The Pennsylvania State University
The Graduate School

PROCEDURES FOR FEATURE SCREENING AND INTERACTION IDENTIFICATION IN HIGH-DIMENSIONAL DATA MODELLING

A Dissertation in
Statistics
by
Ling Zhang

© 2019 Ling Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2019

The dissertation of Ling Zhang was reviewed and approved* by the following:

Runze Li
Eberly Family Chair Professor of Statistics
Dissertation Advisor, Chair of Committee

Yanyuan Ma
Professor of Statistics

Zhibiao Zhao
Associate Professor of Statistics

Rongling Wu
Distinguished Professor of Public Health Sciences and Statistics

Ephraim Hanks
Assistant Professor of Statistics
Chair of Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

Nowadays, rapid developments in computer technologies have greatly reduced the cost of collecting and storing a massive amount of data. As a result, data with ultrahigh dimensionality begins to enter our vision due to a cheaper cost. It makes new levels of scientific discoveries promising, but also brings us new challenges of analyzing and understanding these data. Variable selection methods, feature screening procedures, and random forest algorithms have been widely used in many scientific fields such as computational biology, health studies, and financial engineering. The goal is to recover the underlying model structure and make an accurate prediction when a large number of predictors are introduced at the initial stage, but only a small subset of them are truly associated with the response.

High dimensional survival data analysis is such a scientific field. In the first part of the dissertation, we propose a two-stage feature screening procedure for varying-coefficient Cox model with ultrahigh dimensional covariates. The varying-coefficient model is flexible and powerful for modeling the dynamic effects of coefficients. In the literature, the screening methods for varying-coefficient Cox model are limited to marginal measurements. Distinguished from the marginal screening, the proposed screening procedure is based on the joint partial likelihood of all predictors. Through this, the proposed procedure can effectively identify active predictors that are jointly dependent of, but marginally independent of the response. In order to carry out the proposed procedure, we propose an efficient algorithm and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property: with probability tending to one, the selected variable set includes the actual active predictors. Monte Carlo simulation is conducted to evaluate the finite sample performance of the proposed procedure, with comparison to SIS(Fan and Lv, 2008) procedure and SJS (Yang et al., 2016) for Cox model. The proposed methodology is also illustrated through the analysis of two real data examples.

Although very helpful and computationally efficient, feature screening is not a very powerful method to detect those marginal unimportant variables that participate in high order interaction effects. However, this is the advantage of random forest algorithms because tree structure is a natural and powerful structure for detecting interaction effects. The drawback of the random forest algorithms is that they don't pay enough attention to feature selection, and therefore include lots of redundancy when constructing the forest. This phenomenon will severely influence the interpretability and prediction performance of the forest especially when only a small proportion among large amount of candidate variables are important.

In the second part of the dissertation, we propose combining the advantages of forest algorithm and feature screening for a better understanding of the hidden mechanism. To achieve this, we propose a new two-layer random forest algorithm, "Iteratively Kings' Forests"(iKF), for feature selection and interaction detection in classification and regression problems. In the first layer, we modified the traditional forest constructing process so that we can fully explore the mechanism, both marginal and interaction effects, related to a given important variable(say "King" variable). In the second layer, we iteratively search the next important variable and iterate the process of the first layer for it. Finally, we not only obtain a screened variable index set but also output a short list of ranked highly possible interaction effects. Simulation comparisons are conducted to compare its performance with the feature screening procedure DC-SIS(Li et al., 2012) and random forest algorithm "iRF"(Basu et al., 2018). Also, we apply iKF procedure for an empirical analysis to identify important interactions in an early Drosophila embryo data and compare its performance with "iRF".

Table of Contents

Chapter 1

Chapter 2

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to sincerely thank my advisors, Professor Runze Li, for his valuable inputs and patient guidance throughout my research. He contributed a lot of time and effort in helping me in many aspects such as selecting research topics, developing methods, analyzing real data and interpreting results. I have learned from him a great amount about critical thinking and conducting scientific statistical research. All his help will be beneficial throughout my whole life.

I would also like to thank my committee members: Professor Yanyuan Ma, Zhibiao Zhao and Rongling Wu for their kind support. In particular, I want to thank my collaborators: Dr. Guangren Yang and Dr. Yifan Xia for the opportunities of working on feature screening on varying coefficient Cox model and social inequality issues. The tools and techniques I learned from them have also been applied to my thesis research. Also, I want to thank Zhao Chen and Jiayu Peng for discussing and commenting on research idea with me. Their constructive challenges and suggestions are very helpful.

In addition, I want to thank the statistics department at Penn State University for offering me the opportunity of graduate studies, and providing a friendly and comfortable environment just like home.

Last but not least, I would like to thank my parents for their unconditional love and support during the past five years. My heartfelt thanks also go to all my friends, for their consistent trust and encouragements, and the great time we had together.

This dissertation research was partly supported by Pennsylvania State University, National Science Foundation grants, DMS 1512422 and DMS 1820702, and National Institute on Drug Abuse, NIH grant P50 DA039838. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

Chapter 1

Introduction

1.1 Overview

Nowadays, "Big Data" has become an increasingly popular terminology. Rapid developments in technologies have enabled us to collect, store and process a massive amount of data at relatively low cost, and therefore make "Big Data" a regular resource. From a statistical perspective, "Big Data" is a dataset with a large sample size N or variable dimension p. Large sample size is appealing because it can give accurate estimates for a given model. More importantly, large variable dimension p prevents us from losing important variables and promises new levels of scientific discoveries. However, when a large number of variables are introduced at the initial stage, only a small subset are truly associated with the response. That brings us new challenges of analyzing and understanding these data.

To deal with these issues, variable selection plays an essential role in recovering the underlying model structure. In classical best subset selection, we search all candidate sub-models of the full model and decide the optimal model through criteria such as AIC (Akaike, 1974), BIC (Schwarz et al., 1978), RIC (Foster and George, 1994) and Mallows C_p (Mallows, 1973). However, best subset selection becomes infeasible for large p because the computing time increases at an exponential rate of p. Classical stepwise selection avoids this drawback by adding or subtracting one variable each time based on pre-specified criteria. However, it leads to a local optimal model. To deal with these drawbacks, regularization methods, such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and MCP (Zhang et al., 2010), are developed to simultaneously select important variables and give coefficient estimates. However, when analyzing data with an ultrahigh dimension of features, for example, millions of predictors, the regularization methods may fail due to the tremendous computational burden and result instability.

Recently, feature screening in ultrahigh dimensional data analysis has become one of the most important statistical topics. By effectively reducing ultrahigh dimensionality to a moderate size, feature screening procedures receive considerable attentions in recent literature. Various feature screening methods have been developed in different contexts. Fan and Lv (2008) proposed a marginal screening procedure for ultrahigh dimensional Gaussian linear models, and further demonstrate that it possesses a sure screening property under certain conditions. Feature screening procedures for varying-coefficient model (i.e. VCM) with ultrahigh dimensional covariates have also been proposed in the literature. Liu et al. (2014) developed a SIS procedure for ultrahigh dimensional VCM by taking conditional Pearson correlation coefficients as marginal utility for ranking importance of predictors. Fan et al. (2014) proposed a SIS procedure for ultrahigh dimensional VCM by extending B-spline techniques in Fan et al. (2011) for additive models. Xia et al. (2016) further extend the SIS procedure proposed in Fan et al. (2014) to generalized varying coefficient models (GVCM). Cheng et al. (2016) proposed a forward variable selection procedure for ultrahigh dimensional VCM based on techniques related to B-splines regression and grouped variable selection. Song et al. (2014) extended the proposal of Fan et al. (2014) for longitudinal data without considering within-subject correlation, while Chu et al. (2016) proposed a new SIS procedure that ranks weighted residual sum of squares, utilizes the information of within-subject correlation, and highly improves the accuracy of feature screening for longitudinal data.

Therefore, a computational efficient solution for high dimensional problems is to incorporate a two-stage procedure. At first, feature screening procedures attempt to filter out those predictors that are clearly not important by ranking different utility measurements. After that, we could obtain a substantially smaller subset and then pick the “best” subset of predictors by other methods such as variable selection. By implementing this two-stage procedure, the computing time could be reduced from several days to a few minutes.

Survival analysis is an important subject widely used in medical science, economics, finance, social science, and others. In many studies, survival data have primary outcomes or responses subject to censoring. The Cox model (Cox, 1972) is the most commonly-used regression model for survival data, and the partial likelihood method has become a standard approach for parameter estimation and statistical inference. Fan et al. (2010) extended screening methods SIS and ISIS in GLM to Cox model and employ the maximum of the partial likelihood depending on k-th covariate as k-th marginal untility. By specifying the desired false positive rate, Zhao and Li (2012) proposed a principled sure independence screening (PSIS) procedure for Cox model, which provides a more theoretical solid thresholding criterion to determine the number of variables to retain. Sure joint screening(SJS) proposed by Yang et al. (2016) studied the feature screening for Cox model with ultrahigh dimensional non-independent covariates. However, none of the above have considered a sure joint screening procedure for varying coefficient Cox model. We will address this issue in chapter 3.

Although the two-stage procedure of feature screening procedures followed by variable selection methods is very important and efficient, it usually only works well in finding variables with marginal effects on the response. However, interaction effects exist in many natural phenomena and process. To better understand the mechanism behind them, we want to detect not only the variables with significant marginal effects, but also variables only participating in high-order interaction effects. For the interaction effects, we want to further figure out which variables work together in the mechanism.

