all coefficients. To deal with this inadequacy, various modifications have been proposed, such as the adaptive lasso (Zou, 2006) and non-convex penalties including the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001) and the minimax concave penalty (MCP, Zhang, 2007). These penalties were shown to possess the oracle property, which means the estimator performs as well as if the true underlying model was given in advance.

Another class of lasso extensions for data with a group structure has been studied. Yuan and Lin (2006) proposed the group lasso to select a subset among the predefined groups of variables. The group lasso estimator is defined as the solution to the following problem:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} (\boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{K}_j \boldsymbol{\beta}_j)^{1/2},$$

where $\boldsymbol{\beta}_j$ denotes the coefficients corresponding to the j-th group and $\boldsymbol{K}_j$ is some positive definite matrix for $j = 1, \ldots, J$. By imposing a group-wise L_2-penalty, the variables are selected in groups, and all the coefficients in the selected groups are non-zero. If each group is of size 1, then the group lasso penalty reduces to the lasso penalty. Breheny and Huang (2009) extended the SCAD and MCP to the group selection problems. In genomic analyses, the group penalty is convenient when building a prognostic model based on functionally related gene sets or pathways.

1.1.3 Varying-coefficient models

We are often interested in exploring the non-linear relationships between the predictors and the outcomes of interest. The varying-coefficient model appears to be a useful alternative to the classical linear models when one wishes to examine the change of regression coefficient over different groups characterized by the effect modifiers. Hastie and Tibshirani (1993) introduced a generic form of the varying-coefficient model. Suppose that the distribution of Y depends on η in the following form:

$$\eta = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{U}),$$

where $\boldsymbol{g}$ is a vector of varying coefficients that are modified by an effect modifier $\boldsymbol{U}$. In contrast to the classical linear model with constant regression coefficients, the varying-coefficient model assumes that each regression coefficient varies as a function of another predictor $\boldsymbol{U}$, which permits an interaction between each component of $\boldsymbol{X}$ with $\boldsymbol{U}$. The idea of the varying coefficient is often used in conjunction with survival models. For example, Martinussen et al. (2002), Tian et al. (2005), and Chen et al. (2012) studied the proportional hazards models where the coefficient is a function of time or an exposure variable. Also, additive hazards models with varying coefficients have been studied by McKeague and Sasieni (1994) and Yin et al. (2008).

We review two popular classes of non-parametric estimation techniques for estimating the varying coefficients, namely spline and kernel. The spline-based estimation method is based on piecewise polynomial fitting, which approximates the underlying non-parametric function using piecewise polynomials. A spline curve can be expressed conveniently as a linear combination of the basis functions. Upon specifying the degree of the basis functions and the location of knots, $\boldsymbol{g}(\cdot)$ can be estimated by maximizing the likelihood over all possible linear combinations of the basis functions. On the other hand, the kernel-based estimation method is based on local linear fitting, which estimates $\boldsymbol{g}(\boldsymbol{u})$

using data in the neighborhood of $\boldsymbol{u}$. The basic idea underlying kernel methods is to use a likelihood function that includes not only subjects taken at the target point $\boldsymbol{u}$ but includes (usually with decreasing weights) subjects taken in the neighborhood of $\boldsymbol{u}$.

These non-parametric estimation techniques have been extensively studied for univariate effect modifiers. When one wishes to incorporate multivariate effect modifiers into the regression model, the single-index varying-coefficient model appears to be a convenient choice. The single-index varying-coefficient model is a combination of the varying-coefficient model and the single-index model (Hardle et al., 1993). It assumes that the distribution of Y depends on η in the following form:

$$\eta = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is a vector of index parameters. In this model, the effect of each component in $\boldsymbol{X}$ can vary flexibly with a linear combination of $\boldsymbol{U}$. As the multivariate effect modifiers are collapsed into an index, conventional non-parametric estimation techniques can be applied to estimate the varying coefficients. In the literature, Fan et al. (2003) explored a class of single-index varying-coefficient linear models by considering $\boldsymbol{U} = \boldsymbol{X}$. Xue and Pang (2013) and Zhao et al. (2019) considered the single-index varying-coefficient models for continuous outcomes and proposed kernel-based methods to estimate the index parameters and varying coefficients. Lin et al. (2016) studied a proportional hazards model with single-index varying-coefficient components for censored outcomes.

1.2 Integrative approaches for genomic analysis

In practical applications, clinical and genomic data are commonly used to construct regression models for disease prognosis. Numerous genomic analyses have been conducted to identify the risk factors involved in disease development and make predictions through analyzing the clinical or genomic data separately. As the underlying disease mechanisms

are influenced by a variety of factors, the integration of different data types appears to be more comprehensive. Many studies have shown that the integrative analysis of clinical and genomic data confers greater prognostic power than the analysis of clinical data alone (Li, 2006; Shedden et al., 2008; Bøvelstad et al., 2009; Fan et al., 2011; Zhao et al., 2015). To facilitate information borrowing among multiple data types, it is crucial to study statistical approaches for integrative genomic analysis.

Over the past two decades, integrative analyses for multiple data types have been extensively conducted to improve predictions over the analyses for a single data type. Existing integrative approaches can be predominantly classified into two classes, namely parallel integration and hierarchical integration. In parallel integration, the direct effects of features from each data type on the outcome are assessed. Whereas in hierarchical integration, the indirect regulatory or interaction effects between different data types on the outcome are of interest.

1.2.1 Parallel integration

A straightforward integration strategy is to combine different types of data into a single data set, on which conventional analyses are performed. Bøvelstad et al. (2009) applied the principal component analysis to project the high-dimensional genomic features into a low-dimensional space and then combined the projected data with clinical factors to predict survival outcomes. Fan et al. (2011) combined clinical factors and high-dimensional gene expression modules of breast cancer patients into a data set and built prognostic models using lasso regression. In a more comprehensive study, Zhao et al. (2015) evaluated the predictive powers of clinical and omics data along with their combinations across different cancer types by building prognostic models using principal component analysis, partial least squares, and lasso regression. Although these studies have demonstrated that direct combinations of clinical factors and genomic features improves risk prediction in some of the cancer types, this naive method probably neglects heterogeneity across data types as

each data type appears to have a distinct level of association with the outcome.

Alternatively, one may take into account the difference in prognostic power across different data types through some weighting approach. There are many applications that attempt to integrate heterogeneous data types using the weighted kernel-based methods. The kernel-based integration approach primarily consists of two steps. In the first step, a kernel matrix is constructed for each data type by mapping the data to a feature space using a non-linear feature map. Next, different kernel matrices are combined using some learning methods such as support vector machine. Daemen et al. (2007) proposed a weighted kernel-based method to integrate clinical and microarray data for classification. They demonstrated that models that account for the distinction between clinical and genomic data yield better prediction accuracy over models that treat these data types equally. To fully utilize all available information, weighted kernel-based methods for more than two data types have also been investigated (Lanckriet et al., 2004; Daemen et al., 2009; Seoane et al., 2014). Despite that the kernel-based integration approach improves predictive power, it transforms the data into another structure and makes interpretation difficult.

Methods that incorporate regularized constraints on multiple data types are more appealing for interpreting the effects of individual features on the outcome. Boulesteix et al. (2017) extended the idea of the lasso to an integrative lasso with penalty factors, which assigns distinct penalties to different data types for variable selection. Wong et al. (2019) proposed an integrative boosting approach that leverages the omics data sets with clinical information to predict survival outcomes by learning weights for heterogeneous data types using an iterative procedure. In each iteration, a prediction rule is constructed for different data types, and distinct penalties are assigned to the coefficients for features from different data types. The resulting models obtained using penalized estimation methods are relatively easier to understand than those obtained using the kernel-based methods.

1.2.2 Hierarchical integration

Though accounting for the prognostic difference across multiple data types, the parallel integrative approaches do not consider the interrelationship between features from different data types. To facilitate the understanding of the genetic basis of complex diseases, one may consider a class of hierarchical integrative approaches that accommodates the biological relationship between different types of data when building the prognostic models. Wang et al. (2013) and Zhu et al. (2016) incorporated prior knowledge of the regulatory relationship between different types of genomic data to the regression models. Wang et al. (2013) proposed an integrative Bayesian analysis of genomics data to model the regulatory effects by DNA methylation that are mediated through gene expression for predicting prognostic outcomes. They first considered a mechanistic model to infer the direct effects of DNA methylation on gene expression. With this information, a clinical model was constructed to assess the association between gene expression and survival outcome. Zhu et al. (2016) proposed a different approach to describe the regulation across different types of genomic data through constructing linear regulatory modules, and their proposed approach can accommodate the direct effects of regulators on the prognostic outcomes.

Apart from the regulatory relationship, we are often interested in studying the interaction effects between features from different data types. Many chronic diseases are complex and are caused by genes interacting with clinical factors. In cancer genomics, the effects of genomic features on cancer progression are often modified by clinical factors. For example, Landi et al. (2008) demonstrated that the effects of some gene expressions on the risk of lung cancer mortality vary with tobacco consumption. Also, Chen et al. (2017) and Relli et al. (2018) showed that the molecular mechanisms of carcinogenesis exhibit a high level of heterogeneity between two subtypes of non-small-cell lung carcinoma (NSCLC), and the same set of features could have distinct effects on disease outcome across different subtypes. As the effects of genomic features can vary across different

clinical characteristics, the integrative approaches that account for the differences and regulations across data types are insufficient to uncover the intricate interactions. This emerges the need for integrative approaches that analyze the clinical and genomic data compositely in the form of interaction.

To incorporate interaction effects between clinical and genomic variables in regression analyses, one may take a conventional approach that includes pairwise product terms of the variables into the regression model. Consider a simple linear model with pairwise interaction terms,

$$Y = a + \sum_j b_j X_j + \sum_k c_k U_k + \sum_{j,k} d_{jk} X_j U_k + \varepsilon,$$

where a is an intercept term, b_j, c_k, and d_{jk} are the regression coefficients for the j-th component of $\boldsymbol{X}$, k-th component of $\boldsymbol{U}$, and their product, respectively. Despite the simplicity of the form of interaction, this model may not be ideal for analyzing the interactions between clinical and genomic data. First, inclusion of product terms may greatly expand the model complexity and aggravate the high dimensionality issue. Even though the separate data set consists of a moderate number of variables, the size can grow substantially as all possible product terms between clinical and genomic variables are considered. Second, the scales of (quantitative) clinical and genomic variables are generally incomparable, and modeling interaction effects using pairwise product terms may not be appropriate.

Several integrative approaches that allow for interactions between different features have been proposed. Nevins et al. (2003) and Pittman et al. (2004) developed tree-based classification methods to evaluate the effects of clinical and genomic data on (binary) disease outcomes, allowing for potential interactions among multiple risk factors. However, the estimated model does not have simple interpretations, and the methods may not accommodate a large number of variables. In a recent study, Li et al. (2020) proposed a

penalized framework to select important gene-gene interaction effects on disease outcomes. However, the interactions between features from different data types were not considered.

Towards this end, the varying-coefficient model emerges as a promising approach for characterizing flexible interaction effects between different data types. As mentioned in the preceding section of this chapter, the varying-coefficient model allows the coefficients to vary over the groups stratified by certain effect modifiers, which permits interactions between features from different data types. In particular, the single-index varying-coefficient model has received much attention due to its ability to accommodate multivariate effect modifiers. In genomic analyses, the single-index varying-coefficient model can be used to describe the modifications of the effects of genomic features by clinical factors and introduce interaction effects between each genomic feature and a set of clinical factors. There are two major advantages of using the single-index varying-coefficient model to examine the potential interaction effects between clinical and genomic features. First, this model avoids the curse of dimensionality by projecting the clinical factors to an index so that the number of parameters only increases linearly with the number of genomic features. Second, this model accommodates the difference in scales between clinical and genomic data since the effects of genomic features are formulated as non-parametric functions of the index.

Some estimation methods have been developed to estimate the varying coefficients and index parameters of the single-index varying-coefficient models (see Section 1.1.3). While these methods show relevance, they possess several limitations. First, most of the existing works assume that the covariate effects are generally non-constant. In real applications, the model structure is usually unknown, and misspecifying constant effects as non-constant will probably reduce the estimation efficiency. Huang (2012) constructed a generalized likelihood ratio statistic to test if the coefficient is a constant. However, this approach can be computationally expensive as it requires estimating all possible sub-models. Another limitation is that most existing estimation methods for single-index

varying-coefficient models are proposed for low-dimensional data.

Penalized estimation methods have been developed for model selection and structure identification in single-index varying-coefficient models. Feng and Xue (2013) proposed a penalized estimation method to select and estimate important index parameters and coefficient functions based on spline approximation and the SCAD penalty (Fan and Li, 2001). As an extension, Feng and Xue (2015) imposed an additional penalty to the derivatives of the coefficient functions so as to identify the zero, constant, and varying effects. With the same purpose to distinguish constant and varying covariate effects, Guan (2017) proposed an alternative penalization method using the MCP and extended the penalization framework to a class of generalized linear models in a low-dimensional setting. All existing penalization methods are developed for continuous or binary outcomes, and models for censored event time have not been considered. It is unclear whether existing methods can be extended to accommodate right-censored outcomes, especially under a semi-parametric outcome model with an infinite-dimensional nuisance parameter.

1.3 Organisation of thesis

To address the aforementioned issues, it is highly desirable to develop a penalized regression method to identify the true model structure and estimate the varying coefficients for the censored survival outcomes. We adopt a single-index varying-coefficient model to accommodate the potential interaction effects between two sets of predictors. We propose a penalized estimation method to allow separate selection of constant and varying effects. In an ongoing study, we consider a different approach to accommodate the interaction effects based on a varying-coefficient additive hazards model. We propose a kernel-based estimation method to estimate the constant and varying effects.

This thesis is organized as follows. In Chapter 2, we introduce a novel penalized estimation method for a class of single-index varying-coefficient models. We describe the

model and estimation procedures and provide applications to a motivating cancer genomic study. In Chapter 3, we discuss some practical issues about the integrative approach considered in this thesis and suggest some related directions for further investigation.

Chapter 2

Penalized estimation in single-index varying-coefficient models

2.1 Preliminaries

We propose a penalized (sieve) maximum likelihood estimation method for variable selection and estimation for a class of single-index varying-coefficient models, which accommodates continuous and censored outcomes. We adopt a novel two-part penalty that allows for separate selection of genomic features with non-zero constant effect and those with effects modified by clinical factors. The proposed penalty functions are weighted to unify the degree of shrinkage imposed to the constant and varying effects of a predictor. A coordinate-wise algorithm for computing the penalized estimators is developed. Unlike existing methods, our method accommodates right-censored survival outcomes, which are common in cancer genomic studies. Also, the proposed method is based on convex penalties, which tends to be more computationally stable compared with the existing methods based on the non-convex penalties.

2.2 Methodology

2.2.1 Model, data, and sieve likelihood

Let Y be an outcome of interest, $\boldsymbol{U}$ and $\boldsymbol{Z}$ be two sets of low-dimensional predictors that may overlap, and $\boldsymbol{X} \equiv (X_0, \ldots, X_p)^{\mathrm{T}}$ be a set of potentially high-dimensional predictors with $X_0 = 1$. We are interested in the effect of $(\boldsymbol{X}, \boldsymbol{Z})$ on Y, where the effect of $\boldsymbol{X}$ is allowed to depend on $\boldsymbol{U}$. In genomic studies, Y can be a disease outcome such as time to death, $\boldsymbol{X}$ can be a set of gene expressions, and $\boldsymbol{U}$ and $\boldsymbol{Z}$ can be clinical factors. We assume the following partial linear single-index varying-coefficient model:

$$Y \mid (\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z}) \sim f\left\{ \cdot \, ; \sum_{j=0}^{p} g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}) X_j + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\psi} \right\},$$

where f is a density function, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are regression parameters, and $g_0, \ldots, g_p$ are unspecified smooth functions. For model identifiability, we set $\|\boldsymbol{\beta}\| = 1$, and if $\boldsymbol{U}$ is a subset of $\boldsymbol{Z}$, then we set the component of $\boldsymbol{\psi}$ that corresponds to the last component of $\boldsymbol{U}$ to be 0. This model assumes that the effect of each component of $\boldsymbol{X}$ is characterized by a non-parametric transformation of an index $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}$. If each g_j $(j = 0, \ldots, p)$ is constant, then the model contains only linear effects of $(\boldsymbol{X}, \boldsymbol{Z})$. If g_j is a linear function, then the model contains the linear effect of X_j and the interaction effect of $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}$ and X_j. The proposed model accommodates many different types of outcomes. For continuous or categorical outcomes, we set f to be a density from the exponential family. For survival outcomes, we set f to be the density under the Cox proportional hazards model. Under the Cox model, the conditional hazard function of survival time T given $(\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z})$ takes the form of $h(t) \exp\{\sum_{j=0}^{p} g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}) X_j + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\psi}\}$ for $t \geq 0$, where $h(\cdot)$ is an unspecified baseline hazard function.