Forest algorithms have been proposed for feature and interaction selection in the literature. Breiman (2001) proposed “Random Forest”(RF) for classification and prediction, and use Permutation Variable Importance Measure (PVIM) for ranking variable importances. However, when the number of features is huge and the percentage of truly imformative features is small, the performance of RF declines significantly. To solve this, Díaz-Uriarte and De Andres (2006) selected genes by iteratively fitting RF and dropping a pre-specified proportion of genes each round. Instead of dropping features, Amaratunga et al. (2008) proposed a feature-weighted version of RF for feature ranking and selection under the name “Enriched Random Forests”. Beside feature selection, some forest alogrithms have been developed to further improve our understanding of model mechanism through detecting interaction effects. Assuming both features and the response are binary, Shah and Meinshausen (2014) proposed the “Random Intersection Trees” procedure to discover interactions using the intersection of d randomly chosen sets of active features, whose d corresponding responses are from the same category. By combining feature-weighted RF (Amaratunga et al., 2008) and “Random Intersection Trees”(Shah and Meinshausen, 2014), Basu et al. (2018) proposed an “Iterative Random Forest” algorithm to discover interactions for problems with binary response, and continuous or categorical features. Furthermore, Basu et al. (2018) proposed a “stability score” based on an “outer layer” of bootstrapping to assess the stability of recovered interactions. However, their algorithm just studied the binary response case. Moreover, by removing interactions that are a strict subset of another interaction with high stability score, the algorithm gives up studying the detail structures between variables within this interaction.

1.2 Contributions

In the first part of this dissertation, we propose a new feature screening procedure for ultrahigh dimensional varying-coefficient Cox model. It is distinguished from the SIS procedures proposed by Fan et al. (2010) and Zhao and Li (2012) in that it is based on the joint partial likelihood of potentially important features rather than the marginal partial likelihood of individual features. Non-marginal screening procedures have been demonstrated to have their advantage over the SIS procedures in the context of generalized linear models. For example, Xu and Chen (2014) proposed a feature screening procedure for generalized linear models via the sparsity-restricted maximum likelihood estimator and demonstrate that their approaches perform significantly better than the SIS procedures in some scenarios. Compare to the sure joint screening procedure proposed in Yang et al. (2016), we generalize the model to varying coefficient setting and first employ Hoeffdings inequality for a sequence of martingale differences to establish a concentration inequality for the score function of partial likelihood. Furthermore, we establish the sure screening property for the proposed varying-coefficient sure joint screening (V-SJS) procedure in Cox model.

To evaluate its performance, we conduct Monte Carlo simulation studies for different settings of nonzero coefficient functions to assess the finite sample performance and compare it with the existing sure screening procedures under the constant coefficient assumption. Our numerical results show that the proposed V-SJS procedure outperforms the existing sure independence screening (SIS) and sure joint screening (SJS) procedures. We also demonstrate this with the empirical analysis of two real data.

In the second part, we propose a framework, "Iteratively Kings' Forests", for feature selection and interaction detection in classfication and regresssion problems. The framework is also a two stage process.

For the first stage, given the prior knowledge that one variable is important, we treat it as a "King" and construct an iteratively weighted forest with the "King" as the root node of every tree. In the iteratively reweighted process, we use the "Permutation Variable Importance Measure" (PVIM) of the "King" variable as a criteria to search the trees including other variables that participate in the same interactions with the "King". That is, larger PVIM means a better chance that other variables are modelled in the tree. If the PVIM of "King" is positive, the weights of all variables in this tree will be increased by the amount of PVIM. As a result, variables participating in the same interactions with the "King" will gradually gain larger weights, and therefore be likely to line in the same path of different trees in the forest.

Although the first stage tremendously increases the possiblity of modelling interaction effects related to the "King", other important variables will still be selected into trees and get large weights. In the second stage, we iteratively choose the variable with largest weight from the variables haven't been selected as "King" as the new "King", and construct "King's Forest" for each of them. We keep doing this until some convergence criterion are met.

Through this two stage procedure, we conduct both feature selection and interaction detection at the same time, and could therefore get a thorough and in-depth understand of the mechanism. Furthermore, we propose a criterion to help identify the order of related interactions. Through iteratively constructing forests for important variables, we outline the hidden model structure by selecting important features and interactions.

For the performance evaluation, we conduct Monte Carlo simulation studies to compare it with both DC-SIS(Li et al., 2012) and iRF (Basu et al., 2018) under different settings of regression and classification problems. Our numerical results show that the proposed procedure outperforms them in both feature selection and interaction detection.

In the third part, we propose a multi-dimensional economic dispersion index (MEDI) based on the Lorenz hyper-surface determined by the distributions of multiple social resources to comprehensively evaluate the social inequality level. As we know, the Gini index(Gini, 1912, 2005, 1921) is widely used in economics as a measure of inequality with respect to income or wealth. However, it is not applicable when we consider evaluating the inequality level using more than one social resources. The proposed MEDI is a natural extension of the Gini index, and is equivalent to the Gini index in the present of only one resource. Furthermore, we propose the estimator of MEDI with good statistical properties and develop an algorithm to calculate the estimate. We further apply MEDI for an empirical analysis to evaluate the social inequality level of Chinese provincial capitals. The results reveal some interesting phenomena of Chinese social inequalities, and also demonstrate how MEDI captures more information in complex economic situations than the classic Gini index.

1.3 Organization

The rest of this dissertation proposal is organized as follows. In chapter 2, we review the literature on variable selection methods, feature screening procedures, Cox model in survival data analysis, varying coefficient models, Random forest algorithms and measures of social inequality level. In chapter 3, we propose a new two-stage feature screening procedure for the varying-coefficient Cox model and demonstrate the ascent property of our proposed algorithm. We also study the sampling property of the proposed procedure and establish its sure screening property. Furthermore, we present numerical comparisons through simulation and empirical real data analysis. The chapter 4 gives the proposed framework, “Iteratively Kings’ Forests”, and uses two examples to demonstrate how to use it for feature and interaction selection. Simulation studies are conducted to evaluate its performance by comparing it with DC-SIS and iRF under differents scenarios. In chapter 5, we point out the disadvantages of the Gini index, and derive a multi-dimensional economic dispersion index (MEDI) to more comprehensively evaluate the social inequality level. In addition, we establish the consistent property of emperical MEDI and calculate the emperical MEDI under different scenarios.

Chapter 2

Literature Review

This chapter is organized as follows. First, we review well-established variable selection methods for linear regression and generalized linear regression models, including classical variable selection criteria and penalized regression approaches. Feature screening methods based on different assumptions for ultrahigh-dimensional data analysis are reviewed in section 2.2. Section 2.3 presents a brief review of some basic concepts and commonly-used models in survival data analysis. Furthermore, existing variable selection and feature screening procedures are summarized in this part. In the end, we introduce common structure of varying coefficient models and widely used coefficient function estimation methods. Feature screening procedures and varying coefficient Cox model are reviewed in section 2.4, based on which we will propose a new screening procedure for Cox model in Chapter 3.

2.1 Variable Selection Methods

Variable selection plays an important role in high-dimensional linear regression. To prevent the missing of important variables, we usually try to include every potential influential predictor at the beginning stage of statistical modeling. As a result, it is natural to assume that many predictors do not contribute to the response in the true model. Under this sparsity assumption, statisticians make great efforts on selecting important variables and getting a parsimonious model with good prediction accuracy and interpretability. In this section, we first study the classical variable selection criteria, and then review the regularized variable selection methods via penalized least squares and likelihood. Considering the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ is a $n \times d$ predictor matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d)^T$ denotes the coefficient vector, and $\boldsymbol{\epsilon}$ is the $n \times 1$ independently identical distributed noise vector with mean zero.

2.1.1 Classical Variable Selection Criteria

Classical variable selection is to select the subset of variables with best prediction behavior. To evaluate the performance of regression model, statisticians have developed a variety of variable selection criteria. They are measures of fit with a penalty for the number of parameters. Residual sum of squares defined below is a familiar measure of fit we want to minimize

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.2}$$

where $\hat{y}_i$ is the fitted value for y_i

Denote RSS_p as the residual sum of squares with p variables ($1 \leq p \leq d$) in the model. As we know, RSS_p will certainly decrease as p increase. Therefore, selecting model simply based on RSS will easily lead to the full model suffering from the overfitting problem. To select an optimal sub-model, the intuition is to add a term to penalize the number of parameters to the objective function. Several widely used variable selection criteria such as adjusted R^2, AIC, BIC, RIC and C_p are constructed based on this intuition. All of these criteria are applied through so-call best subset selection. By fitting all subsets of the full model, we can get the best sub-model based on any of given criteria we mention above. However, different criteria usually produce different "best models". To evaluate the performance of selected "best models", we will use the prediction sum of squares (Allen, 1974) defined as

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{ip})^2, \tag{2.3}$$

where $\hat{y}_{ip}$ is the predicted value for y_i based on a subset of p-variables. To get an accurate and stable $PRESS_p$ statistic, cross-validation methods such as leave-one-out cross-validation, k-fold cross-validation and generalized cross-validation are widely used to estimate the PRESS statistic.

All aforementioned variable selection criteria are listed as follow:

- **Adjusted R^2:** The adjusted R^2 statistic with p variables is defined as

  $$AR_p^2 = 1 - (1 - R_p^2)\frac{n-1}{n-p-1} \tag{2.4}$$

  Unlike R^2, the adjusted R^2 can be used not only for how well the fitting of the p predictors is, but also for variable selection via adding a penalty on the number of predictors.

- **Akaikes information criterion(AIC):** The AIC_p statistic for linear model with p variables is defined as

  $$AIC_p = RSS_p + 2p\sigma^2 \tag{2.5}$$

  where σ^2 is usually estimated by $\hat{\sigma^2} = \frac{RSS_d}{n-d}$.

- **Bayesian Information Criterion(BIC):** Similar extension of Akaikes Information Criterion (BIC, Schwarz, 1978) yields BIC_p statistic in linear model, which is defined as

  $$BIC_p = RSS_p + log(n)p\hat{\sigma^2} \tag{2.6}$$

  Compared with AIC, BIC penalizes higher for more complicated models with larger p. Hence the BIC tends to favor smaller models than AIC.

- **Risk Inflation Criterion(RIC):** Risk Inflation Criterion (RIC, Foster and George, 1994) for linear model is defined as

  $$RIC_p = RSS_p + 2log(d)p\hat{\sigma^2} \tag{2.7}$$

- C_p **statistic:** The C_p statistic (Mallows, 1973) is defined as

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \tag{2.8}$$

  Hence, minimization of C_p with respect to p is equivalent to the minimization of AIC_p.

However, when d is large, it is computionally infeasible for the exhaustive search for all 2^d possible subsets. In practice, stepwise selection is widely used to search a good subset instead of best subset selection. Stepwise regression is a fitting procedure in which the choice of predictive variables is automatically carried out. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criteria such as F-tests, t-tests, adjusted R^2, Akaike information criterion, Bayesian information criterion, Mallows's C_p, PRESS, or false discovery rate. The procedure is carried out with three main approaches from two directions

- **Forward selection** starts with no variables in the model, tests the addition of each variable based on a chosen fit criterion, adds the variable gives the most statistically significant improvement of the fit and repeats this process until no variable improves the model to a statistically significant extent.

- **Backward elimination** starts with all candidate variables, tests the deletion of each variable based on a given fit criterion, deletes the variable whose loss results in the most statistically insignificant loss of the fit and repeats this process until no variable can be deleted without a statistically significant loss of fit.

- **Bidirectional elimination**, a combination of the above, tests at each step for variables to be included or excluded.

Technical details are referred to Miller (2002).

2.1.2 Variable Selection via Penalized Least Squares

Subset selection and stepwise selection provide interpretable models with selected most significant predictors. However, they both have inherent drawbacks. As we mention before, it is computionally infeasible for subset selection when d is large. For stepwise selection, the greedy strategy it uses only results in a local optimal model. As d increases, the probability of the consistency between local optimal and global optimal model will sharply decrease. Furthermore, as d increases, both selection procedures are increasingly sensitive to small changes in the data and get very different selected models. To get a stable global optimal model and improve prediction accuracy, penalized least squares (PLS) methods are proposed. Instead of minimizing the residual sum of squares, we obtain the estimate by minimizing a penalized least squares function

$$Q(\beta) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.9}$$

where $p_\lambda(\cdot)$ is the penalty function with tuning parameter λ selected by a data-driven method to control the model complexity. For simplicity, we assume that the tuning parameter λ and penalty function form for all coefficients are the same. To be more specific, we present several commonly used penalty functions and their specific properties.

Frank and Friedman (1993) propose a general framework called the bridge regression with L_q penalty

$$p_\lambda(|\theta|) = \lambda|\theta|^q \tag{2.10}$$

where $0 \leq q \leq 2$. When $q = 0$, the penalty function only penalizes the number of predictors. Previous mentioned best subset selection criteria fall into this category. L_1 and L_2 correspond to famous LASSO (Tibshirani, 1996) and Ridge regression (Hoerl and Kennard, 1970), which will be introduced later. By definition, L_0, L_1 and L_2 penalty are all special cases of L_q penalty. For penalized least squares methods, coefficient estimate is continuous with respect to the OLS estimate only when $q \geq 1$. However, if $q > 1$, the L_q penalty can not produce a sparse solution. L_2 penalty for Ridge regression is a perfect example

$$p_\lambda(|\theta|) = \frac{1}{2}\lambda|\theta|^2 \tag{2.11}$$

Ridge regression is proposed to deal with multi-collinearity problem in predictors. It continuously shrinks coefficients and gets a more stable model with an explicit form of coefficients estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$. However, it can not set any coefficient to 0 and fail to get a sparse model. To solve this, Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator(LASSO) to shrink coefficients and select significant predictors at the same time. The LASSO penalty is

$$p_\lambda(|\theta|) = \lambda|\theta| \tag{2.12}$$

The LASSO estimate continuously shrinks the OLS estimate and sets some coefficients exactly to 0. Hence, it selects important predictors automatically to produce an interpretable model like subset selection and enjoys the stability of ridge regression. However, LASSO estimate is biased.

We have been discussed several penalty functions so far. However, they all have their own drawbacks. L_0 penalty can conduct variable selection, but the computational cost and the result instability make it unfeasible for high dimensional data analysis. L_2 penalty can solve the multi-collinearity problem by shrinking the estimated coefficients to make the result stable, but it cannot do variable selection. L_1 penalty(LASSO) can provide stable estimation and variable selection, but the coefficient estimates are biased, especially for the large true coefficients. Therefore, it is natural to ask: What kind of penalty function is a satisfactory one? To answer this question, Fan and Li (2001) advocated that a good penalty function should result in an estimator with three properties.

1. Unbiasedness. The coefficient estimates are nearly unbiased, especially for those estimates with large true unknown parameters, to avoid unnecessary modeling bias.
2. Sparsity. The coefficient estimates become a thresholding rule. It automatically sets unimportant coefficient estimates to zero and produces a sparse model.
3. Continuity. The coefficient estimates are continuous with a small change in data to avoid instability in model prediction.

To guarantee the above three properties, Antoniadis and Fan (2001) propose the following three conditions

1. Unbiasedness. iff $p_{\lambda}^{'}(|\theta|) = 0$ for large $|\theta|$.
2. Sparsity. if $min_{\theta} p_{\lambda}^{'}(|\theta|) + |\theta| > 0$.
3. Continuity: if $argmin_{\theta}\{p_{\lambda}^{'}(|\theta|) + |\theta|\} = 0$.

As we point out before, all the previous mentioned penalty functions cannot simultaneously satisfy all three conditions for unbiasedness, sparsity, and continuity.

To satisfy all three conditions, Fan and Li (2001) propose a continuously differentiable penalty function, the smoothly clipped absolute deviation (SCAD) penalty, defined by its derivative

$$p_{\lambda}^{'}(\theta) = \lambda I(\theta \leq \lambda) + \frac{a\lambda - \theta}{a-1} I(\theta > \lambda), a > 2 \tag{2.13}$$

The above formula includes three intervals. For the first interval, the derivative is the constant λ, which is equivalent to LASSO. The derivative linearly decreases to 0 for $\lambda \leq \theta \leq a\lambda$ in the second interval. In the third interval, no more penalty is added to the PLS function when θ is larger than $a\lambda$. The resulting solution to SCAD penalty is

$$\hat{\theta}(z) = \begin{cases} 0, & \text{if } |z| \leq \lambda \\ sgn(z)(|z| - \lambda), & \text{if } \lambda < |z| \leq 2\lambda \\ \{(a-1)z - sgn(z)a\lambda\}/(a-2), & \text{if } 2\lambda < |z| \leq a\lambda \\ z, & \text{if } |z| \geq a\lambda \end{cases} \tag{2.14}$$

As we can see, SCAD fixes the biased problem for LASSO by reducing the penalty for large θ. It corresponds to a nonconcave symmetric function singular at the origin with knots at λ and $a\lambda$ and satisfies the condition of unbiasedness, sparsity and continuity. Two unknown parameters λ and a are needed to be decided here. $a = 3.7$ is found to be best and widely used in practice, while λ is the tuning parameter. Furthermore, Fan and Li (2001) established the asymptotic **oracle property** of the SCAD penalty under penalized likelihood setting. That is, the resulting coefficient estimates with the SCAD penalty works so well by giving the exactly correct submodel if the regularization parameter is appropriately chosen.

In addition to SCAD, Zou (2006) also proposed an Adaptive LASSO penalty satisfying the three conditions

$$p_\lambda(|\theta|) = \lambda\hat{\omega}|\theta| \tag{2.15}$$

where $\hat{\omega} = 1/|\hat{\beta}_0|^\gamma$ with $\hat{\beta}_0$ being any consistent estimate. In practice, we use OLS estimate as $\hat{\beta}_0$ here. The adaptive LASSO assigns different weights to different coefficients based on the reciprocal of initial OLS estimate. That is, the coefficients with smaller initial estimates will be largely penalized, while coefficients with large initial estimates will be just slightly penalized. The Adaptive LASSO also enjoys the oracle properties with appropriate λ.

Zhang et al. (2010) proposed the minimax concave penalty (MCP) defined as

$$p_\lambda(\beta_j) = \lambda \int_0^{|\beta_j|} (1 - \frac{\theta}{a\lambda})_+ d\theta \tag{2.16}$$

The MCP is also proved to enjoy the oracle property and three desirable properties.

2.1.3 Technical Details

To get the optimal resulting estimator, computational algorithms and tuning parameter selection are two necessary steps. We will review both aspects as follow.