Suppose there are n observations. For uncensored outcomes, the observed data consist

of $(Y_i, \boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$. The log-likelihood function is

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \mathcal{G}) = \sum_{i=1}^{n} \log f\left\{ Y_i; \sum_{j=0}^{p} g_j(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta}) X_{ij} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\psi} \right\},$$

where $\mathcal{G} = (g_0, \ldots, g_p)$. For a potentially right-censored outcome, let C_i be the censoring time for the i-th subject, $\widetilde{Y}_i = \min(Y_i, C_i)$, and $\Delta_i = I(Y_i \leq C_i)$. The observed data consist of $(\widetilde{Y}_i, \Delta_i, \boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$. We set ℓ_n to be the log-partial-likelihood function, such that

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \mathcal{G}) = \sum_{i=1}^{n} \Delta_i \Bigg(\sum_{j=0}^{p} g_j(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta}) X_{ij} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\psi} - \log \Bigg[\sum_{h:\widetilde{Y}_h \geq \widetilde{Y}_i} \exp\left\{ \sum_{j=0}^{p} g_j(\boldsymbol{U}_h^{\mathrm{T}}\boldsymbol{\beta}) X_{hj} + \boldsymbol{Z}_h^{\mathrm{T}}\boldsymbol{\psi} \right\} \Bigg] \Bigg).$$

Because the likelihood involves the non-parametric functions $(g_0, \ldots, g_p)$, maximum likelihood estimation is not feasible. There are many choices of basis functions, including the Fourier, spline, and wavelet basis functions, for approximating the non-parametric functions. We adopt B-splines for ease of interpretation and efficient computations. We propose to approximate g_j by B-spline functions. Let $(B_1, \ldots, B_d)$ be a set of B-spline functions on a pre-specified set of grid points, such that each function passes through the origin; the construction of the B-spline functions are discussed in the Appendix. For $j = 0, \ldots, p$, we approximate g_j by $\gamma_j + \sum_{k=1}^{d} \alpha_{jk} B_k$, where $(\gamma_j, \alpha_{j1}, \ldots, \alpha_{jd})$ are regression parameters. For right-censored outcomes, we set $\gamma_0 = 0$ for identifiability. Since the number of regression coefficients increases linearly with the number of knots, choosing too many knot points may lead to unstable estimation or overfitting. Thus, we consider a quadratic B-splines with 3 knots in this work.

2.2.2 Penalized sieve likelihood

When p is large, the total number of parameters may be larger than the sample size, and penalization on $\boldsymbol{\gamma} \equiv (\gamma_0, \ldots, \gamma_p)^{\mathrm{T}}$ and $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_{jk})_{j=0,\ldots,p;k=1,\ldots,d}$ could be adopted

for stable estimation and variable selection. We propose to estimate the parameters by maximizing the following penalized log-likelihood function:

$$p\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \ell_n\left\{\boldsymbol{\beta}, \boldsymbol{\psi}, \left(\gamma_j + \sum_{k=1}^{d} \alpha_{jk} B_k\right)_{j=0,\ldots,p}\right\} - \sum_{j=1}^{p} \rho_1(\gamma_j; \lambda_1) - \sum_{j=1}^{p} \rho_2(\boldsymbol{\alpha}_j; \lambda_2),$$

where ρ_1 and ρ_2 are penalty functions, λ_1 and λ_2 are tuning parameters, and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jd})^{\mathrm{T}}$ for $j = 1, \ldots, p$. This formulation allows separate selection of constant and non-constant effects of X_j by separate penalization on γ_j and $\boldsymbol{\alpha}_j$. Let $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}_j$, and $\widehat{\boldsymbol{\alpha}}_j$ denote the penalized estimator of $\boldsymbol{\beta}$, γ_j, and $\boldsymbol{\alpha}_j$, respectively ($j = 0, \ldots, p$). For $j = 1, \ldots, p$, if $\widehat{\gamma}_j = 0$ and $\widehat{\boldsymbol{\alpha}}_j = \mathbf{0}$, then X_j does not have an effect on the outcome in the estimated model. If only $\widehat{\boldsymbol{\alpha}}_j = \mathbf{0}$, then X_j has a constant effect of $\widehat{\gamma}_j$. If $\widehat{\boldsymbol{\alpha}}_j$ is non-zero, then X_j has a non-constant effect indexed by $\boldsymbol{U}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.

Many choices of penalty functions, such as the (group) lasso (Tibshirani, 1996; Yuan and Lin, 2006), SCAD (Fan and Li, 2001; Breheny and Huang, 2009), and MCP (Zhang, 2007; Breheny and Huang, 2009) are possible. In this work, we propose to set $\rho_1(\gamma_j; \lambda_1) = \lambda_1 w_j |\gamma_j|$ and $\rho_2(\boldsymbol{\alpha}_j; \lambda_2) = \lambda_2 w_j (\boldsymbol{\alpha}_j^{\mathrm{T}} \boldsymbol{K}_j \boldsymbol{\alpha}_j)^{1/2}$, where w_j is a weight for the j-th predictor, and $\boldsymbol{K}_j$ is some $(d \times d)$-symmetric matrix; the first penalty is similar to the adaptive lasso penalty (Zou, 2006), and the second penalty is a weighted version of the group lasso.

There are two advantages of the adaptive-lasso type penalty over other commonly used penalty functions. First, the adaptive-lasso type penalty can reduce the estimation bias while retaining the convexity of the regularization term. Although lasso is computationally efficient, the shrinkage introduced by lasso may result in substantial biases for large regression coefficients. Non-convex penalties, such as SCAD and MCP, are designed to diminish this bias while achieving selection consistency. However, this class of penalty functions tends to possess multiple local optima. To preserve stability, we suggest using a weighted convex penalty function. Second, the weighted penalty utilizes the information

that the constant and non-constant effects correspond to the same predictor. Although conventional choices of penalty functions for ρ_1 and ρ_2 can produce sparse estimation of the constant and non-constant effects, they fail to take into account the fact that γ_j and $\boldsymbol{\alpha}_j$ $(j = 1, \ldots, p)$ correspond to the same predictor X_j. The weight w_j is introduced to capture the overall signal strength of g_j and unify the degree of shrinkage of γ_j and $\boldsymbol{\alpha}_j$. In particular, we set $w_j = (\widetilde{\gamma}_j^2 + \|\widetilde{\boldsymbol{\alpha}}_j\|^2)^{-1/2}$, where $\widetilde{\gamma}_j$ and $\widetilde{\boldsymbol{\alpha}}_j$ are estimates of γ_j and $\boldsymbol{\alpha}_j$ obtained from maximizing the penalized log-likelihood with $w_j = 1$ for $j = 1, \ldots, p$. If the initial estimates $\widetilde{\gamma}_j$ and $\widetilde{\boldsymbol{\alpha}}_j$ are accurate in that variables with stronger signal receive smaller weights, then the weighted estimators would yield better variable selection and estimation accuracy than unweighted estimators.

2.2.3 Estimation

We propose to compute the estimates of $(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$ using an alternating algorithm. In particular, we initialize $\boldsymbol{\beta}$ as some unit vector such that $\|\boldsymbol{\beta}\| = 1$ and update the parameter estimates of $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$ and $\boldsymbol{\beta}$ alternatively until convergence. The algorithm is as follows:

Step 1: Initialize $\widehat{\boldsymbol{\beta}}^{(0)}$ as some unit vector such that $\|\widehat{\boldsymbol{\beta}}^{(0)}\| = 1$. Set $m = 1$.

Step 2: Update $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$ by

$$\big(\widehat{\boldsymbol{\gamma}}^{(m)}, \widehat{\boldsymbol{\alpha}}^{(m)}, \widehat{\boldsymbol{\psi}}^{(m)}\big) \equiv \underset{(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})}{\arg\max}\, p\ell_n\big(\widehat{\boldsymbol{\beta}}^{(m-1)}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi}\big).$$

For fixed $\boldsymbol{\beta}$, the objective function $p\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$ is essentially a penalized log-likelihood function for a conventional regression model under a group lasso penalty, and $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$ can be updated using existing algorithms for the group lasso (Breheny and Huang, 2009).

Step 3: Update $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}}^{(m)} \equiv \underset{\|\boldsymbol{\beta}\|=1}{\arg\max}\, \ell_n\left\{\boldsymbol{\beta}, \widehat{\boldsymbol{\psi}}^{(m)}, \left(\widehat{\gamma}_j^{(m)} + \sum_{k=1}^{d} \widehat{\alpha}_{jk}^{(m)} B_k\right)_{j=0,\ldots,p}\right\}.$$

For fixed $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$, $\boldsymbol{\beta}$ can be updated using the Lagrange multiplier method, and $\widehat{\beta}$ is defined as the solution of the equations:

$$\begin{cases} \frac{\partial}{\partial \boldsymbol{\beta}} \ell_n\left\{\boldsymbol{\beta}, \boldsymbol{\psi}, \left(\gamma_j + \sum_{k=1}^{d} \alpha_{jk} B_k\right)_{j=0,\ldots,p}\right\} + c\boldsymbol{\beta} = \mathbf{0} \\ \|\boldsymbol{\beta}\|^2 - 1 = 0, \end{cases}$$

where c is the Lagrange multiplier. We can solve for $\boldsymbol{\beta}$ and c simultaneously using the Newton-Raphson algorithm.

Step 4: Set $m = m + 1$. Repeat Steps 2–4 until convergence.

The performance of the proposed methods depends critically on the choice of tuning parameters. We propose to select the tuning parameters λ_1 and λ_2 using a version of the Bayesian information criterion (BIC), defined as

$$-2\ell_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}, \widehat{\mathcal{G}}) + q \log(n^*),$$

where $\widehat{\mathcal{G}} = (\widehat{\gamma}_j + \sum_{k=1}^{d} \widehat{\alpha}_{jk} B_k)_{j=0,\ldots,p}$, q is the effective degrees of freedom, and n^* is the effective sample size. Specifically, $n^* = n$ for uncensored outcomes, and n^* is the number of uncensored observations for right-censored outcomes. Following Breheny and Huang (2009), we define the effective degrees of freedom as

$$q = \sum_{j=1}^{p} \left(\frac{\widehat{\gamma}_j}{\widehat{\gamma}_j^*} + \sum_{k=1}^{d} \frac{\widehat{\alpha}_{jk}}{\widehat{\alpha}_{jk}^*}\right),$$

where $(\widehat{\gamma}_j, \widehat{\alpha}_{jk})$ denote the estimated value of (γ_j, α_{jk}), $\widehat{\gamma}_j^*$ denote the maximizer of the unpenalized log-likelihood function with respect to γ_j with other parameters fixed at the

estimated value, and $\widehat{\alpha}^*_{jk}$ denote the maximizer of the unpenalized log-likelihood function with respect to α_{jk} with other parameters fixed at the estimated value. We select (λ_1, λ_2) that yield the minimum modified BIC value.

In conventional group lasso problems, the predictor matrix of the j-th group, denoted by $\boldsymbol{W}_j$, is typically transformed such that $\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_j$ is a diagonal matrix with equal diagonal elements. This is equivalent to setting $\boldsymbol{K}_j$ to be (a scaled version of) $\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_j$. In the current problem, however, the "predictor matrix," which consists of rows $(X_{ij}, B_1(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij}, \ldots, B_d(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij})$ $(i = 1, \ldots, n)$, depends on the unknown parameter $\boldsymbol{\beta}$. One estimation strategy is to set $\boldsymbol{K}_j$ based on the predictor matrix evaluated at some initial estimator of $\boldsymbol{\beta}$, such as that obtained under $\boldsymbol{K}_j = \boldsymbol{I}$. Another strategy is to update $\boldsymbol{K}_j$ with $\boldsymbol{\beta}$ after each iteration; this can be thought of as setting $\boldsymbol{K}_j$ based on the converged value of $\boldsymbol{\beta}$. Another difficulty that arises from the unknown $\boldsymbol{\beta}$ is that the converged estimates may vary with the initial value of $\boldsymbol{\beta}$. We propose to consider multiple initial values and select the final estimates that yield the smallest modified BIC. In the simulation studies, we considered 5 initial values of $\boldsymbol{\beta}$ and updated $\boldsymbol{K}$ along with $\boldsymbol{\beta}$ at each iteration, and the algorithm converged at almost all replicates.

2.3 Simulation studies

We set the dimension of $\boldsymbol{U}$ to be 4 and generated components of $\boldsymbol{U}$ as i.i.d. standard normal variables. We set $\boldsymbol{Z} = \boldsymbol{U}$ and generated $\boldsymbol{X}$ from the p-variate standard normal distribution. We set $\boldsymbol{\beta} = (0.4, -0.4, 0.2, -0.8)^{\mathrm{T}}$, $\boldsymbol{\psi} = (0.2, -0.2, 0.5, -0.5)^{\mathrm{T}}$, and $g_1, \ldots, g_{20}$ to be non-zero constant, linear, or non-linear functions; the functions are plotted in Figure 2.3. We set g_0 and $g_{21}, \ldots, g_p$ to be constant at 0. We considered a continuous outcome variable and a right-censored outcome variable. For the continuous outcome, we set $f(y; \mu) = (2\pi)^{-1/2} \exp\{-(y-\mu)^2/2\}$, so that conditional on $(\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z})$, Y follows the normal distribution with unit variance. For the right-censored outcome, we set

$f(y;\mu) = h(y)\exp(\mu)\exp\big\{-\exp(\mu)\int_0^y h(t)\,\mathrm{d}t\big\}$, where h is the baseline hazard function with $h(t) = t$. The censoring time was generated from an exponential distribution with the mean chosen to yield a censoring rate of about 30%. In each setting, we considered a sample size of 500 and $p = 20$, 50, and 100.

We compare the proposed methods with conventional regression models with or without interaction terms. For the proposed methods, we simply set the degree of the B-spline functions to be 2 and the knots at $-\max_i \|\boldsymbol{U}_i\|_2$, 0, and $\max_i \|\boldsymbol{U}_i\|_2$. We considered the proposed weighted approach and an unweighted approach with $w_j = 1$ $(j = 1, \ldots, p)$. We also considered the lasso regression on the linear predictors $(\boldsymbol{X}, \boldsymbol{Z})$ and the lasso regression on $\boldsymbol{X}$, $\boldsymbol{Z}$, and pairwise interactions between components of $\boldsymbol{X}$ and $\boldsymbol{U}$; in both cases, coefficients of $\boldsymbol{Z}$ were not penalized. In addition, we considered adaptive lasso for the models with or without interactions, where the weights are the inverse of the absolute value of the corresponding lasso estimates. In all methods, the tuning parameters were selected using the modified BIC.

We evaluate the performance of each method in terms of variable selection and prediction. For variable selection, we report the sensitivity and the false discovery rate (FDR). Sensitivity is the proportion of correctly identified signal variables among all true signal variables. FDR is the proportion of noise variables that are incorrectly identified as signal variables among all selected variables. For the proposed methods, a variable X_j is selected if either γ_j or $\boldsymbol{\alpha}_j$ is estimated as non-zero $(j = 1, \ldots, p)$. For the proposed methods and lasso with interactions, we also report the sensitivity and FDR with respect to the selection of non-constant effects, where for the proposed methods, the non-constant effect of X_j is selected if $\widehat{\boldsymbol{\alpha}}_j \neq \mathbf{0}$, and for lasso with interactions, the non-constant effect is selected if the coefficient of the product of X_j and any component of $\boldsymbol{U}$ is non-zero. In addition, we report the total numbers of the selected variables and the number of variables identified to have non-constant effects.

For prediction, we report the mean-squared error (MSE), defined as $\mathrm{E}(\widehat{\eta} - \eta_0)^2$, where

$\eta_0 = \eta(\boldsymbol{\beta}_0, \mathcal{G}_0, \boldsymbol{\psi}_0)$, $\eta(\boldsymbol{\beta}, \mathcal{G}, \boldsymbol{\psi}) \equiv \sum_{j=1}^{p} g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta})X_j + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\psi}$, and $(\boldsymbol{\beta}_0, \mathcal{G}_0, \boldsymbol{\psi}_0)$ denote the true parameter values. For the proposed methods, $\widehat{\eta} = \eta(\widehat{\boldsymbol{\beta}}, \widehat{\mathcal{G}}, \widehat{\boldsymbol{\psi}})$, where $(\widehat{\boldsymbol{\beta}}, \widehat{\mathcal{G}}, \widehat{\boldsymbol{\psi}})$ denote the estimated parameter values. For lasso with and without interaction effects, $\widehat{\eta} = \sum_j \widehat{b}_j X_j + \sum_k \widehat{c}_k Z_k + \sum_{j,l} \widehat{d}_{jl} X_j U_l$ and $\widehat{\eta} = \sum_j \widetilde{b}_j X_j + \sum_k \widetilde{c}_k Z_k$, respectively, where $\widehat{b}_j$, $\widehat{c}_k$, $\widehat{d}_{jl}$, $\widetilde{b}_j$, and $\widetilde{c}_k$ are the corresponding estimated regression parameters. For the right-censored outcome, we also compute the concordance index (C-index) (Harrell et al., 1982), defined as $\mathrm{P}(\eta_i > \eta_j \mid \widetilde{Y}_i < \widetilde{Y}_j)$ for two generic independent subjects indexed by i and j. C-index typically takes values between 0.5 and 1, where a value of 0.5 indicates no discrimination and a value of 1 indicates perfect discrimination. We compute the MSE and C-index on a set of independently generated data set of size 5000. For the proposed methods, we also report the absolute inner product $|\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}|$ to assess the estimation accuracy of $\widehat{\boldsymbol{\beta}}$. The simulation results for the continuous and right-censored outcomes based on 100 replicates are summarized in Tables 2.1 and 2.2, respectively. Figures 2.1 – 2.6 show the average estimated values of $g_1, \ldots, g_{20}$ under different settings.

Table 2.1: Simulation results for the continuous outcome.

		$p = 20$			$p = 50$			$p = 100$		
		Proposed	Main	Interaction	Proposed	Main	Interaction	Proposed	Main	Interaction
Unweighted										
SEN	Overall	0.990	0.800	0.978	0.978	0.776	0.966	0.962	0.754	0.960
	Non-constant	0.999	-	0.949	0.987	-	0.917	0.983	-	0.912
FDR	Overall	0	0	0	0.377	0.295	0.504	0.543	0.479	0.706
	Non-constant	0.253	-	0.426	0.436	-	0.718	0.530	-	0.833
NS	Overall	19.80	16.01	19.57	31.57	22.21	39.13	42.44	29.30	65.60
	Non-constant	13.48	0	16.65	17.70	0	32.84	21.17	0	54.84
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.997	-	-	0.996	-	-	0.996	-	-
MSE		0.378	1.583	0.990	0.569	1.754	1.160	0.698	1.764	1.206
Weighted										
SEN	Overall	0.948	0.657	0.908	0.915	0.649	0.900	0.898	0.638	0.897
	Non-constant	0.979	-	0.848	0.951	-	0.826	0.915	-	0.834
FDR	Overall	0	0	0	0.155	0.096	0.324	0.247	0.200	0.540
	Non-constant	0.069	-	0.243	0.154	-	0.552	0.222	-	0.709
NS	Overall	18.96	13.14	18.15	21.79	14.47	26.90	24.11	16.19	39.43
	Non-constant	10.57	0	11.36	11.33	0	18.83	11.93	0	29.19
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.998	-	-	0.997	-	-	0.997	-	-
MSE		0.313	1.602	1.008	0.408	1.724	1.135	0.544	1.730	1.198

NOTE: "SEN" represents sensitivity; "NS" represents number of selected variables; "Main" represents lasso regression model without interactions; "Interaction" represents lasso regression model with interactions; "Overall" gives values of corresponding measures concerning all components of $\boldsymbol{X}$; "Non-constant" gives values of corresponding measures concerning components of $\boldsymbol{X}$ with non-constant effects on the outcome.

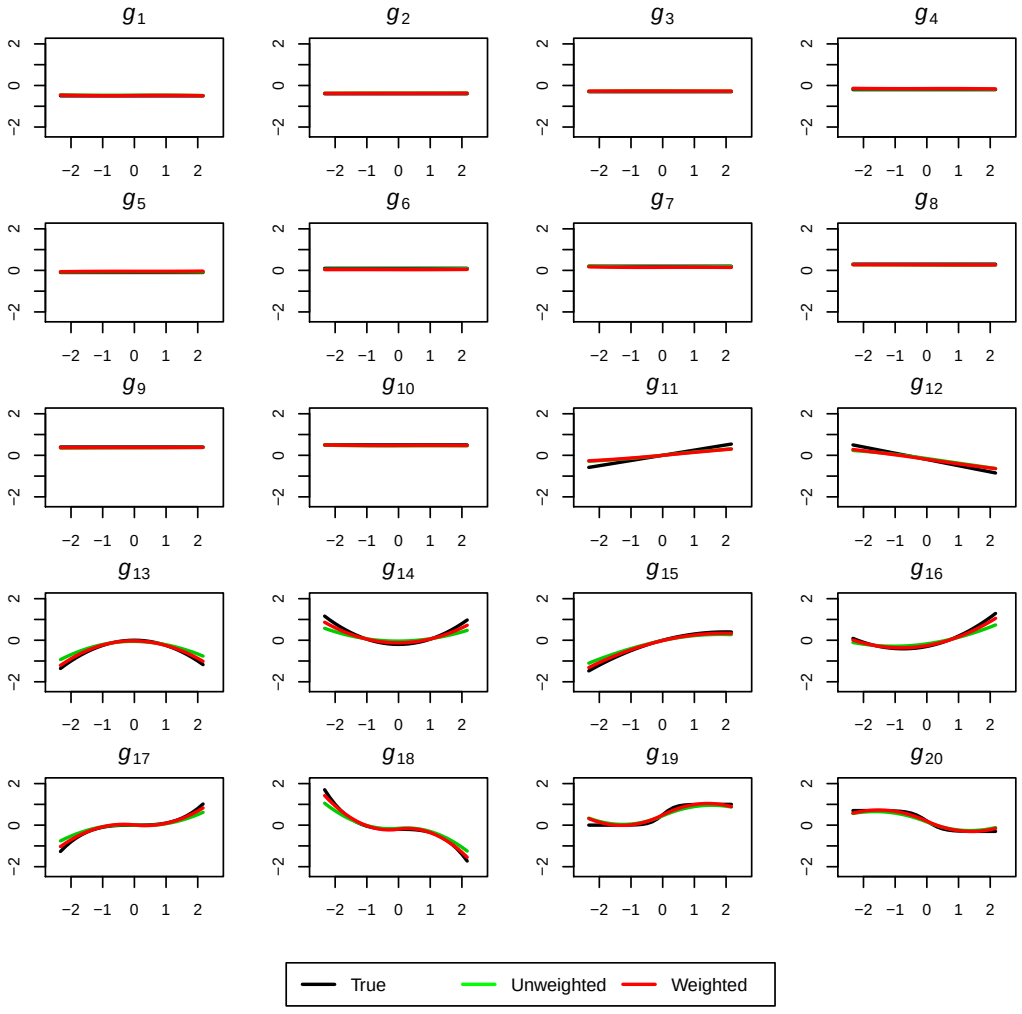

Figure 2.1: Estimated coefficients for the continuous outcome under $p = 20$.

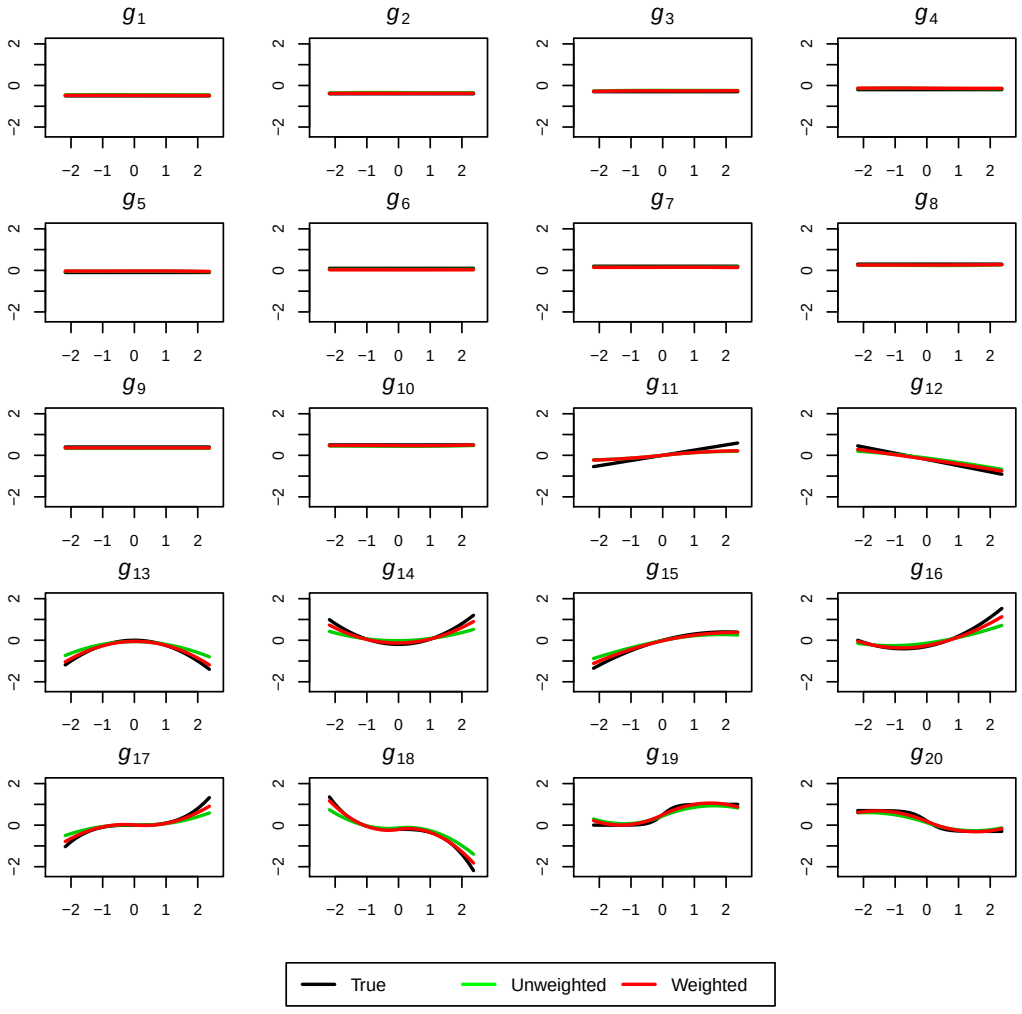

Figure 2.2: Estimated coefficients for the continuous outcome under $p = 50$.

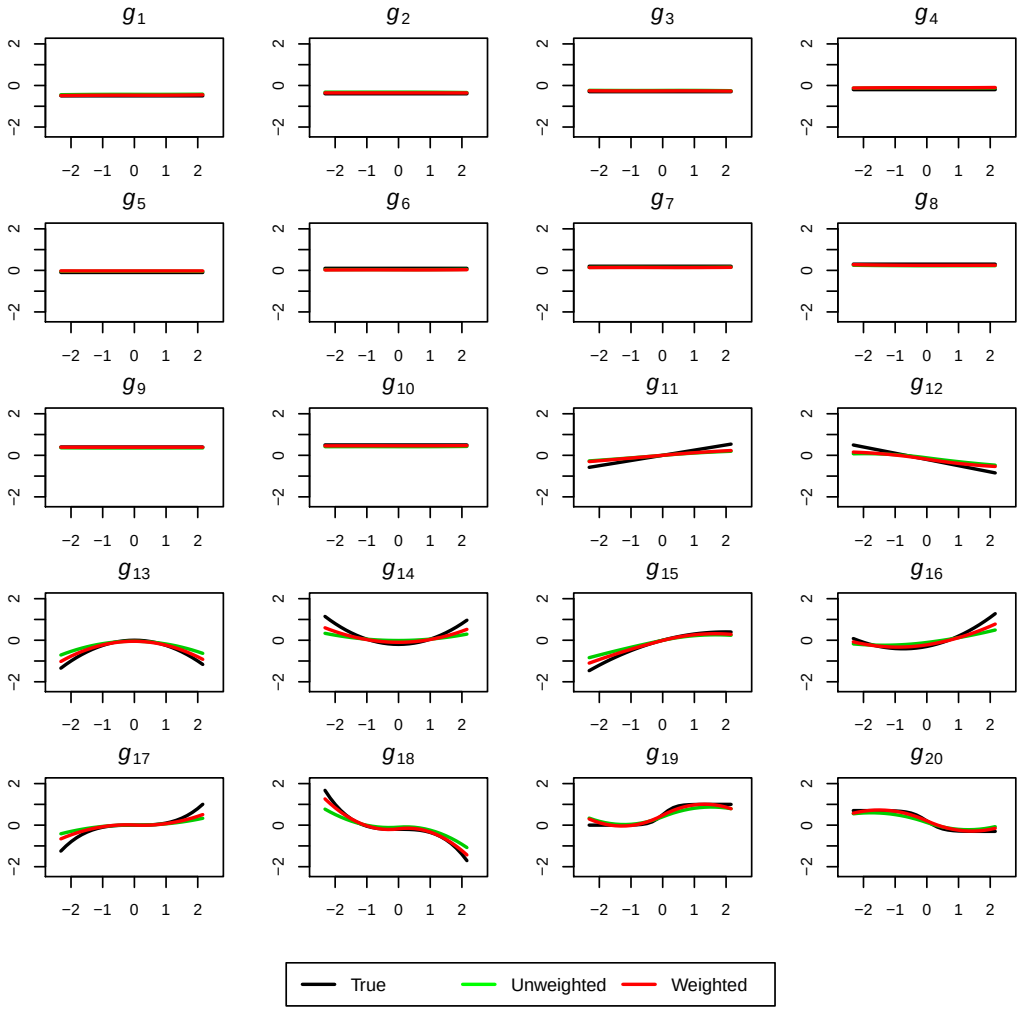

Figure 2.3: Estimated coefficients for the continuous outcome under $p = 100$.

Table 2.2: Simulation results for the right-censored outcome.

		$p = 20$			$p = 50$			$p = 100$		
		Proposed	Main	Interaction	Proposed	Main	Interaction	Proposed	Main	Interaction
Unweighted										
SEN	Overall	0.962	0.795	0.948	0.909	0.737	0.914	0.856	0.689	0.874
	Non-constant	0.916	-	0.895	0.805	-	0.852	0.702	-	0.794
FDR	Overall	0	0	0	0.319	0.297	0.446	0.477	0.480	0.629
	Non-constant	0.176	-	0.392	0.283	-	0.661	0.386	-	0.776
NS	Overall	19.24	15.90	18.97	26.95	21.11	33.28	33.06	26.78	48.24
	Non-constant	11.28	0	14.91	11.42	0	25.51	11.84	0	36.66
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.993	-	-	0.974	-	-	0.957	-	-
MSE		0.843	1.873	1.419	1.357	2.142	1.701	1.518	2.134	1.746
C-index		0.772	0.716	0.743	0.758	0.716	0.738	0.745	0.708	0.727
Weighted										
SEN	Overall	0.883	0.631	0.864	0.831	0.615	0.832	0.763	0.580	0.787
	Non-constant	0.806	-	0.780	0.708	-	0.732	0.593	-	0.687
FDR	Overall	0	0	0	0.136	0.139	0.303	0.239	0.273	0.489
	Non-constant	0.073	-	0.238	0.148	-	0.518	0.217	-	0.658
NS	Overall	17.66	12.62	17.27	19.35	14.39	24.13	20.37	16.19	31.22
	Non-constant	8.78	0	10.40	8.42	0	15.60	7.64	0	21.03
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.994	-	-	0.991	-	-	0.982	-	-
MSE		0.691	1.822	1.209	0.929	1.983	1.414	1.166	1.947	1.528
C-index		0.773	0.714	0.743	0.766	0.716	0.740	0.754	0.710	0.726

NOTE: See NOTE to Table 2.1.