2.1.3.1 Computational Algorithms

The penalty functions aforementioned could be categorized into two mainstream types, convex penalty such as LASSO (L_1) and the nonconvex penalization such as SCAD and MCP. The convex penalty, for example LASSO, is popular due to its computational efficiency. Fu (1998) provided a shooting algorithm for L_q penalized regression, wihle the Least Angel Regression (LARS) was suggested by Efron et al. (2004), which could be applied to solve the optimization problem of LASSO type penalty. Furthermore, the coordinate descent algorithm has been shown to be very efficient for convex penalization penalization problems.

The nonconcave penalty functions SCAD and MCP have recently achieved great success because they can eliminate the estimation bias for parameters and attain oracle properties. However, since nonconvex penalty usually has multiple local minimizers, it will be more difficult to arrive at the global optimal point. In this subsection, we would discuss three main computational algorithms, LARS, LQA and LLA for the minimization problems of PLS. Among them, LARS is commonly used for convex penalty, while LQA and LLA are widely used for nonconvex penalty terms.

Least Angle Regression (LARS): Efron et al. (2004) propose the least angle regression (LARS) algorithm for penalized variable selection. This fast and efficient algorithm that can produce the entire LASSO solution path is defined as:

1. Start with all coefficients β_j equal to zero.
2. Find the predictor x_j most correlated with y.
3. Increase the coefficient β_j in the direction of the sign of its correlation with y. Take residuals $r = y - \hat{y}$ along the way. Stop when some other predictor x_k has as much correlation with r as x_j has.
4. Increase (β_j, β_k) in their joint least squares direction, until some other predictor x_m has as much correlation with the residual r.
5. Continue until: all predictors are in the model.

The LARS-Lasso relationship: Denote $\hat{\boldsymbol{\beta}}$ as a Lasso solution, with $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$. Efron et al. (2004) shows that the sign of any nonzero coordinate $\hat{\beta}_j$ of Lasso solution must agree with the sign s_j of the current correlation $\hat{c}_j = \mathbf{x}_j^{'}(\mathbf{y} - \hat{\boldsymbol{\mu}})$

$$sign(\hat{\beta}_j) = sign(\hat{c}_j) = s_j, \tag{2.17}$$

However, the LARS algorithm does not enforce restriction (2.17). Therefore, by imposing a minor modification on the LARS algorithm, the full set of Lasso solutions can be generated.

Define $\hat{\mathbf{d}}$ to be a m dimensional equiangular vector making equal angles with the columns of m selected predictors. Moving in the positive γ direction along the LARS line, we see that

$$\boldsymbol{\mu}(\gamma) = X\boldsymbol{\beta}(\gamma), \text{ where } \beta_j(\gamma) = \hat{\beta}_j + \gamma\hat{d}_j$$

Therefore, $\beta_j(\gamma)$ will change sign at

$$\gamma_j = -\hat{\beta}_j/\hat{d}_j, \tag{2.18}$$

the first such change occurring at

$$\tilde{\gamma} = min_{\gamma_j>0}\{\gamma_j\}, \tag{2.19}$$

$\tilde{\gamma}$ equals infinity by definition if there is no $\gamma_j > 0$.

If $\tilde{\gamma}$ is less than $\hat{\gamma}$, then $\beta_j(\gamma)$ cannot be a Lasso solution for $\gamma > \tilde{\gamma}$ since the sign restriction (2.17) will be violated. Therefore, we can get the Lasso Modification for LARS. **Lasso Modification:** If $\tilde{\gamma} < \hat{\gamma}$, stop the ongoing LARS step at $\gamma = \tilde{\gamma}$ and remove the previous included index.

Local Quadratic Approximation (LQA): Fan and Li (2001) propose the Local Quadratic Approximation (LQA) to optimize nonconvex penalized objective function. Consider an initial estimate $\boldsymbol{\beta}^0$, for example $\boldsymbol{\beta}^{OLS}$, that is close to the minimizer of (2.9). Set $\hat{\beta}_j = 0$ if β_j^0 is smaller than a given thresholding value. Otherwise, approximate the penalty function by a local quadratic functions as

$$p_\lambda(|\beta_j|) = p_\lambda(|\beta_j^0|) + \frac{1}{2}\{p_\lambda^{'}(|\beta_j^0|)/|\beta_j^0|\}(\beta_j^2 - (\beta_j^0)^2). \tag{2.20}$$

Note that the residual sum of squares term $\frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$ is convex with respect to $\boldsymbol{\beta}$, therefore (2.9) could be locally approximated by

$$Q(\beta) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \frac{n}{2}\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}^0)\boldsymbol{\beta}, \tag{2.21}$$

where $\Sigma_\lambda(\boldsymbol{\beta}^0) = diag\{p_\lambda^{'}(|\beta_1^0|)/|\beta_1^0|, \ldots, p_\lambda^{'}(|\beta_d^0|)/|\beta_d^0|\}$, β_j^0 are the jth component of initial value $\boldsymbol{\beta}_0$. The solution for (2.21) can be found by iteratively computing the ridge regression

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{X}^T\mathbf{y}, k = 0, 1, \ldots \tag{2.22}$$

Local Linear Approximation (LLA): However, once LQA algorithm delete a covariate at any step, it wouldn't be included in the final selected model again. To resolve the drawback of the LQA, Zou and Li (2008) propose an efficient one-step sparse estimation procedure based on Local Linear Approximation (LLA). According to LLA, the penalty function could be locally approximated by

$$p_\lambda(|\beta_j|) = p_\lambda(|\beta_j^0|) + p_\lambda^{'}(|\beta_j - \beta_j^0|), \text{ for } \beta_j \approx \beta_j^0.$$

Similar to the LQA algorithm, we could set the OLS estimates as initial value $\boldsymbol{\beta}^0$ and then repeatedly solve the local linear approximation function for $k = 0, 1, \ldots$

$$\boldsymbol{\beta}^{k+1} = argmin_\beta\{\frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{d} p_\lambda^{'}(|\beta_j^k|)|\beta_j|\}. \tag{2.23}$$

From (2.23), we can see that the LLA algorithm transform the nonconvex penalty into a convex LASSO-type penalty. As a result, the minimization could be efficiently solved by the existing algorithms such as the previous reviewed LARS. Moreover, LLA automatically adopts a sparse estimator. The one-step LLA estimator ($k = 0$ for (2.23)) is as efficient as the fully iterative one given a good initial value, which avoids expensive computational cost.

Several other useful algorithms are also proposed to compute the nonconcave penalized estimators. Zhang et al. (2010) propose a PLUS algorithm to solve the penalized least squares with MCP penalty and SCAD penalty. Breheny and Huang (2011) not only apply coordinate descent algorithms to the SCAD and the MCP penalization problems, but also gave a frequently used R package "ncvreg" for R users. Fan et al. (2014) show that if the LLA algorithm is initialized at a LASSO optimum that satisfies certain properties, then the two-stage procedure produces an oracle solution for various nonconcave penalties. To get the global optimum from several local ones, Wang et al. (2013) study a calibrating CCCP algorithm to find a consistent solution path with probability approaching to one to contain the oracle estimator.

2.1.3.2 Tuning Parameters Selection

In this part, we briefly review the selection of tuning parameter λ. Given λ, PLS estimate $\hat{\beta}_\lambda$ could be obtained by applying previous mentioned algorithms. Model selection critera is a function of $\hat{\beta}_\lambda$, which is determined by λ. In practice, four popular tuning parameter selectors, AIC, BIC, K-fold cross validation (CV) and GCV, are widely used in practice

1. $AIC(\lambda) = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda||^2 + 2df_\lambda\hat{\sigma}^2$,
2. $BIC(\lambda) = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda||^2 + log(n)df_\lambda\hat{\sigma}^2$,
3. $CV(\lambda) = \sum_{k=1}^{K} ||\mathbf{y}_k - \mathbf{X}_k^T\hat{\boldsymbol{\beta}}_\lambda^k||^2$,
4. $GCV(\lambda) = \frac{1}{n}\frac{||\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda||^2}{\{1-df_\lambda/n\}^2}$

where $df_\lambda = trace\{\mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda + n\Sigma_\lambda(\hat{\beta}_\lambda))^{-1}\mathbf{X}_\lambda\}$ is the estimated degree of freedom in selected model depending on λ. And $\mathbf{y}_k$, $\mathbf{X}_k^T\hat{\boldsymbol{\beta}}_\lambda^k$ in $CV(\lambda)$ correspond to the testing set and the estimated model coefficients.

Intuitively, we could select the optimal tuning parameter λ by

$$\hat{\lambda} = argmin_\lambda\{Selector(\lambda)\} \tag{2.24}$$

Wang et al. (2007) demonstrate that the GCV-selector for the PLS with SCAD penalty can not select the tuning parameter consistently. They show that GCV behaved similarly to AIC and usually produced an overfitting selected model. Instead, Wang et al. (2007) further propose a high dimensional BIC-type tuning parameter selector as $HBIC_\lambda^* = log(\hat{\sigma}_\lambda^2) + \frac{log(n)}{n}df_\lambda$. Furthermore, they prove that $HBIC_\lambda^*$ owned the desirable oracle property and could consistently identify the true model.

2.1.4 Regularized Variable Selection via Penalized Likelihood

The methodology and algorithms in PLS can be generalized directly to likelihood-based generalized linear model, in which statistical inferences are based on the likelihood functions. The penalized maximum likelihood estimator can be defined to select significant variables.