In terms of prediction, both the weighted and unweighted versions of the proposed methods correctly identify the interaction structure between $\boldsymbol{X}$ and $\boldsymbol{U}$ and yield higher prediction accuracy than other methods. In particular, they yield lower MSE in all settings and higher C-index for the right-censored outcome. In addition, the estimated value of $\boldsymbol{\beta}$ is close to the true value, indicating that the proposed methods can correctly identify the composition of the index. The weighted estimators are generally accurate, whereas the unweighted estimators tend to be biased towards zero due to the uniform shrinkage imposed on all parameters. Lasso with interaction terms generally yields smaller MSE than lasso with main effects alone, suggesting that a varying-coefficient model can be approximated by a conventional regression model with pairwise interaction terms. Nevertheless, possibly due to the complexity of the interaction model, the performance of lasso with interaction is substantially worse than that of the proposed methods.

In terms of variable selection, both the proposed methods and lasso with interactions have substantially higher sensitivity than lasso with main effects alone. The FDR is lower under the proposed methods than lasso with interactions, indicating that the proposed methods tend to yield more interpretable models. The FDR for the proposed methods is higher than those for lasso with main effects alone under some settings, possibly because lasso with main effects alone generally selects much fewer variables. For all methods, the weighted estimators yield substantially lower FDR than the unweighted estimators. By setting higher penalty for noise variables and lower penalty for signal variables, the weighted method yields higher variable selection accuracy.

2.4 Real data examples

2.4.1 TCGA NSCLC data set

We demonstrate the application of the proposed methods using a set of NSCLC patients from TCGA. The data set consists of two subtypes of lung cancer, namely lung adeno-

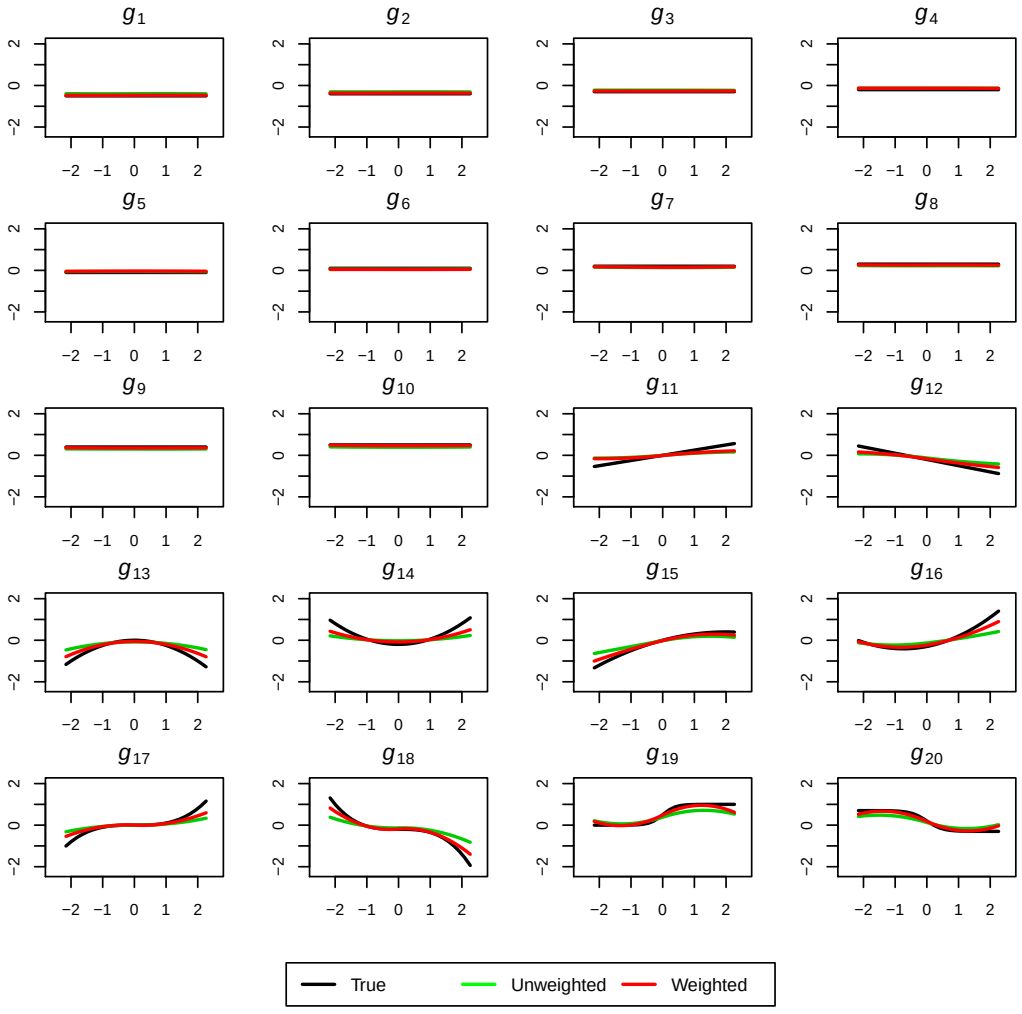

Figure 2.4: Estimated coefficients for the right-censored outcome under $p = 20$.

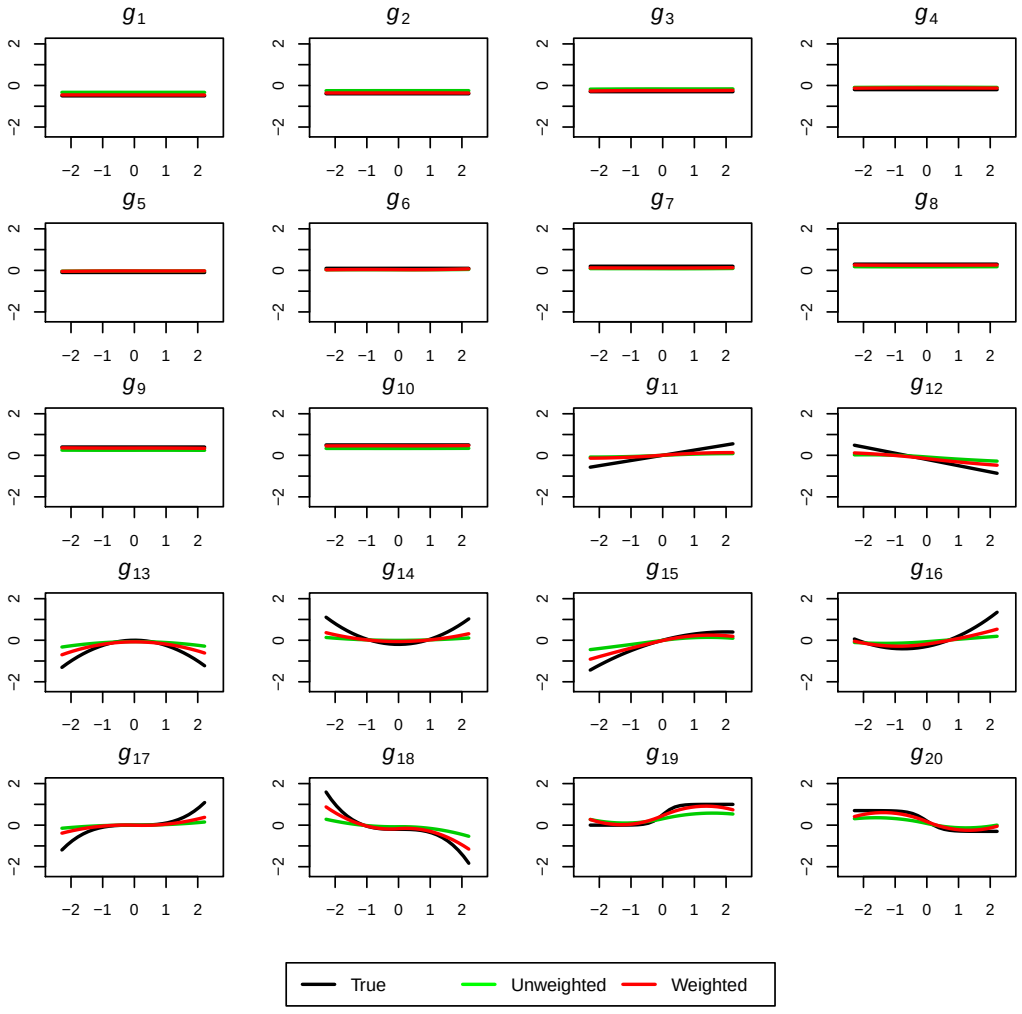

Figure 2.5: Estimated coefficients for the right-censored outcome under $p = 50$.

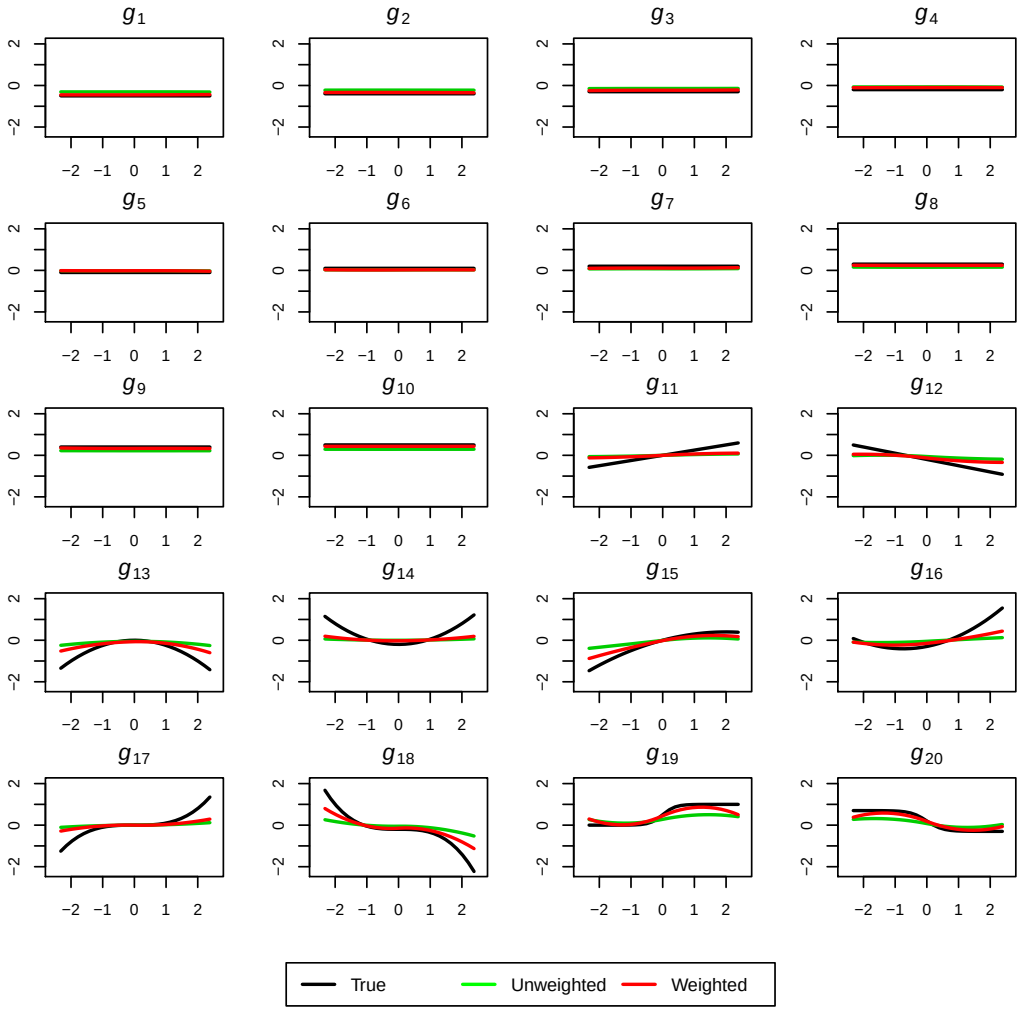

Figure 2.6: Estimated coefficients for the right-censored outcome under $p = 100$.

carcinoma (LUAD) and lung squamous cell carcinoma (LUSC). We are interested in the potential risk factors associated with pulmonary function, measured by the percentage of expiratory volume in one second (FEV1); a higher FEV1 represents larger lung capacity, and patients with severely impaired lung function have an increased risk of mortality (Hole et al., 1996). In particular, we investigated the effects of gene expressions and clinical variables on FEV1, allowing for interactions between the two types of variables. We fit the proposed model with $\boldsymbol{U}$ consisting of age, number of packs of cigarettes smoked per year or pack-year smoked (PYS), cancer subtype, tumor stage, and gender; tumor stage is dichotomized into stage I versus stage II or above. This formulation allows the effects of genomic factors to be modified by clinical variables. We set $\boldsymbol{Z} = \boldsymbol{U}$ to allow linear effects of clinical variables on FEV1. After discarding genes with zero expressions for 30% or more subjects, the data set consists of 17148 gene expressions and the number of gene expressions is substantially greater than the sample size. For stable estimation, we performed screening to filter out the gene expressions that appear to contribute little signal to the variations in FEV1. We set $\boldsymbol{X}$ to consist of 300 gene expressions that have the most significant marginal association with FEV1 (adjusted for clinical variables). After removing subjects with missing data, the sample size is 353, with 185 and 168 LUAD and LUSC patients, respectively. Following the simulation studies, we set the degree of the B-spline functions to be 2 and the knots at $-\max_i \|\boldsymbol{U}_i\|_2$, 0, and $\max_i \|\boldsymbol{U}_i\|_2$. We adopted the weighted penalty approach. We standardized all variables to have zero mean and unit variance.

We identified 17 gene expressions to be associated with FEV1. The selected gene expressions and their estimated coefficients are shown in Table 2.3. Among the selected gene expressions, EIF4A3 was known to be involved in the development of NSCLC, and KCNK2 and N4BP1 were known as prognostic factors in some cancer types (Innamaa et al., 2013; Xu et al., 2017; Lin et al., 2018; Li et al., 2019). The effects of CDK11A and LRRC29 were identified to vary with the clinical variables; CDK11A has previously been

Table 2.3: Selected gene expressions for TCGA NSCLC analysis.

Gene	Coefficient
ANKRD13D	0.143
CDK11A	(varying)
CRELD2	1.671
C12orf56	−1.611
C8orf38	2.008
C8orf58	2.348
EIF4A3	−0.069
KCNK2	−2.506
LOC642826	1.233
LRRC29	(varying)
LRRTM2	−1.168
NRN1L	1.553
N4BP1	1.449
PLEKHG4B	0.724
RNF122	−3.429
THAP4	1.591
ZNF75D	2.656

NOTE: "varying" represents a varying coefficient. The estimated coefficient functions are given in Figure 2.7.

shown to be associated with many cancer types (Zhou et al., 2016). The estimated index parameters $\boldsymbol{\beta}$ for age, PYS, gender, tumor stage, and cancer subtype are 0.199, 0.637, 0.157, −0.548, and −0.479, respectively. The index is dominated by PYS, tumor stage, and cancer subtype, suggesting that the effects of CDK11A and LRRC29 mainly depend on these three clinical factors. Figure 2.7 displays the estimated values of g_0 and the g functions for CDK11A and LRRC29 under different index values.

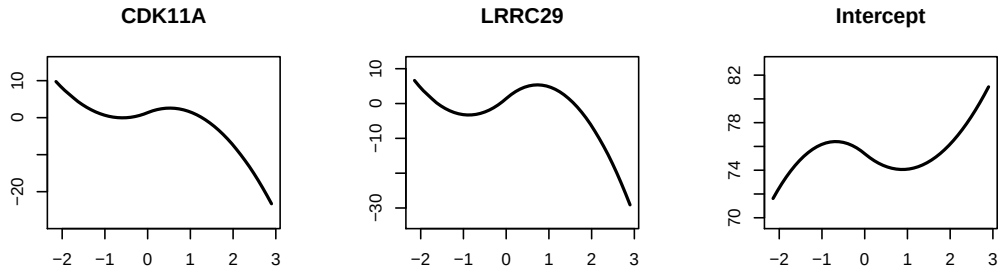

Figure 2.7: Estimated coefficients for TCGA NSCLC analysis.

2.4.2 TCGA LGG data set

We also applied the proposed methods to identify potential risk factors associated with the survival of patients diagnosed with lower-grade glioma (LGG) in TCGA. The data set consists of grade II and grade III tumors. Instead of integrating clinical and a single type of genomic variables, we investigated the effects of protein expressions, gene expressions, and clinical variables on time to death since initial diagnosis, allowing for interactions between protein and gene expressions. After discarding genes with zero expressions for 30% or more subjects, the data set consists of 17238 gene expressions. We set the overall survival time to be the outcome of interest, which is potentially right-censored. We reduced the dimension of gene expressions using principal component analysis and set $\boldsymbol{U}$ to be the first 7 principal components, which account for over 50% of the total variability. The set of linear predictors $\boldsymbol{Z}$ consists of $\boldsymbol{U}$, age, histological grade, and gender. The set of predictors $\boldsymbol{X}$ includes the expressions of 209 proteins or phospho-proteins. After removing subjects with missing data, the sample size is 423. The median time to censoring or death is 630 days, and the censoring rate is 76.83%.

We identified 7 important protein expressions to be associated with overall survival. The selected protein expressions and their estimated coefficients are shown in Table 2.4. Some of the selected proteins, including FoxM1, HSP70, and Cyclin B1, have previously

Table 2.4: Selected protein expressions for TCGA LGG analysis.

Protein	Coefficient
Cyclin B1	(varying)
FoxM1	-0.001
HER3	-0.149
HSP70	-0.393
MRE11	-0.241
Stathmin	0.251
ERCC5	0.257

NOTE: "varying" represents a varying coefficient. The estimated coefficient functions are given in Figure 2.8.

been shown to be associated with the survival of glioma patients (Chen et al., 2008; Beaman et al., 2014; Zhang et al., 2017). The effect of Cyclin B1 was identified to vary with the gene expressions. The estimated index parameters $\boldsymbol{\beta}$ correspond to the first 7 principal components are 0.192, 0.224, -0.320, 0.059, 0.343, 0.114, and 0.823 respectively. Figure 2.8 displays the estimated values of g_0 and the g function for Cyclin B1 under different index values.

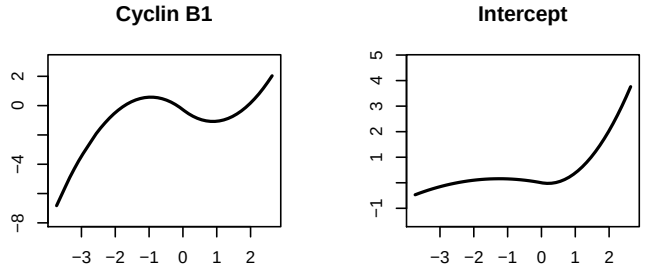

Figure 2.8: Estimated coefficients for TCGA LGG analysis.

2.5 Discussion and future work

In this chapter, we considered a single-index varying-coefficient model for the integration of clinical and genomic features, where the effects of genomic features are allowed to vary with clinical factors. The effects of genomic features are set as non-parametric functions of (a projection of) the clinical factors to accommodate intrinsically different scales of measurements between clinical and genomic data. Unlike most existing estimation methods for varying-coefficient models, our penalized approach separately selects for predictors with constant effects and those with varying effects. Numerical studies illustrate that the proposed methods effectively distinguish zero, constant, and non-constant effects and yield accurate predictions.

The proposed methods are general and can be applied with different choices of penalty functions, basis functions, or outcome distributions. As shown in the simulation studies, the adaptive-lasso-type estimators are generally more accurate than the lasso estimators. Nevertheless, the weighted method is computationally expensive as it requires model estimation twice. Other penalties that attempt to reduce the estimation bias, including SCAD and MCP, can be chosen for the constant or varying effects. But these non-convex penalties tend to possess multiple local optima and result in non-convergence. Also, different families of splines, such as integrated B-splines (Ramsay, 1988) and convex splines (Meyer, 2008), can be used to approximate the non-parametric functions. For example, we may adopt the integrated B-splines to enforce the monotonicity of splines. However, the estimation of regression coefficients may be challenging since the associated regression coefficients should be with the same sign (except for the constant term). Furthermore, different outcome models, such as the accelerated failure time model or additive hazards model, can be adopted.

There are several possible directions for future research. First, we may be interested in the interaction between two types of high-dimensional predictors, in which case the

predictor vector $\boldsymbol{U}$ is high-dimensional. One possible approach is to project $\boldsymbol{U}$ to a low-dimensional space prior to fitting the proposed model. For example, as in the analysis of the LGG data, the projection can be performed by principal component analysis. However, the projected features may not have simple interpretations because it turns the original features into components that are combinations of different features. Another possible approach is to perform variable selection on $\boldsymbol{U}$ by introducing an extra penalty on $\boldsymbol{\beta}$ (Peng and Huang, 2011; Feng and Xue, 2013, 2015; Radchenko, 2015). This approach would involve substantial computational difficulty due to the introduction of an extra penalty term. Second, it is of interest to consider more than two data types. A possibility is to introduce extra indices that correspond to the extra data types so that the effect of a variable may be a function of multiple indices. This approach, however, faces enormous computational challenges because it involves multivariate non-parametric functions.

2.6 Appendix: Construction of basis functions

We discuss the construction of 2-degree basis functions that pass through the origin and are continuously differentiable; basis functions of a general degree can be constructed analogously. Let $(k_1, \ldots, k_d)$ be an ordered set of grid points, where the number of grid points d is odd and is larger than 2, and $k_{(d+1)/2} = 0$. Let $d' = (d+1)/2$, $(\widetilde{L}_1, \ldots, \widetilde{L}_{d'})$ be a set of 2-degree B-spline functions on $(0, -k_{d'-1}, \ldots, -k_1)$, and $(R_1, \ldots, R_{d'})$ be a set of 2-degree B-spline functions on $(0, k_{d'+1}, \ldots, k_d)$. All B-spline functions do not have an intercept, such that $\widetilde{L}_1(0) = \cdots = \widetilde{L}_{d'}(0) = R_1(0) = \cdots = R_{d'}(0) = 0$. Let $L_j = \widetilde{L}_j(-x)$ for $j = 1, \ldots, d'$. The set of continuously differentiable spline functions spanned by these B-spline functions is therefore

$$\left\{ f = \sum_{j=1}^{d'} c_j L_j + \sum_{j=1}^{d'} c_{j+d'} R_j : (c_1, \ldots, c_{2d'}) \in \mathbb{R}^{2d'}, \sum_{j=1}^{d'} c_j L_j^{(1)}(0) = \sum_{j=1}^{d'} c_{j+d'} R_j^{(1)}(0) \right\},$$

where $h^{(1)}$ denotes the first derivative of the function h. We can then construct the basis function as

$$\Big(L_1 + \frac{k_{d'+1}}{k_{d'-1}} R_1, L_2, \ldots, L_{d'}, R_2, \ldots, R_{d'}\Big).$$

Chapter 3

Parameter estimation in the varying-coefficient additive hazards model

3.1 Preliminaries

In the previous chapter, we have discussed an integrative approach to characterize the interaction effects between two data types by mapping the features from one data type (known as the effect modifiers) linearly into an index and allowing the covariate effects of the features from another data type to vary with the index. This approach is based on the assumption that a unique index can be used to describe the varying covariate effects of different covariates. Through summarizing the effect modifiers in an index, the resulting model is easy to interpret because we can visualize the variation of the covariate effects at different index values. However, in some applications, it is not appropriate to assume that all covariate effects are modified by a unique index. For example, in cancer genomics, a gene mutation can affect the behavior of genes as well as their effects on tumor progression in many ways. In this case, it is rigid to use the same composition

of gene mutations as the index to describe the changes of covariate effects for different genes.

Apart from the single-index varying-coefficient model, other forms of varying-coefficient models have been explored and used to accommodate the potential interaction effects between two heterogeneous data types. For example, Ni et al. (2019) studied a varying-sparsity accelerated failure time model and proposed a Bayesian framework to identify relevant protein-gene interactions on cancer progression. While allowing for the effect of proteins on patients' survival to be modified by different gene sets, their method requires prior knowledge on the coding genes specific to each protein. Also, Ma and Song (2015) considered a varying index coefficient model that allows the covariate effects to be modified by distinct index values. However, incorporating multiple indices faces enormous computational challenges, and it is even infeasible when the number of covariates is large. On top of the computational challenges, summarizing the effect modifiers linearly can also be restrictive.

We may consider an alternative way to accommodate multivariate effect modifiers instead of mapping them into an index. As reviewed in Chapter 1, spline and kernel are two popular non-parametric estimation techniques that can be used to estimate the varying coefficients. Indeed, these techniques have been extended to accommodate multivariate variables for density estimation and can be generalized to approximate the varying coefficients. In spline regression, a (multivariate) spline is determined by a given grid configuration. For example, a univariate spline is a piecewise polynomial separated by the knots, a bivariate spline is a piecewise polynomial in two variables separated by a grid of curves, and so on. In kernel regression, the basic difference between multivariate kernel from its univariate analog is the bandwidth estimator. Silverman (1986) considered a multivariate kernel and defined the bandwidth estimator as a diagonal matrix with size depending on the number of effect modifiers. In our context, the multivariate kernel is preferred to the multivariate spline for involving fewer hyperparameters. For the

multivariate spline, we need to specify a set of grid points in each dimension. Whereas for the multivariate kernel, we only need to specify a bandwidth in each dimension. Therefore, we focus on the multivariate kernel smoothing technique to estimate the varying coefficients characterized by multivariate effect modifiers.

In this ongoing study, we consider a varying-coefficient model for the right-censored survival outcomes, which is of most practical interest in genomic studies. We adopt our idea on an additive hazards model by Lin and Ying (1994) with time-independent covariate effects. The additive hazards model is a useful alternative to the proportional hazards model when the proportionality assumption is in doubt. With the purpose to derive a computationally efficient estimator for the varying covariate effects, the additive hazards model is more appealing than the proportional hazards model for several reasons. First, the additive hazards model with time-independent covariate effects has an explicit closed-form solution that may lower the incurred computational cost for estimation. Second, the estimator from the additive hazards model is relatively easier to study theoretically as its asymptotic covariance matrix does not depend on the regression parameter (see Lin and Ying, 1994).

3.2 Models and estimations

3.2.1 Lin and Ying's additive hazards model

Before introducing the proposed model, we briefly review Lin and Ying's (1994) additive hazards model. For an event time T and a vector of possibly time-dependent covariates $\boldsymbol{X}$ with corresponding parameter $\boldsymbol{\beta}$, Lin and Ying (1994) considered the following hazard function:

$$\lambda(t \mid \boldsymbol{X}) = \lambda_0(t) + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}(t),$$

where λ_0 is an unspecified baseline hazard function. In the following discussion, we assume $\boldsymbol{X}$ is time-independent and drop the time factor t. Since the likelihood function

is difficult to work with owing to the non-parametric baseline hazard function, Lin and Ying (1994) considered a least-squares argument based on the counting process and martingale theory.

We introduce the usual notation for the counting process. For a sample of size n, suppose that we observe $(T_i, \Delta_i, \boldsymbol{X}_i)$, where T_i is the observed time, Δ_i is an event indicator with $\Delta_i = 1$ for failure and $\Delta_i = 0$ for right-censoring, and $\boldsymbol{X}_i$ is a vector of p-dimensional covariates for $i = 1, \ldots, n$. We denote $N_i(t) = \Delta_i I(T_i \leq t)$ and $Y_i(t) = I(T_i \geq t)$ as the observed event process and the at-risk process for the i-th subject, respectively. By the Doob–Meyer decomposition, the observed event process $N_i(t)$ can be uniquely decomposed that for every i and t,

$$N_i(t) = M_i(t) + \int_0^t Y_i(s)\,\mathrm{d}\Lambda(s \mid \boldsymbol{X}_i),$$

where $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$. Lin and Ying (1994) proposed to estimate the cumulative baseline hazard, $\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$, by solving $\sum_{i=1}^n \mathrm{d}M_i(t) = 0$. Thus, the estimator for the cumulative baseline hazard can be expressed as

$$\widehat{\Lambda}_0(\widehat{\boldsymbol{\beta}}, t) = \int_0^t \frac{\sum_{i=1}^n \{\mathrm{d}N_i(s) - Y_i(s)\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}_i\,\mathrm{d}s\}}{\sum_{i=1}^n Y_i(s)}.$$

To estimate $\boldsymbol{\beta}$, they introduced the following semi-parametric estimating function:

$$\sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t)\}\{\mathrm{d}N_i(t) - Y_i(t)\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i\,\mathrm{d}t\}, \tag{3.1}$$

where τ is the maximum follow-up time and $\bar{\boldsymbol{X}}(t) = \sum_{i=1}^n \boldsymbol{X}_i Y_i(t) / \sum_{i=1}^n Y_i(t)$ is the average covariate vector for the subjects at risk at time t with the convention that $0/0 = 0$. Solving (3.1) equals to zero yields the following closed-form solution for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t)\}^{\otimes 2}\,\mathrm{d}t\right]^{-1} \sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t)\}\,\mathrm{d}N_i(t),$$

where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$. With the standard martingale theory, Lin and Ying (1994) have shown that $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with a covariance matrix that does not depend on $\boldsymbol{\beta}$. Through some algebraic manipulation, we can write (3.1) as $\boldsymbol{b} - \boldsymbol{V}\boldsymbol{\beta}$ with

$$\boldsymbol{b} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t)\}\,\mathrm{d}N_i(t),$$
$$\boldsymbol{V} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t)\}^{\otimes 2}\,\mathrm{d}t.$$

Since $\boldsymbol{V}$ is positive semidefinite, integrating $-(\boldsymbol{b} - \boldsymbol{V}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ leads to a least-squares-type loss function given by $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\beta}/2 - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\beta}$.

3.2.2 Varying-coefficient additive hazards model

To allow the covariate effects of $\boldsymbol{X}$ to be modified by another predictor, we study a varying-coefficient additive hazard model. We introduce a set of q-dimensional covariates $\boldsymbol{U}$ into the model, which characterizes the effect of $\boldsymbol{X}$ on the hazard rate. Also, we consider a set of r-dimensional covariates $\boldsymbol{Z}$ that is assumed to have a constant effect $\boldsymbol{\alpha}$ on the hazard rate. The covariates $\boldsymbol{U}$ and $\boldsymbol{Z}$ may overlap. In genomic studies, $\boldsymbol{X}$ can be a set of gene expressions, $\boldsymbol{U}$ can be gene mutations, and $\boldsymbol{Z}$ can be clinical factors. We consider the following partial linear varying-coefficient additive hazards model:

$$\lambda(t \mid \boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}(\boldsymbol{U})^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z},$$

where $\boldsymbol{\beta}(\boldsymbol{U}) \equiv (\beta_1(\boldsymbol{U}), \ldots, \beta_p(\boldsymbol{U}))^{\mathrm{T}}$ is a vector of unknown coefficient functions that allows the effect of each component of $\boldsymbol{X}$ to vary with $\boldsymbol{U}$. In the following discussion, we introduce a naive kernel estimation method and a proposed method to estimate the covariate effects.