Assume we have independently collected data $\{\mathbf{x}_i, y_i\}$. Given $\mathbf{x}_i$, y_i has a density $f_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i)$ with a known link function g. Therefore, the penalized likelihood can be defined as

$$PL(\boldsymbol{\beta}) = \sum_{i=1}^{n} log(f_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i)) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.25}$$

By maximizing (2.25), we can obtain the penalized MLE with respect to $\boldsymbol{\beta}$. Similar to PLS, we apply LLA and LQA algorithm to objective function $PL(\boldsymbol{\beta})$.

LLA approximates (2.25) by

$$\ell(\boldsymbol{\beta}^0)+\nabla\ell(\boldsymbol{\beta}^0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T\nabla^2\ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta}-\boldsymbol{\beta}^0)-n\sum_{j=1}^{d}p_\lambda^{'}(|\beta_j^0|)|\beta_j|, \tag{2.26}$$

where $\nabla\ell(\boldsymbol{\beta}^0)$ and $\nabla^2\ell(\boldsymbol{\beta}^0)$ are the first two partial derivatives of the likelihood function in (2.25). Since $\boldsymbol{\beta}^0 = \hat{\boldsymbol{\beta}}_{MLE}$ and $\nabla\ell(\hat{\boldsymbol{\beta}}_{MLE})^T = 0$, (2.26) could be expressed as

$$\ell(\boldsymbol{\beta}^0)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T\nabla^2\ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta}-\boldsymbol{\beta}^0)-n\sum_{j=1}^{d}p_\lambda^{'}(|\beta_j^0|)|\beta_j| \tag{2.27}$$

Through applying LLA, we transform the penalty term into a convex form. Hence local maximum of (2.25) could be obtained.

LQA algorithm could be applied if the likelihood function has first two partial derivatives continuous with respect to β. Applying LQA algorithm, (2.25) could be locally approximated by

$$\ell(\boldsymbol{\beta}^0)+\nabla\ell(\boldsymbol{\beta}^0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T\nabla^2\ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta}-\boldsymbol{\beta}^0)-\frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}^0)\boldsymbol{\beta}, \tag{2.28}$$

where $\Sigma_\lambda(\boldsymbol{\beta}^0) = diag\{p_\lambda^{'}(|\beta_1^0|)/|\beta_1^0|,\ldots,p_\lambda^{'}(|\beta_d^0|)/|\beta_d^0|\}$.

By implementing the Newton-Raphson algorithm, the quadratic maximization problem (2.28) yields the solution

$$\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 - \{\nabla^2\ell(\boldsymbol{\beta}^0) - n\Sigma_\lambda(\boldsymbol{\beta}^0)\}^{-1}\{\nabla\ell(\boldsymbol{\beta}^0) - \Sigma_\lambda(\boldsymbol{\beta}^0)\boldsymbol{\beta}^0\} \tag{2.29}$$

Under the likelihood setting, the definition of AIC, BIC and GCV statistic are also generalized to

$$AIC(\lambda) = -2\ell(\hat{\boldsymbol{\beta}}) + 2df_\lambda \tag{2.30}$$

$$BIC(\lambda) = -2\ell(\hat{\boldsymbol{\beta}}) + log(n)df_\lambda \tag{2.31}$$

$$CV(\lambda) = \frac{1}{n}\frac{-\ell(\hat{\boldsymbol{\beta}})}{\{1-df_\lambda/n\}^2} \tag{2.32}$$

where $df_\lambda = trace[\{\nabla^2\ell(\hat{\boldsymbol{\beta}}) + \Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\nabla^2\ell(\hat{\boldsymbol{\beta}})]$

2.2 Ultra-High Dimensional Feature Screening

Various variable selection methods for high dimensional data analysis have been studied in the previous section. For high dimension, we usually mean that the dimension p increases with sample size n at a polynomial rate: $p_n = O(n^\alpha)$ for some $\alpha > 0$. However, if the dimension p_n, as a function of n, increases with sample size n at an exponential rate: $p_n = O(\exp(an))$ for some $0 < a < 1$, variable selection methods can be computationally infeasible due to heavy computational cost and algorithmic instability (Fan et al., 2014).

Recently, data with ultrahigh dimensionality begins to enter our vision due to a cheaper cost. For example, in genome-wide association studies (GWAS), we could get inexpensive measurement of the whole genome that enables the generation of hundreds of thousands of single-nucleotide polymorphisms (SNPs). As a result, a new data analysis techniques to study ultrahigh dimensional data is increasingly demanded in practice. To solve this, Fan and Lv (2008) proposed the idea of feature screening. Feature screening is a two-stage computationally efficient procedure that first removes unimportant predictors to produce a moderate scale subset that contains all the active predictors with high probability, and then apply more sophisticated variable selection techniques to identify important predictors.

2.2.1 Sure Independence Screening for Linear Model

Fan and Lv (2008) proposed Sure Independence Screening (SIS) method based on marginal correlation ranking between covariates and the response y in linear model setting. Consider the following linear model

$$\boldsymbol{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.33}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is a $n \times 1$ response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$ is a $n \times p$ design matrix. It is assumed that $\mathbf{x}_i's$ are independent from each other and standardized to have mean 0 and standard deviation 1. We have the sparity assumption that only a small subset of $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^T$ is truly associated with the response. That is, $\boldsymbol{\beta}$ is sparse under $p \gg n$ setting. Sometimes, we further assumed the number of nonzero components $m < n$ and $m = [a\frac{n}{log(n)}] = [a\gamma(n)n]$, where a is a scale parameter usually setted as 1. Fan and Lv (2008) define the marginal correlation by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$

$$\boldsymbol{\omega} = \mathbf{X}^T\mathbf{y}, \tag{2.34}$$

Thus, for any given $\gamma \in [0, 1]$, a submodel can be defined as

$$\mathcal{M}_m = \{1 \leq j \leq p : |\omega_j| \text{ is among the first m largest of all}\}, \tag{2.35}$$

Thus, we effectively reduce the full model with dimensionality p to a submodel with m nonzero components.

However, we don't want to miss important variables during this process. To deal with this concern, Fan and Lv (2008) prove that SIS enjoys the sure screening property.

Sure Screening Property: Define $\mathbf{z} = \Sigma^{-1/2}\mathbf{x}$ and $\mathbf{Z} = \mathbf{X}\Sigma^{-1/2}$ where $\Sigma = cov(\mathbf{x})$. Five regularity conditions are needed to establish sure screening property:

C1 For $p > n$, there exist $\xi > 0$ such that $log(p) = O(n^{\xi})$. This condition shows that SIS is suitable for the ultrahigh dimensional cases.

C2 $\mathbf{z}$ has a spherically symmetric distribution, while $\mathbf{Z}$ has a concentration property. That is, there exists some c, $c_1 > 1$ and $C_1 > 0$ satisfying the following inequality for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{z}}$ of $\mathbf{z}$ with $cn < \tilde{p} \leq p$

$$P(\lambda_{max}(\tilde{p}^{-1}\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T) > c_1 \text{ and } \lambda_{min}(\tilde{p}^{-1}\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T) < 1/c_1) \leq e^{-C_1 n}.$$

where $\lambda_{max}(A)$ and $\lambda_{min}(A)$ are the largest and smallest eigenvalue of matrix A. This condition makes restriction on ξ through concentration property.

C3 Assume $var(Y) = O(1)$, there exists some $\kappa \geq 0$ and $c_2, c_3 > 0$,

$$\min_{i\in\mathcal{M}_0} |\beta_j| \geq \tfrac{c_2}{n^{\kappa}} \text{ and } \min_{i\in\mathcal{M}_0} |cov(\beta_j^{-1}\mathbf{Y}, X_i)| \geq c_3$$

This condition avoids the case in which a significant variable is marginally uncorrelated but jointly correlated with y .

C4 For some $\tau \geq 0$ and $c_4 \geq 0$, we have

$$\lambda_{max}(\Sigma) \leq c_4 n^{\tau}.$$

This condition rules out strong collinearity.

C5 $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. This is a common restriction on the error distribution.

Let $\mathcal{M}^*$ represent the true sparse model. Suppose conditions (C1)-(C5) are satisfied. If $2\kappa + \tau < 1$, then for some $\theta < 1 - 2\kappa - \tau$ such that when $\gamma \sim cn^{-\theta}$ with $c > 0$, we have for some $C > 0$,

$$P(\mathcal{M}^* \in \hat{\mathcal{M}}_{\gamma}) = 1 - O(\exp(-Cn^{1-2\kappa}/log(n))) \tag{2.36}$$

Therefore, the screened submodel will contain the true model with probability tending to one.

$$P(\mathcal{M}^* \subset \mathcal{M}_m) \to 1, \text{ as } n \to \infty. \tag{2.37}$$

Since the submodel with moderate size m will contain the true model with an overwhelming probability, many standard variable selection methods such as stepwise selection (Miller, 2002), SCAD (Fan and Li, 2001) and adaptive lasso (Zou, 2006) can be applied to further simply the model and produce coefficient estimates.

Motivated by SIS, Wang (2009) borrow the idea of forward regression (FR) and proposed FR screening method for the ultrahigh dimensional situation. Let $\mathcal{M}^{(k)}$ be the index set of kth submodel. The FR algorithm is introduced as follow.

Step 1 Set initial index set $\boldsymbol{\mathcal{M}}^{(0)} = \emptyset$.