Naive kernel estimation method

A naive approach to estimate $\boldsymbol{\beta}(\cdot)$ and $\boldsymbol{\alpha}$ is to treat $\boldsymbol{\alpha}$ as varying and then combine the estimates by taking weighted average. Let $\boldsymbol{W} \equiv (\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\theta}(\cdot) \equiv (\boldsymbol{\beta}(\cdot)^{\mathrm{T}}, \boldsymbol{\alpha}(\cdot)^{\mathrm{T}})^{\mathrm{T}}$. For any vector $\boldsymbol{u}$ in the support of $\boldsymbol{U}$, the varying covariate effect $\boldsymbol{\theta}(\boldsymbol{u})$ can be estimated by minimizing the following kernel-weighted loss function:

$$\ell\{\boldsymbol{\theta}(\boldsymbol{u})\} = \frac{1}{2}\boldsymbol{\theta}(\boldsymbol{u})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{u})\boldsymbol{\theta}(\boldsymbol{u}) - \boldsymbol{b}(\boldsymbol{u})^{\mathrm{T}}\boldsymbol{\theta}(\boldsymbol{u}),$$

with

$$\begin{aligned}
\boldsymbol{b}(\boldsymbol{u}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{W}_i - \bar{\boldsymbol{W}}(\boldsymbol{u},t)\}K_{\boldsymbol{H}}(\boldsymbol{U}_i - \boldsymbol{u})\,\mathrm{d}N_i(t),\\
\boldsymbol{V}(\boldsymbol{u}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\{\boldsymbol{W}_i - \bar{\boldsymbol{W}}(\boldsymbol{u},t)\}^{\otimes 2}K_{\boldsymbol{H}}(\boldsymbol{U}_i - \boldsymbol{u})\,\mathrm{d}t,\\
\bar{\boldsymbol{W}}(\boldsymbol{u},t) &= \frac{\sum_{i=1}^{n} Y_i(t)\boldsymbol{W}_i K_{\boldsymbol{H}}(\boldsymbol{U}_i - \boldsymbol{u})}{\sum_{i=1}^{n} Y_i(t)K_{\boldsymbol{H}}(\boldsymbol{U}_i - \boldsymbol{u})},
\end{aligned}$$

where $K_{\boldsymbol{H}}(\boldsymbol{x}) \equiv |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}\boldsymbol{x})$ is a kernel function with bandwidth matrix $\boldsymbol{H}$. The estimator of $\boldsymbol{\theta}$ evaluated at $\boldsymbol{u}$ is given by $\widehat{\boldsymbol{\theta}}(\boldsymbol{u}) = \boldsymbol{V}(\boldsymbol{u})^{-1}\boldsymbol{b}(\boldsymbol{u})$.

A possible way to estimate $\boldsymbol{\alpha}$ is to compute the weighted average of $\widehat{\boldsymbol{\alpha}}(\boldsymbol{u})$ over the support of $\boldsymbol{U}$ (Yin et al., 2008). After obtaining $\widehat{\boldsymbol{\theta}}(\boldsymbol{u})$ at different values of $\boldsymbol{u}$, we can form a global estimate of $\widehat{\boldsymbol{\alpha}}$ given by

$$\widehat{\boldsymbol{\alpha}}_w = \int_{\mathcal{U}} \boldsymbol{w}(\boldsymbol{v})\widehat{\boldsymbol{\alpha}}(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v},$$

where $\mathcal{U}$ denotes the support of $\boldsymbol{U}$ and $\boldsymbol{w}(\boldsymbol{v})$ is a weight matrix satisfying $\int_{\mathcal{U}} \boldsymbol{w}(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = \boldsymbol{I}_{r\times r}$ as an identity matrix. A natural choice of the weight matrix is the standardized inverse covariance matrix of $\widehat{\boldsymbol{\alpha}}(\boldsymbol{u})$, that is, $\boldsymbol{w}(\boldsymbol{u}) = \{\int_{\mathcal{U}} \boldsymbol{J}(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}\}^{-1}\boldsymbol{J}(\boldsymbol{u})$, where $\boldsymbol{J}(\boldsymbol{u})$ is the inverse of the sub-matrix of the asymptotic covariance matrix that corresponding to

$\boldsymbol{\alpha}$.

In fact, the naive kernel estimation method is based on an assumption that the baseline hazard function is a function of time t and effect modifier $\boldsymbol{U}$. Consider a special case that the bandwidth is a matrix with all zeros, then the estimation of $\boldsymbol{\beta}(\boldsymbol{u})$ will only based on the subjects taken at a target point $\boldsymbol{u}$. In this case, the hazard rate at $\boldsymbol{u}$ is estimated individually without borrowing the information from the neighborhood of $\boldsymbol{u}$. This naive method will probably leads to loss of efficiency as it does not utilize the information that the λ_0 is the same over different values of $\boldsymbol{U}$.

Proposed method

To motivate an estimation method that accounts for the fact that λ_0 is the same over different values of $\boldsymbol{U}$, we consider a modification of the kernel estimation method based on a block-wise formulation. We first consider the model without constant covariate effect for simple illustration. Suppose that $\boldsymbol{U}$ is discrete and the subjects can be stratified into K non-overlapping groups based on the values of $\boldsymbol{U}$, so that we can partition $\boldsymbol{X}$ into $\big(\boldsymbol{X}^{(1)\mathrm{T}}, \ldots, \boldsymbol{X}^{(K)\mathrm{T}}\big)^{\mathrm{T}}$ and there are n_k subjects in the k-th block for $k = 1, \ldots, K$. As it was assumed that the effects of $\boldsymbol{X}$ vary over different values of $\boldsymbol{U}$, we consider a block-wise formulation to separate the effects across strata. Let $\boldsymbol{B} \equiv \mathrm{diag}\big(\boldsymbol{X}^{(1)\mathrm{T}}, \ldots, \boldsymbol{X}^{(K)\mathrm{T}}\big)$ denote a block diagonal matrix and $\boldsymbol{\beta}^{(k)}$ denote the covariate effect corresponding to the subjects in the k-th block. The covariate effects over the K blocks, $(\boldsymbol{\beta}^{(1)\mathrm{T}}, \ldots, \boldsymbol{\beta}^{(K)\mathrm{T}})^{\mathrm{T}}$, can be estimated using the estimation procedures for Lin and Ying's (1994) additive hazards model by replacing $\boldsymbol{X}$ with $\boldsymbol{B}$.

Let $N_i^{(k)}(t)$ and $Y_i^{(k)}(t)$ denote the event process and at-risk process for the i-th subject in the k-th block at time t, respectively. We can rewrite the estimating function in the following block-wise expression:

$$\bar{\boldsymbol{B}}(t) = \frac{\sum_{i=1}^{n} Y_i(t)\boldsymbol{B}_i}{\sum_{i=1}^{n} Y_i(t)}$$

$$= \left(\frac{\sum_{i=1}^{n_1} Y_i^{(1)}(t) \boldsymbol{X}_i^{(1)\mathrm{T}}}{\sum_{i=1}^{n} Y_i(t)}, \ldots, \frac{\sum_{i=1}^{n_K} Y_i^{(K)}(t) \boldsymbol{X}_i^{(K)\mathrm{T}}}{\sum_{i=1}^{n} Y_i(t)} \right)^{\mathrm{T}}$$
$$= \left(\bar{\boldsymbol{X}}^{(1)}(t)^{\mathrm{T}}, \ldots, \bar{\boldsymbol{X}}^{(K)}(t)^{\mathrm{T}} \right)^{\mathrm{T}}.$$

Hence, $\boldsymbol{b} \equiv n^{-1} \sum_{i=1}^{n} \int_0^\tau \{\boldsymbol{B}_i - \bar{\boldsymbol{B}}(t)\} \, \mathrm{d}N_i(t)$ is a vector with the k-th element given by

$$\frac{1}{n} \sum_{i=1}^{n_k} \int_0^\tau \boldsymbol{X}_i^{(k)} \, \mathrm{d}N_i^{(k)}(t) - \frac{1}{n} \sum_{j=1}^{K} \sum_{i=1}^{n_j} \int_0^\tau \bar{\boldsymbol{X}}^{(k)}(t) \, \mathrm{d}N_i^{(j)}(t),$$

and $\boldsymbol{V} \equiv n^{-1} \sum_{i=1}^{n} \int_0^\tau Y_i(t) \{\boldsymbol{B}_i - \bar{\boldsymbol{B}}(t)\}^{\otimes 2} \, \mathrm{d}t$ is a block matrix with the k-th diagonal block given by

$$\frac{1}{n} \int_0^\tau \left\{ \sum_{i=1}^{n_k} Y_i^{(k)}(t) \boldsymbol{X}_i^{(k)} \boldsymbol{X}_i^{(k)\mathrm{T}} \right\} \mathrm{d}t - \frac{1}{n} \int_0^\tau \frac{1}{\sum_{i=1}^{n} Y_i(t)} \left\{ \sum_{i=1}^{n_k} Y_i^{(k)}(t) \boldsymbol{X}_i^{(k)} \right\}^{\otimes 2} \mathrm{d}t$$

and the (k, k')-th off-diagonal element ($k \neq k'$) given by

$$-\frac{1}{n} \int_0^\tau \frac{1}{\sum_{i=1}^{n} Y_i(t)} \left\{ \sum_{i=1}^{n_k} Y_i^{(k)}(t) \boldsymbol{X}_i^{(k)} \right\} \left\{ \sum_{i=1}^{n_{k'}} Y_i^{(k')}(t) \boldsymbol{X}_i^{(k')\mathrm{T}} \right\} \mathrm{d}t.$$

In contrast to building separate model for each block using the naive method, estimating the covariate effects among finite number of groups in a block-wise manner allows us to take into account the fact that λ_0 is the same among subjects regardless of their values of $\boldsymbol{U}$.

To accommodate continuous $\boldsymbol{U}$, we mimic this block-wise formulation and consider a set of modified kernel-weighted functions. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ be a set of grid points. The covariate effects evaluated at the grid points, $\boldsymbol{\beta}(\boldsymbol{u}_1), \ldots, \boldsymbol{\beta}(\boldsymbol{u}_M)$, can be estimated simultaneously by $\widetilde{\boldsymbol{V}}^{-1}\widetilde{\boldsymbol{b}}$, where $\widetilde{\boldsymbol{V}}$ is a block matrix with $\boldsymbol{V}(\boldsymbol{u}_m, \boldsymbol{u}_m)$ as the m-th diagonal block and $\boldsymbol{V}(\boldsymbol{u}_m, \boldsymbol{u}_{m'})$ as the (m, m')-th off-diagonal block and $\widetilde{\boldsymbol{b}} \equiv (\boldsymbol{b}(\boldsymbol{u}_1)^{\mathrm{T}}, \ldots, \boldsymbol{b}(\boldsymbol{u}_M)^{\mathrm{T}})^{\mathrm{T}}$

with

$$
\begin{aligned}
\boldsymbol{b}(\boldsymbol{u}_m) =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\frac{\boldsymbol{X}_i K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_m)}{\widehat{n}_1}\,\mathrm{d}N_i(t) \\
& -\frac{1}{n}\int_0^{\tau}\frac{1}{\sum_{i=1}^{n}Y_i(t)}\frac{\sum_{i=1}^{n}Y_i(t)\boldsymbol{X}_i K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_m)}{\widehat{n}_1}\sum_{i=1}^{n}\mathrm{d}N_i(t), \\
\boldsymbol{V}(\boldsymbol{u}_m,\boldsymbol{u}_m) =& \frac{1}{n}\int_0^{\tau}\sum_{i=1}^{n}\frac{Y_i(t)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_m)}{\widehat{n}_1}\,\mathrm{d}t \\
& -\widehat{p}_1\int_0^{\tau}\frac{1}{\sum_{i=1}^{n}Y_i(t)}\left\{\frac{\sum_{i=1}^{n}Y_i(t)\boldsymbol{X}_i K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_m)}{\widehat{n}_1}\right\}^{\otimes 2}\mathrm{d}t, \\
\boldsymbol{V}(\boldsymbol{u}_m,\boldsymbol{u}_{m'}) =& -\widehat{p}_1\int_0^{\tau}\frac{1}{\sum_{i=1}^{n}Y_i(t)}\left\{\frac{\sum_{i=1}^{n}Y_i(t)\boldsymbol{X}_i K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_m)}{\widehat{n}_1}\right\} \\
& \times\left\{\frac{\sum_{i=1}^{n}Y_i(t)\boldsymbol{X}_i^{\mathrm{T}}K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_{m'})}{\widehat{n}_1}\right\}\mathrm{d}t, \\
\widehat{p}_1 =& \frac{\sum_{i=1}^{n}K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_1)}{\sum_{m=1}^{M}\sum_{i=1}^{n}K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_m)}, \\
\widehat{n}_1 =& \sum_{i=1}^{n}K_{\boldsymbol{H}}(\boldsymbol{U}_i-\boldsymbol{u}_1).
\end{aligned}
$$

The term $\widehat{n}_1$ can be viewed as an estimate of the aggregated weight indicating the contribution of subjects over a neighborhood of $\boldsymbol{u}_1$. If all the components in $\boldsymbol{H}$ are set to be 0, then this formulation reduces to the estimating function under the block-wise formulation.

In analogy with the kernel-weighted functions motivated by the block-wise formulation for estimating the varying covariate effect, we propose to estimate $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}(\boldsymbol{u}_1)^{\mathrm{T}},\ldots,\boldsymbol{\beta}(\boldsymbol{u}_M)^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}$ simultaneously by

$$
\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{V}} & \widetilde{\boldsymbol{V}}_{\boldsymbol{\alpha}} \\ \widetilde{\boldsymbol{V}}_{\boldsymbol{\alpha}}^{T} & \boldsymbol{V}_{\boldsymbol{\alpha},\boldsymbol{\alpha}} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\boldsymbol{b}} \\ \boldsymbol{b}_{\boldsymbol{\alpha}} \end{pmatrix},
$$

where $\widetilde{\boldsymbol{V}}_{\boldsymbol{\alpha}} = (\boldsymbol{V}_{\boldsymbol{\alpha}}(\boldsymbol{u}_1)^{\mathrm{T}}, \ldots, \boldsymbol{V}_{\boldsymbol{\alpha}}(\boldsymbol{u}_M)^{\mathrm{T}})^{\mathrm{T}}$ with

$$
\begin{aligned}
\boldsymbol{b}_{\boldsymbol{\alpha}} =& \frac{1}{n^2}\frac{1}{\widehat{p}_1}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)\}\,\mathrm{d}N_i(t),\\
\boldsymbol{V}_{\boldsymbol{\alpha},\boldsymbol{\alpha}} =& \frac{1}{n^2}\frac{1}{\widehat{p}_1}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\{\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}\,\mathrm{d}t,\\
\boldsymbol{V}_{\boldsymbol{\alpha}}(\boldsymbol{u}_m) =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\Bigg[\frac{\{\boldsymbol{X}_i\boldsymbol{Z}_i^{\mathrm{T}} - \boldsymbol{X}_i\bar{\boldsymbol{Z}}(t)^{\mathrm{T}}\}K_{\boldsymbol{H}}(\boldsymbol{U}_i - \boldsymbol{u}_m)}{\widehat{n}_1} - \bar{\boldsymbol{X}}(\boldsymbol{u}_m, t)\boldsymbol{Z}_i^{\mathrm{T}}\\
&+ \bar{\boldsymbol{X}}(\boldsymbol{u}_m, t)\bar{\boldsymbol{Z}}(t)^{\mathrm{T}}\Bigg]\,\mathrm{d}t,\\
\bar{\boldsymbol{X}}(\boldsymbol{u}_m, t) =& \frac{\sum_{i=1}^{n}\boldsymbol{X}_i(t)Y_i(t)K_{\boldsymbol{H}}(\boldsymbol{U}_i - \boldsymbol{u}_m)}{\widehat{n}_1\sum_{i=1}^{n}Y_i(t)},\\
\bar{\boldsymbol{Z}}(t) =& \frac{\sum_{i=1}^{n}Y_i(t)\boldsymbol{Z}_i}{\sum_{i=1}^{n}Y_i(t)}.
\end{aligned}
$$

3.3 Simulation studies

We assessed the performance of the proposed method in estimating the constant and varying covariate effects. For simplicity, we considered a univariate $\boldsymbol{X}$, and its effect on the outcome is modified by a univariate $\boldsymbol{U}$ with $q = 1$ and bivariate $\boldsymbol{U}$ with $q = 2$, respectively. The components of X and $\boldsymbol{U}$ were generated independently from a uniform distribution on $(0, 1)$. Also, we considered a univariate linear predictor $\boldsymbol{Z}$ that was generated from a Bernoulli distribution with probability 0.5. We used the following two examples to illustrate the performance of our method:

Model 1: $\beta(\boldsymbol{u}) = 3\bar{u}^3$,

Model 2: $\beta(\boldsymbol{u}) = 3\{1 - \sin(\bar{u}\pi)\}$,

where $\bar{u}$ is the sample mean of the components of $\boldsymbol{u}$. The varying-coefficient functions $\beta(\boldsymbol{u})$ under Model 1 and 2 are plotted in Figure 3.1. In both models, we set α to be 0.5. In the following simulations, we set $\lambda_0(t) = t$, and the censoring time follows an exponential distribution with the mean chosen to yield a censoring rate of about 30%. In

each setting, we considered sample sizes $n = 200$, 500, and 1000.