Step 2 At the kth($k \geq 1$) step, for every $j \in \{1, \ldots, p\} \setminus \mathcal{M}^{(k-1)}$, fit a candidate model with variables from the index set $\mathcal{M}_j = \mathcal{M}^{(k-1)} \bigcup \{j\}$. Then compute RSS of the candidate model $\mathbf{y} = \mathbf{x}_{\mathcal{M}_j}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for all j and get $RSS_j = \mathbf{y}^T(\mathbf{I}_n - \mathbf{x}_{\mathcal{M}_j}(\mathbf{x}_{\mathcal{M}_j}^T\mathbf{x}_{\mathcal{M}_j})^{-1}\mathbf{x}_{\mathcal{M}_j}^T)\mathbf{y}$, where $\mathbf{I}_n$ is a $n \times n$ identity matrix. Choose the model with the smallest RSS_j and update the index set to $\mathcal{M}^{(k)}$.

Step 3 Repeat **Step 2** for n times and gets a solution path with n nested models $\mathbb{M} = \{\mathcal{M}^{(k)} : k = 1, \dots, n\}$.

Under certain conditions, the solution path $\mathbb{M}$ is defined to achieve screening consistency

$$P(\mathcal{M}_0 \subset \mathcal{M}^{(k)} \in \mathbb{M} \text{ for some } 1 \leq k \leq n) \to 1 \text{ if } n \to \infty, \tag{2.38}$$

Denote k^* as the smallest integer satisfying (2.38), then every model $\mathcal{M}^{(k)}$ with $k > k^*$ enjoys the screening consistency. To decide the optimal model, Wang (2009) suggest to use the following BIC criterion (Chen and Chen, 2008) for the model $\mathcal{M}^{(k)}$

$$BIC(\mathcal{M}^{(k)}) = \log(\frac{1}{n} RSS(\mathcal{M}^{(k)})) + \frac{1}{n} |\mathcal{M}^{(k)}| (\log(n) + 2\log(p)) \tag{2.39}$$

The optimal model with the smallest BIC can be selected from the solution path $\mathbb{M}$. Denote the optimal model as $\hat{\mathcal{M}}$. Wang (2009) show that the model $\hat{\mathcal{M}}$ is screening consistent.

2.2.2 Sure Independence Screening for Generalized Linear Model

SIS proposed in Fan and Lv (2008) is developed under linear model assumption. However, its property highly depends on the joint normality assumption, which limits its use to data from other distributions and models. Fan and Song (2010) generalize independence screening to generalized linear models and proposed a more general independence screening method by ranking the maximum marginal likelihood estimator (MMLE) or maximum marginal likelihood itself. Consider the case where the response Y comes from an exponential family with the canonical form

$$f_Y(y|\mathbf{x}) = \exp\{y\mathbf{x}^T\boldsymbol{\beta} - b(\mathbf{x}^T\boldsymbol{\beta}) + c(y)\}, \tag{2.40}$$

where $\mathbf{x}$ is a $(p+1)$-dimensional predictor with the first element equals to 1, and $b(\cdot)$ and $c(\cdot)$ are known functions. Based on (2.40), the MMLE $\hat{\boldsymbol{\beta}}_j^M$ is defined as the maximizer of the marginal log-likelihood function:

$$\hat{\boldsymbol{\beta}}_j^M = (\hat{\beta}_{j0}^M, \hat{\beta}_j^M) = \arg\max_{\beta_0,\beta_j} \sum_{i=1}^n \ell(\beta_0 + \beta_j x_{ij}, y_i), j = 1, \ldots, p \tag{2.41}$$

where $\ell(\beta_0 + \beta_j x_{ij}, y_i) = (\beta_0 + \beta_j x_{ij})y_i - b(\beta_0 + \beta_j x_{ij})$. Based on $\hat{\boldsymbol{\beta}}_j^M$, we can rank the importance of features based on the magnitudes of marginal regression coefficients $|\hat{\boldsymbol{\beta}}_j^M|$ and produce a submodel:

$$\mathcal{M}_{\gamma_n} = \{1 \le j \le p : |\hat{\beta}_j^M| \ge \gamma_n\}, \tag{2.42}$$

where γ_n is a predefined threshold value. Under certain conditions, Fan and Song (2010) prove that MMLE has sure screening property and size control property if γ_n follows an ideal rate. The size control property states that the size of the submodel will be at most $O(n^{2\kappa}\lambda_{max}(\Sigma))$ if $\log(p) = o(n^{1-2\kappa})$. Here $\kappa < 1/2$ and $\Sigma = \text{cov}(\mathbf{x}_i)$.

Instead of independent outcome data, Xu et al. (2014) gave a screening method for longitudinal data with correlated outcomes based on generalized estimating equation (GEE). Assume the i-th subject is observed at J_i discrete time points. Let $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iJ_i})^T$ be the corresponding vector of response and $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ_i})^T$ be the corresponding $J_i \times P$ matrix of covariates, where $\mathbf{x}_{ik} = (X_{ik1}, \ldots, X_{ikp})^T$. Given $\mathbf{x}_{ik}$, the conditional mean of Y_{ik} is $E(Y_{ik}|\mathbf{x}_{ik}) = g^{-1}(\mathbf{x}_{ik}^T\boldsymbol{\beta})$, in which $g(\cdot)$ is a known link function. Assume that Y_{ik} is from an exponential family with a canonical link function. The GEE is defined as

$$G(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^T \mathbf{a}_i^{1/2}(\boldsymbol{\beta})\mathbf{R}_i^{-1}(\boldsymbol{\theta})\mathbf{a}_i^{-1/2}(\boldsymbol{\beta})(\mathbf{y}_i - \mu_i(\boldsymbol{\beta})) = 0, \tag{2.43}$$

where $\mathbf{a}_i(\boldsymbol{\beta})$ is a $J_i \times J_i$ diagonal matrix with kth diagonal element $\text{cov}(Y_{ik}|\mathbf{x}_{ik})$, and $\mathbf{R}_i(\boldsymbol{\theta})$ is a $J_i \times J_i$ working correlation matrix determined by parameters $\boldsymbol{\theta}$. To measure the importance of each covariate, Xu et al. (2014) define a population version of $G(\boldsymbol{\beta})$ as follow

$$g(\boldsymbol{\beta}) = E[\mathbf{x}^T\mathbf{a}^{1/2}(\boldsymbol{\beta})\mathbf{R}^{-1}(\boldsymbol{\theta})\mathbf{a}^{-1/2}(\boldsymbol{\beta})(\mathbf{y} - \mu(\boldsymbol{\beta}))], \tag{2.44}$$

thus, the marginal dependence between $\mathbf{y}$ and $\mathbf{x}_j$ can be measured by $g_j(\mathbf{0})$, the jth element of $g(\boldsymbol{\beta})$ when $\boldsymbol{\beta} = \mathbf{0}$. Based on an empirical estimate of the working correlation matrix $\mathbf{R}(\hat{\boldsymbol{\theta}})$, we can estimate $g(\mathbf{0})$ by $\hat{G}(\mathbf{0})$:

$$\hat{G}(\mathbf{0}) \equiv (\hat{G}_1(\mathbf{0}), \ldots, \hat{G}_p(\mathbf{0}))^T = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{a}_i^{1/2}(\mathbf{0}) \mathbf{R}_i^{-1}(\hat{\boldsymbol{\theta}}) \mathbf{a}_i^{-1/2}(\mathbf{0})(\mathbf{y}_i - \mu_i(\mathbf{0})). \tag{2.45}$$

Note that this procedure is computational efficient since it only produces one single evaluation $\hat{G}(\mathbf{0})$ of the GEE function $G(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = 0$ instead of p separate marginal models for $g_j(\mathbf{0})$, $j = 1, \ldots, p$. Furthermore, the authors show that GEE owns sure screening property even when the working correlation matrix is misspecified.

2.2.3 Joint Screening via Sparse MLE

Instead of ranking the importance of predictors based on marginal correlations or marginal magnitudes, Xu and Chen (2014) propose a new screening procedure considering the joint effects of features via the sparsity-restricted maximum likelihood estimator (SMLE) under generalized linear model setting. Specifically, based on the log-likelihood function $\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{(\mathbf{x}_i^T \boldsymbol{\beta}) y_i - b(\mathbf{x}_i^T \boldsymbol{\beta})\}$, the SMLE is defined as

$$\hat{\boldsymbol{\beta}}_{[k]} = \arg\max_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) \text{ subject to } ||\boldsymbol{\beta}||_0 \leq m, \tag{2.46}$$

where $|| \cdot ||_0$ denotes the number of nonzero components, and m is a pre-set value known to be greater than the true model size. This model provides the chance to find and explain the joint effects among features when screening.

To solve (2.46) and get $\hat{\boldsymbol{\beta}}_{[k]}$, Xu and Chen (2014) propose an iterative hard-thresholding algorithm (IHT) for GLM. First, define function $h_n(\boldsymbol{\gamma}; \boldsymbol{\beta})$ to approximate the log likelihood function $\ell_n(\cdot)$ at a neighbor of $\boldsymbol{\beta}$:

$$h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) - (\boldsymbol{\gamma} - \boldsymbol{\beta})^T S_n(\boldsymbol{\beta}) - \frac{u}{2} ||\boldsymbol{\gamma} - \boldsymbol{\beta}||_2^2, \tag{2.47}$$

where $S_n(\boldsymbol{\beta}) = \ell_n^{'}(\boldsymbol{\beta})$ is the score function and u a positive scaling parameter. Note that $h_n(\boldsymbol{\gamma}; \boldsymbol{\beta})$ well approximates $\ell_n(\boldsymbol{\beta})$ when $\boldsymbol{\gamma}$ is close to $\boldsymbol{\beta}$. Especially, when $\boldsymbol{\gamma} = \boldsymbol{\beta}$, $h_n(\boldsymbol{\beta}; \boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta})$. Therefore, we can obtain an approximate solution to (2.46) through the following iterative procedure:

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \arg\max_{\boldsymbol{\gamma}} h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}^{(t)}) & (2.48)\\ &= \arg\min_{\boldsymbol{\gamma}} ||\boldsymbol{\gamma} - \frac{1}{u}\{u\boldsymbol{\beta}^{(t)} + \mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{b}^{'}(\mathbf{x}\boldsymbol{\beta}^{(t)})\}||_2^2, & (2.49)\end{aligned}$$

subject to $||\boldsymbol{\gamma}||_0 \leq k$. The solution of $\boldsymbol{\gamma}$ is obtained by keeping the k components with largest absolute values of $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta}^{(t)} + u^{-1}\mathbf{x}^T\{\mathbf{y} - b^{'}(\mathbf{x}\boldsymbol{\beta}^{(t)})\}$. For each iteration, the regularization term $\frac{u}{2}||\boldsymbol{\gamma} - \boldsymbol{\beta}||_2^2$ in (2.47) penalizes the step size between $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\beta}^{(t)}$. The iteration will stop when $||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}||_2$ falls below the pre-set thresholding value. Furthremore, Xu and Chen (2014) prove that the objective function $\ell_n(\boldsymbol{\beta}(t))$ increases with t and certainly converges to a local maximum. Also, SMLE based screening method is shown to own the sure screening property.

2.2.4 Screening Methods for Additive Model

By extending the idea of SIS to nonparametric setting, Fan et al. (2011) propose nonparametric independence screening method for additive models through ranking the importance of predictors based on a measure of the goodness of fit of their marginal models. The nonparametric additive model is specified as follow

$$Y = \sum_{j=1}^{p} m_j(X_j) + \epsilon, \tag{2.50}$$

where $\{m_j(X_j)\}_{j=1}^p s$ are unknown smooth functions and ϵ have conditional mean 0. To identify important variables, we consider the following marginal non-parametric regression models for $j = 1, \ldots, p$:

$$\min_{m_j \in L_2(P)} E(Y - m_j(X_j))^2, \tag{2.51}$$

where P is the joint distribution of $(\mathbf{x}, Y)$ and $L_2(P)$ is the class of squares integrable functions under the measure P. The solution to the objective function (2.51) is $m_j(X_j) = E(Y|X_j)$, which is the projection of Y onto X_j. Thus, the marginal utility of X_j measured by $E[m_j^2(X_j)]$ can be used to rank the importance of predictors.

To estimate $E[m_j^2(X_j)]$, Fan et al. (2011) applied a normalized B-spline basis functions $\{\Psi_{jk}, k = 1, \ldots, d_n\}$ to approximate $\{m_j(X_j)\}_{j=1}^p$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$. Then, for any $m_{nj} \in \mathcal{S}_n, j = 1, \ldots, p$, we have

$$m_{nj}(x) = \sum_{k=1}^{d_n} \beta_{jk} \Psi_{jk}(x). \tag{2.52}$$

Given certain conditions, $m_j(X_j)$ can be well approximated by $m_{nj}(X_j)$. Therefore, the minimization problem (2.51) can be formulated as

$$\min_{m_{nj} \in S_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{nj}(X_{ij}))^2 = \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{d_n}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{\Psi}_{ij}^T \boldsymbol{\beta}_j)^2, \tag{2.53}$$

where $\boldsymbol{\Psi}_{ij} = (\Psi_1(X_{ij}), \ldots, \Psi_{d_n}(X_{ij}))^T$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^T$. Let $\boldsymbol{\Psi}_j = (\boldsymbol{\Psi}_{1j}, \cdots, \boldsymbol{\Psi}_{nj})^T$, the minimizer of (2.53) can be obtained as $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{\Psi}_j^T \boldsymbol{\Psi}_j)^1 \boldsymbol{\Psi}_j^T \mathbf{y}$. Thus $\hat{f}_{nj}(X_{ij}) = \boldsymbol{\Psi}_{ij}^T \hat{\boldsymbol{\beta}}_j$, and then $||\hat{f}_{nj}||_n^2 = \frac{1}{n} \sum_{i=1}^n \hat{f}_{nj}^2(X_{ij})$ can be used to estimate the marginal utility $E[m_j^2(X)]$.

Similar rule as (2.42) can be applied to select a submodel with $|\hat{\beta}_j^M|$ replaced by $||\hat{m}_{nj}||_n^2$. Note that it is also equivalent to rank the residual sum of squares of the marginal regression model, where $RSS_j = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_{nj}(X_{ij}))^2$. Fan et al. (2011) also point out that the marginal signal of the true predictors $\{E[m_j^2], j \in \mathcal{M}_0\}$ does not vanish. Under some regularity conditions, $||\hat{f}_{nj}||$ uniformly converge to $||f_{nj}||$ and the sure screening property hold.

2.2.5 Model-Free Feature Screening

All aforementioned model-based screening methods can be biased if the underlying model is incorrectly specified. To avoid this, some more general screening methods are studied without model specification.

Hall and Miller (2009) define a generalized empirical correlation between the response and predictors based on a vector space of functions $\mathcal{H}$ that include all constants and linear functions. Thus, a model-free feature screening submodel can be approached by ranking the generalized correlation. Given i.i.d. $\{Y_i, \mathbf{x}_i\}_{i=1}^n$, the generalized correlation is defined as

$$\psi_j = \sup_{h \in \mathcal{H}} \frac{cov[h(X_{1j}), Y_1]}{\sqrt{var[h(X_{1j})]var(Y_1)}}, \tag{2.54}$$

which can be estimated by

$$\hat{\psi}_j = \sup_{h \in \mathcal{H}} \frac{\sum_i [h(X_{ij}) - \bar{h}_j](Y_i - \bar{Y})}{\sqrt{\sum_i^n [h(X_{ij})^2 - \bar{h}_j^2](Y_i - \bar{Y})^2}}. \tag{2.55}$$

Here $\bar{h}_j = \frac{1}{n}\sum_i h(X_{ij})$. Furthermore, since it is quite challenging to calculate $\hat{\psi}_j$, Hall and Miller (2009) proved that the maximizer of $\hat{\psi}_j$ is equivalent to the minimizer in the following problem

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{n} [Y_i - h(X_{ij})]^2, \tag{2.56}$$

where $\mathcal{H}$ is a finite-dimensional function space. In practice, we impose certain restrictions on $\mathcal{H}$ for an explicit analytic solution. For example, $\hat{\psi}_j$ will be our familiar Pearson correlation if we constrain $\mathcal{H}$ to only linear functions. To determine the screened submodel, Hall and Miller (2009) adopt bootstrap procedure to decide the cutoff point for ranking. Details are refered to Hall and Miller (2009).

Although Pearson correlation is frequently used measure the strength of association between the response and covariates, it still has some inherent drawbacks. For example, Pearson correlation is very sensitive to the outliers and influential points. Furthermore, it is good at detecting linear association, while it fails to discover nonlinear relationship time to time. To deal with these drawbacks, Li et al. (2012) propose a robust rank correlation screening (RRCS) method by ranking Kendall τ correlation coefficient. Given pairs of data $\{Y_i, X_{ij}\}_{i=1}^n$, the marginal Kendall τ correlation coefficient between Y and X_j is defined as

$$\omega_j = \frac{1}{n(n-1)} \sum_{i \neq k}^{n} I(X_{ij} < X_{kj}) I(Y_i < Y_k) - \frac{1}{4}, j = 1, \ldots, p. \tag{2.57}$$

We can rank the importance of predictors by ranking the magnitudes of ω_j and get a submodel defined in the same way as (2.35).

RRCS outperforms the Pearson correlation based SIS in three different aspects. First, RRCS is robust with respect to the outliers and influential points. Second, the rank of ω_j is invariant under monotonic transformation. As a result, RRCS can discover nonlinear relationship and deal with semiparametric models without more complicated nonparametric estimation. Third, RRCS makes use of the ranking information to greatly simplify its theoretical derivation and achieve sure screening property with only a moment condition.

Distance correlation (DC) is studied by Székely et al. (2007) to measure the dependence between two random vectors. Based on that, Li et al. (2012) propose another completely model-free screening procedure. Denote $\phi_{\boldsymbol{\mu}}(\mathbf{t})$ and $\phi_{\boldsymbol{\nu}}(\mathbf{s})$ as the respective characteristic functions for the d_μ dimensional random vector $\boldsymbol{\mu}$ and d_ν dimensional random vector $\boldsymbol{\nu}$. $\phi_{\boldsymbol{\mu},\boldsymbol{\nu}}(\mathbf{t},\mathbf{s})$ is the joint characteristic function for $(\boldsymbol{\mu},\boldsymbol{\nu})$. The distance covariance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is defined as

$$dcov^2(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_{\mathbb{R}^{d_n+d_\nu}} ||\phi_{\boldsymbol{\mu},\boldsymbol{\nu}}(\mathbf{t},\mathbf{s}) - \phi_{\boldsymbol{\mu}}(\mathbf{t})\phi_{\boldsymbol{\nu}}(\mathbf{s})||^2 \omega(\mathbf{t},\mathbf{s}) d\mathbf{t} d\mathbf{s} \tag{2.58}$$

where $\omega(\mathbf{t},\mathbf{s}) = 1/[c_{d_\mu} c_{d_\nu} ||\mathbf{t}||_{d_\mu}^{1+d_\mu} ||\mathbf{s}||_{d_\nu}^{1+d_\nu}]$ with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$. Here, $||\mathbf{x}||_d$ represents the Euclidean norm for $\mathbf{x} \in \mathbb{R}^d$, and $||\psi||^2 = \phi\bar{\phi}$, where $\bar{\phi}$ is the conjugate of ϕ if ϕ is a complex valued function. Thus, the distance correlation (DC) between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is defined as

$$dcorr(\boldsymbol{\mu},\boldsymbol{\nu}) = \frac{dcov(\boldsymbol{\mu},\boldsymbol{\nu})}{dcov(\boldsymbol{\mu},\boldsymbol{\mu}) dcov(\boldsymbol{\nu},\boldsymbol{\nu})}. \tag{2.59}$$

Based on $dcorr(\boldsymbol{\mu},\boldsymbol{\nu})$, a marginal utility is naturally defined as $u_j = dcorr^2(X_j, y)$ and used for screening.