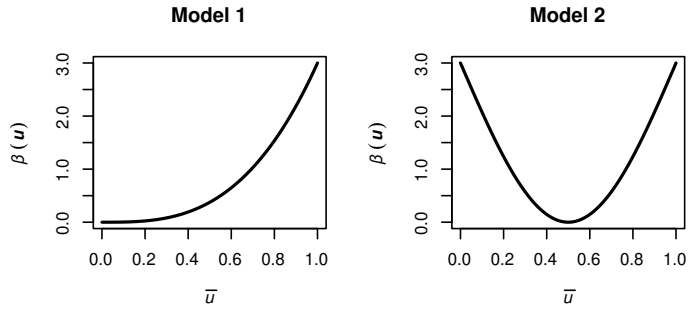

Figure 3.1: Varying-coefficient functions under Model 1 and 2.

In addition to the proposed method, we considered two alternative methods: the regression model on linear predictors X and Z, and the varying-coefficient model, on which the naive kernel estimation method is adopted. For the kernel-weighted methods, we adopted a Gaussian kernel function given by $K(\boldsymbol{x}) = (2\pi)^{-q/2} \exp\{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/2\}$. Following Silverman (1986), the bandwidth estimator was approximated by $\boldsymbol{H}^{1/2} = \mathrm{diag}(h_1, \ldots, h_q)$ with $h_j = \widehat{\sigma}_j \left[4/\{n(q+2)\}\right]^{1/(q+4)}$ and $\widehat{\sigma}_j$ is the empirical standard deviation of $\boldsymbol{U}_{\cdot j}$, for $j = 1, \ldots, q$.

We assess the prediction performance of each method using an independent test sample of size 1000. For the naive kernel estimation method, we adopted a two-step estimation procedure to obtain the estimates for α and $\beta(\cdot)$. We first obtain the estimates $\widehat{\alpha}$ over the support of $\boldsymbol{U}$ and compute the weighted average of $\widehat{\alpha}_w$ using the training data. Given the information of $\widehat{\alpha}_w Z$, we can then estimate the varying covariate effect $\beta(\boldsymbol{u})$ on the validation data. For the proposed method, we adopted the linear interpolation to obtain the estimates for the points off the grid using the local estimates $\beta(\boldsymbol{u}_1), \ldots, \beta(\boldsymbol{u}_M)$. For a univariate $\boldsymbol{U}$, we obtained the local estimate $\widehat{\beta}$ at $u = \{0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1\}$ and used linear interpolation to obtain the estimates for the points off the grid. Similarly, for a bivariate $\boldsymbol{U}$, we obtained the

local estimate $\widehat{\beta}$ by considering $\{0, 0.25, 0.5, 0.75, 1\}$ on each dimension of $\boldsymbol{u}$, and we used bilinear interpolation to obtain the estimates for the points off the grid.

The prediction performance of each method is assessed using the following measures. We report the mean-squared error (MSE), defined as $\mathrm{E}[\{\widehat{\beta}(\boldsymbol{U}) - \beta_0(\boldsymbol{U})\}X + (\widehat{\alpha} - \alpha_0)Z]^2$, where $\widehat{\beta}(\boldsymbol{U})$ and $\beta_0(\boldsymbol{U})$ denote the estimated and true values of β evaluated at $\boldsymbol{U}$, respectively; $\widehat{\alpha}$ and α_0 denote the estimated and true values of α, respectively. To evaluate the ability of $\widehat{\beta}(\cdot)$ and $\widehat{\alpha}$ in discriminating individual risk, we compute the C-index (Harrell et al., 1982) by comparing $\eta_i = \widehat{\beta}(\boldsymbol{U}_i)X_i + \widehat{\alpha}Z_i$ for $i = 1, \ldots, 1000$. Also, we report the biases and standard deviations of the estimates of α and the estimates of β at some grids. The mean of each measure over 100 replicates under different settings are summarized in Tables 3.1–3.6.

Table 3.1: Simulation results under Model 1.

		$n = 200$	$n = 500$	$n = 1000$
		$q = 1$		
MSE	Constant	0.326	0.305	0.287
	Varying	0.157	0.074	0.042
	Proposed	0.086	0.045	0.022
C-index	Constant	0.575	0.578	0.579
	Varying	0.614	0.621	0.624
	Proposed	0.623	0.626	0.628
		$q = 2$		
MSE	Constant	0.166	0.125	0.115
	Varying	0.332	0.139	0.087
	Proposed	0.129	0.044	0.027
C-index	Constant	0.593	0.598	0.600
	Varying	0.581	0.601	0.610
	Proposed	0.611	0.623	0.627

NOTE: "Constant" represents regression model on linear predictors X and Z; "Varying" represents the varying-coefficient model where the naive kernel estimation method is adopted; "Proposed" represents the varying-coefficient model where the modified kernel estimation method is adopted.

Table 3.2: Biases under Model 1 with $q = 1$ (Standard deviations in parentheses).

	u	Constant	Varying	Proposed
		$n = 200$		
α	-	0.032 (0.188)	0.011 (0.182)	0.026 (0.183)
$\beta(u)$	0.2	0.481 (0.302)	−0.005 (0.417)	0.036 (0.320)
	0.4	0.313 (0.302)	0.060 (0.441)	0.053 (0.316)
	0.6	−0.143 (0.302)	−0.038 (0.480)	0.027 (0.394)
	0.8	−1.031 (0.302)	−0.004 (0.702)	−0.013 (0.479)
		$n = 500$		
α	-	0.001 (0.119)	−0.004 (0.115)	−0.002 (0.122)
$\beta(u)$	0.2	0.464 (0.220)	0.047 (0.300)	0.029 (0.234)
	0.4	0.296 (0.220)	−0.023 (0.323)	0.031 (0.238)
	0.6	−0.160 (0.220)	0.009 (0.344)	0.020 (0.242)
	0.8	−1.048 (0.220)	−0.023 (0.482)	−0.017 (0.300)
		$n = 1000$		
α	-	0.007 (0.075)	0.004 (0.076)	0.007 (0.075)
$\beta(u)$	0.2	0.476 (0.152)	0.028 (0.223)	0.013 (0.168)
	0.4	0.308 (0.152)	0.046 (0.249)	0.040 (0.170)
	0.6	−0.148 (0.152)	0.012 (0.268)	0.051 (0.206)
	0.8	−1.036 (0.152)	0.009 (0.360)	0.007 (0.222)

NOTE: See NOTE to Table 3.1.

Table 3.3: Biases under Model 1 with $q = 2$ (Standard deviations in parentheses).

	$\boldsymbol{u}$	Constant	Varying	Proposed
		$n = 200$		
α	-	-0.007 (0.187)	-0.033 (0.179)	0.008 (0.183)
$\beta(\boldsymbol{u})$	(0.25,0.25)	0.460 (0.349)	0.137 (0.720)	0.182 (0.485)
	(0.25,0.75)	0.132 (0.349)	-0.033 (0.832)	0.127 (0.508)
	(0.75,0.25)	0.132 (0.349)	0.092 (0.723)	0.182 (0.506)
	(0.75,0.75)	-0.759 (0.349)	0.121 (0.969)	0.084 (0.612)
		$n = 500$		
α	-	-0.004 (0.112)	-0.023 (0.112)	0.003 (0.114)
$\beta(\boldsymbol{u})$	(0.25,0.25)	0.376 (0.205)	-0.041 (0.449)	0.046 (0.251)
	(0.25,0.75)	0.048 (0.205)	0.079 (0.520)	0.046 (0.336)
	(0.75,0.25)	0.048 (0.205)	-0.042 (0.511)	0.077 (0.303)
	(0.75,0.75)	-0.842 (0.205)	-0.047 (0.695)	-0.006 (0.419)
		$n = 1000$		
α	-	0.012 (0.072)	0 (0.073)	0.015 (0.072)
$\beta(\boldsymbol{u})$	(0.25,0.25)	0.379 (0.125)	0.041 (0.374)	0.019 (0.201)
	(0.25,0.75)	0.051 (0.125)	-0.040 (0.412)	0.035 (0.232)
	(0.75,0.25)	0.051 (0.125)	-0.060 (0.382)	0.008 (0.247)
	(0.75,0.75)	-0.839 (0.125)	0.065 (0.482)	0.082 (0.321)

NOTE: See NOTE to Table 3.1.

Table 3.4: Simulation results under Model 2.

		$n = 200$	$n = 500$	$n = 1000$
		$q = 1$		
MSE	Constant	0.419	0.391	0.374
	Varying	0.214	0.106	0.064
	Proposed	0.130	0.077	0.044
C-index	Constant	0.588	0.589	0.589
	Varying	0.624	0.636	0.640
	Proposed	0.636	0.642	0.643
		$q = 2$		
MSE	Constant	0.202	0.172	0.163
	Varying	0.378	0.154	0.094
	Proposed	0.141	0.052	0.032
C-index	Constant	0.599	0.603	0.605
	Varying	0.584	0.600	0.612
	Proposed	0.611	0.621	0.626

NOTE: See NOTE to Table 3.1.

Table 3.5: Biases under Model 2 with $q = 1$ (Standard deviations in parentheses).

	u	Constant	Varying	Proposed
		$n = 200$		
α	-	0.025 (0.204)	0.005 (0.199)	0.021 (0.198)
$\beta(u)$	0.2	−0.463 (0.330)	−0.221 (0.603)	−0.185 (0.415)
	0.4	0.627 (0.330)	0.149 (0.480)	0.097 (0.340)
	0.6	0.627 (0.330)	0.063 (0.430)	0.071 (0.374)
	0.8	−0.463 (0.330)	−0.101 (0.616)	−0.148 (0.434)
		$n = 500$		
α	-	−0.002 (0.130)	−0.007 (0.121)	−0.001 (0.131)
$\beta(u)$	0.2	−0.463 (0.225)	−0.052 (0.446)	−0.093 (0.316)
	0.4	0.627 (0.225)	0.028 (0.328)	0.059 (0.253)
	0.6	0.627 (0.225)	0.069 (0.299)	0.052 (0.233)
	0.8	−0.463 (0.225)	−0.097 (0.438)	−0.118 (0.279)
		$n = 1000$		
α	-	0.005 (0.088)	0.003 (0.084)	0.007 (0.088)
$\beta(u)$	0.2	−0.455 (0.173)	−0.028 (0.335)	−0.055 (0.246)
	0.4	0.634 (0.173)	0.081 (0.256)	0.079 (0.184)
	0.6	0.634 (0.173)	0.050 (0.234)	0.079 (0.191)
	0.8	−0.455 (0.173)	−0.043 (0.331)	−0.057 (0.218)

NOTE: See NOTE to Table 3.1.

Table 3.6: Biases under Model 2 with $q = 2$ (Standard deviations in parentheses).

	$\boldsymbol{u}$	Constant	Varying	Proposed
		$n = 200$		
α	-	−0.004 (0.180)	−0.024 (0.186)	0.010 (0.179)
$\beta(\boldsymbol{u})$	(0.25,0.25)	−0.397 (0.315)	0.113 (0.962)	0.156 (0.657)
	(0.25,0.75)	0.482 (0.315)	0.041 (0.756)	0.182 (0.446)
	(0.75,0.25)	0.482 (0.315)	0.185 (0.658)	0.228 (0.434)
	(0.75,0.75)	−0.397 (0.315)	0.073 (0.852)	0.047 (0.534)
		$n = 500$		
α	-	−0.004 (0.113)	−0.024 (0.110)	0.003 (0.116)
$\boldsymbol{\beta}(\boldsymbol{u})$	(0.25,0.25)	−0.474 (0.192)	−0.076 (0.576)	0.031 (0.346)
	(0.25,0.75)	0.405 (0.192)	0.114 (0.468)	0.083 (0.301)
	(0.75,0.25)	0.405 (0.192)	0.021 (0.454)	0.117 (0.264)
	(0.75,0.75)	−0.474 (0.192)	−0.074 (0.621)	−0.031 (0.361)
		$n = 1000$		
α	-	0.007 (0.071)	−0.003 (0.071)	0.011 (0.069)
$\boldsymbol{\beta}(\boldsymbol{u})$	(0.25,0.25)	−0.479 (0.120)	0.025 (0.515)	0.006 (0.259)
	(0.25,0.75)	0.400 (0.120)	−0.008 (0.361)	0.067 (0.195)
	(0.75,0.25)	0.400 (0.120)	−0.005 (0.329)	0.049 (0.212)
	(0.75,0.75)	−0.479 (0.120)	0.033 (0.441)	0.047 (0.287)

NOTE: See NOTE to Table 3.1.

Overall, the proposed method yields higher prediction accuracy than the alternatives. In all simulation scenarios, the proposed method has indicated desirable prediction performance for the lowest MSE and highest C-index values. The proposed method outperforms the naive method, possibly because the proposed method can account for the fact that all subjects share the same baseline hazard function. Also, it can be seen that the proposed method yields much higher predictive power than the naive method when the sample size is small. This demonstrates the importance of utilizing the information of the baseline hazard function from the neighborhoods in parameter estimation especially when sample size is limited. The biases under the two kernel-based estimation methods are similar, but the standard deviations of the estimates under the proposed method are less than those under the naive method.

3.4 Discussion and future work

In this chapter, we considered a varying-coefficient additive hazards model to accommodate the interaction effects between two sets of predictors. We investigated two kernel-based methods for estimating the varying covariate effects. We first considered an intuitive way that simply multiplies a kernel weight to each term associated with an observation before computing the weighted average. Thus, both the varying covariate effects and the baseline hazards function evaluated at a target point are estimated in a local sense. This naive method suffers a loss of efficiency because it does not account for the fact that the underlying baseline hazard functions are the same among all subjects. To overcome this inadequacy, we proposed a modified kernel estimating function that utilizes the information of the baseline hazard function from all subjects rather than the observations near the target point. Preliminary results illustrated that the proposed method indeed yields better prediction accuracy than the naive method.

The proposed method can be applied to integrative genomic analyses by specifying each

component of $\boldsymbol{U}$ to represent the information of a data type, thereby accommodating the interaction effect of X_j with multiple data types for $j = 1, \ldots, p$. Since the (multivariate) kernel-based method allows each component of $\boldsymbol{U}$ to modify the covariate effect flexibly, it may accommodate the difference in scales among multiple data types.

There are several directions for future research. First, it is desirable to study the asymptotic properties of the proposed estimators under appropriate choices of bandwidth and grid points. In particular, we will examine the asymptotic properties for the case when the number of covariates p increases with the sample size n. The asymptotic properties can be developed using the counting process and martingale theories. Second, we will allow penalization on the constant and varying covariate effects to accommodate high-dimensional covariates. Third, we will study methods for identifying covariates whose effects on the hazard rate can be reasonably assumed to be constant. It is desirable to develop a method that identifies which feature should be fitted as a linear predictor.

3.5 Appendix: Additional simulations

To evaluate the sensitivity of the number of grids on the model fit, we conducted additional simulations to assess the performance of the the proposed method under different number of grids for a univariate $\boldsymbol{U}$. We considered to estimate $\widehat{\beta}(u_1), \ldots, \widehat{\beta}(u_M)$ simultaneously for $M = 5$, 9, and 13. The additional simulation results are summarized in Tables 3.7–3.10.

Under Model 1, where β is monotonically increasing for $0 \leq u \leq 1$, the MSE and C-index values are almost the same over different choices of grids. When we increase the number of grid points, the biases reduce in nearly all settings as more interpolation grids can generally capture the curvilinear structure of the true function and improve model fit. Nevertheless, the improvement in prediction accuracy is not substantial under Model 1. If the true function is steadily increasing or decreasing, then a small number of grid

Table 3.7: Additional simulation results under Model 1 with $q = 1$.

		$n = 200$	$n = 500$	$n = 1000$
MSE	Constant	0.326	0.305	0.287
	Varying	0.157	0.074	0.042
	Proposed ($M = 5$)	0.089	0.046	0.023
	Proposed ($M = 9$)	0.086	0.045	0.022
	Proposed ($M = 13$)	0.087	0.045	0.022
C-index	Constant	0.575	0.578	0.579
	Varying	0.614	0.621	0.624
	Proposed ($M = 5$)	0.624	0.627	0.628
	Proposed ($M = 9$)	0.623	0.626	0.628
	Proposed ($M = 13$)	0.622	0.626	0.627

NOTE: See NOTE to Table 3.1.

Table 3.8: Additional simulation results: Biases under Model 1 with $q = 1$ (Standard deviations in parentheses).

	u	Constant	Varying	Proposed		
				$M = 5$	$M = 9$	$M = 13$
			$n = 200$			
α	-	0.032 (0.188)	0.011 (0.182)	0.027 (0.183)	0.026 (0.183)	0.025 (0.188)
$\beta(u)$	0.2	0.481 (0.302)	−0.005 (0.417)	0.059 (0.323)	0.036 (0.320)	0.020 (0.323)
	0.4	0.313 (0.302)	0.060 (0.441)	0.096 (0.313)	0.053 (0.316)	0.040 (0.321)
	0.6	−0.143 (0.302)	−0.038 (0.480)	0.122 (0.356)	0.027 (0.394)	0.010 (0.402)
	0.8	−1.031 (0.302)	−0.004 (0.702)	−0.019 (0.476)	−0.013 (0.479)	−0.023 (0.488)
			$n = 500$			
α	-	0.001 (0.119)	−0.004 (0.115)	−0.003 (0.122)	−0.002 (0.122)	−0.002 (0.122)
$\beta(u)$	0.2	0.464 (0.220)	0.047 (0.300)	0.045 (0.242)	0.029 (0.234)	0.017 (0.236)
	0.4	0.296 (0.220)	−0.023 (0.323)	0.074 (0.231)	0.031 (0.238)	0.017 (0.242)
	0.6	−0.160 (0.220)	0.009 (0.344)	0.099 (0.234)	0.020 (0.242)	0.003 (0.246)
	0.8	−1.048 (0.220)	−0.023 (0.482)	−0.032 (0.287)	−0.017 (0.300)	−0.025 (0.310)
			$n = 1000$			
α	-	0.007 (0.075)	0.004 (0.076)	0.006 (0.076)	0.007 (0.075)	0.007 (0.075)
$\beta(u)$	0.2	0.476 (0.152)	0.028 (0.223)	0.034 (0.168)	0.013 (0.168)	0.001 (0.173)
	0.4	0.308 (0.152)	0.046 (0.249)	0.092 (0.169)	0.040 (0.170)	0.027 (0.172)
	0.6	−0.148 (0.152)	0.012 (0.268)	0.123 (0.187)	0.051 (0.206)	0.039 (0.211)
	0.8	−1.036 (0.152)	0.009 (0.360)	−0.009 (0.209)	0.007 (0.222)	−0.004 (0.236)

NOTE: See NOTE to Table 3.1.

Table 3.9: Additional simulation results under Model 2 with $q = 1$.

		$n = 200$	$n = 500$	$n = 1000$
MSE	Constant	0.419	0.391	0.374
	Varying	0.214	0.106	0.064
	Proposed ($M = 5$)	0.129	0.082	0.050
	Proposed ($M = 9$)	0.130	0.077	0.044
	Proposed ($M = 13$)	0.136	0.081	0.046
C-index	Constant	0.588	0.589	0.589
	Varying	0.624	0.636	0.640
	Proposed ($M = 5$)	0.636	0.642	0.643
	Proposed ($M = 9$)	0.636	0.642	0.643
	Proposed ($M = 13$)	0.635	0.641	0.643

NOTE: See NOTE to Table 3.1.

Table 3.10: Additional simulation results: Biases under Model 2 with $q = 1$ (Standard deviations in parentheses).

	u	Constant	Varying	Proposed		
				$M = 5$	$M = 9$	$M = 13$
			$n = 200$			
α	-	0.025 (0.204)	0.005 (0.199)	0.020 (0.198)	0.021 (0.198)	0.021 (0.198)
$\beta(u)$	0.2	−0.463 (0.330)	−0.221 (0.603)	−0.126 (0.399)	−0.185 (0.415)	−0.234 (0.428)
	0.4	0.627 (0.330)	0.149 (0.480)	0.271 (0.332)	0.097 (0.340)	0.047 (0.347)
	0.6	0.627 (0.330)	0.063 (0.430)	0.283 (0.340)	0.071 (0.374)	0.018 (0.382)
	0.8	−0.463 (0.330)	−0.101 (0.616)	−0.104 (0.433)	−0.148 (0.434)	−0.189 (0.444)
			$n = 500$			
α	-	−0.002 (0.130)	−0.007 (0.121)	−0.003 (0.130)	−0.001 (0.131)	−0.001 (0.130)
$\beta(u)$	0.2	−0.463 (0.225)	−0.052 (0.446)	−0.070 (0.325)	−0.093 (0.316)	−0.133 (0.324)
	0.4	0.627 (0.225)	0.028 (0.328)	0.254 (0.247)	0.059 (0.253)	0.008 (0.258)
	0.6	0.627 (0.225)	0.069 (0.299)	0.247 (0.228)	0.052 (0.233)	0.002 (0.238)
	0.8	−0.463 (0.225)	−0.097 (0.438)	−0.103 (0.270)	−0.118 (0.279)	−0.154 (0.289)
			$n = 1000$			
α	-	0.005 (0.088)	0.003 (0.084)	0.007 (0.091)	0.007 (0.088)	0.007 (0.088)
$\beta(u)$	0.2	−0.455 (0.173)	−0.028 (0.335)	−0.043 (0.232)	−0.055 (0.246)	−0.091 (0.256)
	0.4	0.634 (0.173)	0.081 (0.256)	0.279 (0.188)	0.079 (0.184)	0.033 (0.185)
	0.6	0.634 (0.173)	0.050 (0.234)	0.272 (0.188)	0.079 (0.191)	0.035 (0.197)
	0.8	−0.455 (0.173)	−0.043 (0.331)	−0.052 (0.208)	−0.057 (0.218)	−0.092 (0.231)

NOTE: See NOTE to Table 3.1.

points may be sufficient to explain the variation in β and the prediction performance is not very sensitive to the number of grids. Under Model 2, where β is quadratic for $0 \leq u \leq 1$, there is a slight improvement in MSE and considerable improvement in bias when we increase the number of grids from 5 to 9. If the true function is fluctuating, then more grid points are required to capture the variation of β over different values of u. Since the computational costs increase with the number of grids, appropriate choices of grids is essential to balance the trade-off between prediction accuracy and computational efficiency.

Chapter 4

Conclusion

In this thesis, we have studied varying-coefficient models to characterize the interaction effects among features from different data types. In Chapter 2, we have proposed a penalized integrative regression framework based on the single-index varying-coefficient model to select clinically relevant genomic features. Our method allows flexible modeling of clinical-genomic interactions as well as inducing sparsity in the single-index varying-coefficient models. The proposed method is generic and can accommodate various types of outcome variables, such as continuous and censored outcomes. One potential limitation of this method is that the model suffers a lack of flexibility as we need to summarize the multivariate effect modifiers linearly into a unique index for all covariate effects.

In Chapter 3, we have studied a varying-coefficient additive hazards model to accommodate the interaction effects on the censored outcome. We have considered a more versatile approach to incorporate multivariate effect modifiers based on the kernel smoothing technique. This approach overcomes the limitation of restricting all covariate effects to be modified by a unique composition of effect modifiers. We have proposed a kernel-based estimation method to estimate the constant and varying covariate effects simultaneously. Further investigation is needed for penalized estimation methods to accommodate high-dimensional covariates that are frequently encountered in genomic studies.

References

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925.

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

Beaman, G. M., Dennison, S. R., Chatfield, L. K., and Phoenix, D. A. (2014). Reliability of HSP70 (HSPA) expression as a prognostic marker in glioma. *Molecular and Cellular Biochemistry*, 393:301–307.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017(7691937).

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models-a comparative study. *BMC Bioinformatics*, 10(413).

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2:369–380.

Chen, H., Huang, Q., Dong, J., Zhai, D.-Z., Wang, A.-D., and Lan, Q. (2008). Overexpression of CDC2/CyclinB1 in gliomas, and CDC2 depletion inhibits proliferation of human glioma cells in vitro and in vivo. *BMC Cancer*, 8(29).

Chen, K., Lin, H., and Zhou, Y. (2012). Efficient estimation for the Cox model with varying coefficients. *Biometrika*, 99:379–392.

Chen, M., Liu, X., Du, J., Wang, X.-J., and Xia, L. (2017). Differentiated regulation of immune-response related genes between LUAD and LUSC subtypes of lung cancers. *Oncotarget*, 8:133–144.

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

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

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352.

Daemen, A., Gevaert, O., and De Moor, B. (2007). Integration of clinical and microarray data with kernel methods. In *Proceedings of the IEEE Engineering in Medicine and Biology Society*, pages 5411–5415. IEEE.

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C., Machiels, J.-P., Haustermans, K., and De Moor, B. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(39).

Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., and Perou, C. M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Medical Genomics*, 4(3).

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

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B*, 65:57–80.

Feng, S. and Xue, L. (2013). Variable selection for single-index varying-coefficient model. *Frontiers of Mathematics in China*, 8:541–565.

Feng, S. and Xue, L. (2015). Model detection and estimation for single-index varying coefficient model. *Journal of Multivariate Analysis*, 139:227–244.

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

Guan, S. (2017). *Variable Selection in Varying Multi-index Coefficient Models with Applications to Gene-environmental Interactions*. PhD dissertation, Michigan State University.

Hardle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21:157–178.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of American Medical Association*, 247:2543–2546.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55:757–779.

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

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82.

Hole, D., Watt, G., Davey-Smith, G., Hart, C., Gillis, C., and Hawthorne, V. (1996). Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. *BMJ*, 313:711–715.

Huang, Z. (2012). Inference for nonparametric parts in single-index varying-coefficient model. *Communications in Statistics - Theory and Methods*, 41:1214–1227.

Innamaa, A., Jackson, L., Asher, V., Van Schalkwyk, G., Warren, A., Keightley, A., Hay, D., Bali, A., Sowter, H., and Khan, R. (2013). Expression and effects of modulation of the K2P potassium channels TREK-1 (KCNK2) and TREK-2 (KCNK10) in the normal human ovary and epithelial ovarian cancer. *Clinical and Translational Oncology*, 15:910–918.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated data*. Springer.

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

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635.

Landi, M. T., Dracheva, T., Rotunno, M., Figueroa, J. D., Liu, H., Dasgupta, A., Mann, F. E., Fukuoka, J., Hames, M., Bergen, A. W., et al. (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS One*, 3(2).

Li, L. (2006). Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22:466–471.

Li, W.-C., Xiong, Z.-Y., Huang, P.-Z., Liao, Y.-J., Li, Q.-X., Yao, Z.-C., Liao, Y.-D., Xu, S.-L., Zhou, H., Wang, Q.-L., et al. (2019). KCNK levels are prognostic and diagnostic markers for hepatocellular carcinoma. *Aging (Albany NY)*, 11:8169–8182.

Li, Y., Wang, F., Li, R., and Sun, Y. (2020). Semiparametric integrative interaction analysis for non-small-cell lung cancer. *Statistical Methods in Medical Research*, 29:2865–2880.

Lin, D. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81:61–71.

Lin, H., Tan, M. T., and Li, Y. (2016). A semiparametrically efficient estimator of single-index varying coefficient Cox proportional hazards models. *Statistica Sinica*, 26:779–807.

Lin, Y., Zhang, J., Cai, J., Liang, R., Chen, G., Qin, G., Han, X., Yuan, C., Liu, Z., Li, Y., et al. (2018). Systematic analysis of gene expression alteration and co-expression network of eukaryotic initiation factor 4A-3 in cancer. *Journal of Cancer*, 9:4568–4577.

Ma, S. and Huang, J. (2005). Lasso method for additive risk models with high dimensional covariates. Technical report, Department of Statistics and Actuarial Science, University of Iowa.

Ma, S. and Song, P. X.-K. (2015). Varying index coefficient models. *Journal of the American Statistical Association*, 110:341–356.

Martinussen, T., Scheike, T. H., and Skovgaard, I. M. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scandinavian Journal of Statistics*, 29:57–74.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81:501–514.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Annals of Applied Statistics*, 2:1013–1033.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135:370–384.

Nevins, J. R., Huang, E. S., Dressman, H., Pittman, J., Huang, A. T., and West, M. (2003). Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics*, 12:R153–R157.

Ni, Y., Stingo, F. C., Ha, M. J., Akbani, R., and Baladandayuthapani, V. (2019). Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, 114:48–60.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141:1362–1379.

Pittman, J., Huang, E., Dressman, H., Horng, C.-F., Cheng, S. H., Tsou, M.-H., Chen, C.-M., Bild, A., Iversen, E. S., Huang, A. T., et al. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences*, 101:8431–8436.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, 3:425–441.

Relli, V., Trerotola, M., Guerra, E., and Alberti, S. (2018). Distinct lung cancer subtypes associate to distinct drivers of tumor progression. *Oncotarget*, 9:35528–35540.

Seoane, J. A., Day, I. N., Gaunt, T. R., and Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, 30:838–845.

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M.-S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Giordano, T. J., Misek, D. E., et al. (2008). Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14:822–827.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC press.

Tian, L., Zucker, D., and Wei, L. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100:172–183.

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

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

Tsiatis, A. A. et al. (1981). A large sample study of Cox's regression model. *The Annals of Statistics*, 9:93–108.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29:149–159.

Wong, K. Y., Fan, C., Tanioka, M., Parker, J. S., Nobel, A. B., Zeng, D., Lin, D.-Y., and Perou, C. M. (2019). I-Boost: an integrative boosting approach for predicting survival time with multiple genomics platforms. *Genome Biology*, 20(52).

Xu, J., Jiang, N., Shi, H., Zhao, S., Yao, S., and Shen, H. (2017). miR-28-5p promotes the development and progression of ovarian cancer through inhibition of N4BP1. *International Journal of Oncology*, 50:1383–1391.

Xue, L. and Pang, Z. (2013). Statistical inference for a single-index varying-coefficient model. *Statistics and Computing*, 23:589–599.

Yin, G., Li, H., and Zeng, D. (2008). Partially linear additive hazards regression with varying coefficients. *Journal of the American Statistical Association*, 103:1200–1213.

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

Zhang, C. H. (2007). Penalized linear unbiased selection. Technical report, Department of Statistics and Bioinformatics, Rutgers University.

Zhang, X., Qiao-Li, L., Huang, Y.-T., Zhang, L.-H., and Zhou, H.-H. (2017). Akt/FoxM1 signaling pathway-mediated upregulation of MYBL2 promotes progression of human glioma. *Journal of Experimental & Clinical Cancer Research*, 36(105).

Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., and Ma, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics*, 16:291–303.

Zhao, Y., Xue, L., and Feng, S. (2019). Estimation for a partially linear single-index varying-coefficient model. *Communications in Statistics - Simulation and Computation*. doi: 10.1080/03610918.2019.1680691.

Zhou, Y., Shen, J. K., Hornicek, F. J., Kan, Q., and Duan, Z. (2016). The emerging roles and therapeutic potential of cyclin-dependent kinase 11 (CDK11) in human cancer. *Oncotarget*, 7:40846–40859.

Zhu, R., Zhao, Q., Zhao, H., and Ma, S. (2016). Integrating multidimensional omics data for cancer outcome. *Biostatistics*, 17:605–618.

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