There are several advantages of using distance correlation to measure marginal association. First, both X_j and y can be multivariate with different dimensions, regardless of whether it is continuous, discrete or categorical. Therefore, DC-SIS can deal with multivariate response and groupwise predictors without model specification. Second, when X_j and y are normally distributed, their distance correlation is a strictly increasing function with respect to $|\rho|$, which means that SIS can be regarded as a special case of DC-SIS when the response and predictors are normally distributed. Finally, $dcorr(X_j, y) = 0$ if and only if $f(X_j)$ and y

are independent, where $f(\cdot)$ is a strictly monotone function. This feature allows exploration to nonlinear relationship between X_j and y, which is more powerful than traditional SIS developed based on linear model assumption.

2.3 Ultra-High Dimensional Survival Data Analysis

In this section, we first briefly introduce the basic definitions and concepts together with commonly used Cox model in survival analysis. Variable selection and feature screening procedures are then discussed.

2.3.1 Background and Definition

Survival analysis is introduced to analyze survival data. The response in survival data, which is referred as a event time, survival time or failure time, is usually continuous. However, the survival time may be incompletely determined for some subjects. If we have study dropout, we know that the survival time is larger than or equal to the dropout time t. For other subjects, we would know their exact survival time. We define these incompletely observed subjects as censored. For those subjects without censoring, we can apply the regular regression procedures. However, time to event is restricted to be positive and has a skewed distribution, which doesn't satisfy the normality assumption. Furthermore, in survival data analysis, we are more interested in the probability of surviving after a certain time point than the expected failure time. The hazard function, used for regression in survival analysis, can provide more insight for the failure mechanism.

For the censoring mechanism, it is assumed to be noninformative and caused by something other than the impending event throughout our discussion. Censoring might occur due to the following reasons:

1. A subject withdraws from the study;
2. A subject does not experience the event before the study ends;
3. A subject is lost to follow-up during the study period.

All three examples above are right-censoring, which commonly happens in the real life.

Notation: To represent the right-censored survival data, we introduce the terminology as follows. T_i denotes the event time for the i-th subject; C_i denotes the censoring time; and $\delta_i = I(T_i \leq C_i)$ is the event indicator. That is, $\delta_i = 1$ if events happen, while $\delta_i = 0$ if events are censored. Z_i is the observed response defined by $Z_i = \min(T_i, C_i)$.

Regarding the survival time T, survival analysis has some special ways to describe the probability distribution of T. The cumulative distribution function $F(t)$ is the probability of survival time $T \leq t$ and is defined by $F(t) = P(T \leq t)$. The corresponding pdf $f(t)$ is defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t), \tag{2.60}$$

In survival analysis, the probability of surviving at a given time t is more of interest. Therefore, the survival function $S(t)$ is defined as

$$S(t) = P(T > t) = 1 - F(t), \tag{2.61}$$

$S(t)$ is a non-increasing function of time t with $S(0) = 1$ and $S(\infty) = 0$.

The hazard function $h(t)$ is the rate of events occur at time t conditioning on no previous event

$$\begin{aligned} h(t) &= lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < T < t + \Delta t | T > t) & (2.62) \\ &= lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t < T < t + \Delta t)}{P(T \geq t)} = \frac{f(t)}{S(t)}. & (2.63) \end{aligned}$$

Based on $h(t)$, the cumulative hazard function $H(t)$, which describes the accumulated risk before time t, is defined as

$$H(t) = \int_0^t h(\mu) d\mu. \tag{2.64}$$

The relationships between these previous defined functions $f(t)$, $h(t)$, $S(t)$, and

$H(t)$ are as follow

$$f(t) = -\frac{dS(t)}{dt} \tag{2.65}$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\log(S(t))\} \tag{2.66}$$

$$H(t) = \int_0^t h(\mu)d\mu = \int_0^t -\frac{d}{d\mu}\{\log(S(\mu))\}d\mu = -\log(S(t)) \tag{2.67}$$

$$S(t) = \exp\{-H(t)\}, \tag{2.68}$$

Given a survival data, the survival or hazard function can then be estimated through an empirical way or a parametric model. For the first way, Kaplan-Meier estimator for the survival function and Nelson-Aalen estimator for cumulative hazard are developed to gives robust estimates with few assumptions. For the second way, we can specify a parametric model for hazard function $h(t)$ based on a particular density and use the maximum likelihood estimators(MLE) to estimate the unknown coefficients. However, this approach may be too restrictive to get inappropriate conclusions. In real life, it is natural to assume that the survival time can be affected by a vector of covariates. At the same time, some nonparametric unspecified baseline function changing with time t can be combined with the effect of covariates. As a result, the widely used Cox model will be introduced in the following part.

2.3.2 Cox Model in Survival Data Analysis

In this part, we consider survival model with right-censored observations. Their responses are the waiting time until the occurrence of an event, and the covariates whose effects on the survival time are of interest. Denote $\mathbf{x}_i$ as the covariates for i-th subject, where $\mathbf{x}_i$ can be continuous, discrete and time-varying. The goal of survival analysis is to model the effects of significant covariates given a survival data $\{(\mathbf{x}_i, Z_i, \delta_i)\}_n$.

The most popular framework for right-censored survival data is the Cox's proportional hazards model (Cox, 1972), where the hazard function $h(t|\mathbf{x})$ for a subject with covarties $\mathbf{x}$ is defined as

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}). \tag{2.69}$$

$h(t|\mathbf{x})$ contains the parametric term $\exp(\mathbf{x}^T\boldsymbol{\beta})$ and the non-parametric unspecified baseline hazard function $h_0(t)$ who serves as a reference group. Hence, Cox's proportional hazards model is a semi-parametric model. Under the proportional hazard assumption, the relative hazard rate between two groups is the same at all durations t. That is, for two subjects with fixed covariates, their relative risk only depends on their corresponding covariates. To integrate both sides of (2.69) from 0 to t, we obtain the cumulative hazard function

$$H(t|\mathbf{x}) = H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.70}$$

which is also proportional. Applying the relationship (2.68), the survival function $S(t|\mathbf{x})$ can be then determined uniquely by

$$S(t|\mathbf{x}) = \exp(-H(t|\mathbf{x})) = S_0(t)^{\exp(\mathbf{X}^T\boldsymbol{\beta})}, \tag{2.71}$$

where $S_0(t) = \exp(-H_0(t))$ is the baseline survival function. Thus, the effects of covariates $\mathbf{x}$ on the survival function is to raise it to a power given by relative risk $\exp(\mathbf{x}^T\boldsymbol{\beta})$.

2.3.3 Variable Selection in Cox's Survival Data Analysis

In this part, we review the variable selection techniques via penalization to survival analysis setting with right-censored data. Assume that subjects $\{(\mathbf{x}_i, Z_i, \delta_i)\}_n$ are i.i.d. and T_i and C_i are independent conditioning on $\mathbf{x}$, a full likelihood function can be written as

$$L = \prod_{\mu} f(Z_i|\mathbf{x}_i) \prod_{c} \bar{F}(Z_i|\mathbf{x}_i) = \prod_{\mu} h(Z_i|\mathbf{x}_i) \prod_{i=1}^{n} \bar{F}(Z_i|\mathbf{x}_i), \tag{2.72}$$

Here the subscript c and μ denote the censored and uncensored data. $f(Z_i|\mathbf{x}_i)$, $F(Z_i|\mathbf{x}_i)$ and $h(Z_i|\mathbf{x}_i)$ are the conditional density function, the conditional survival function and the conditional hazard function of T given $\mathbf{x}$, respectively. Denote $t_1 < \ldots < t_N$ as the ordered observed failure times. Thus, the covariates associated with the N failures are $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$. R_j is the risk set right before time t_j, namely, $R_j = \{i : Z_i \geq t_j\}$.

Based on the Cox model assumption, $h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta})$, the likelihood function in (2.72) becomes

$$L(h_0(t), \boldsymbol{\beta}) = \prod_{\mu} h_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\prod_{i=1}^{n}\exp\{-H_0(Z_i)exp(\mathbf{x}_i^T\boldsymbol{\beta})\}. \tag{2.73}$$

Here $H_0(\cdot)$ is the cumulative baseline hazard function. Usually, the estimate of $\boldsymbol{\beta}$ is what of interest and $h_0(t)$ is treated as a nuisance parameter. Following Breslow's idea, we can estimate $h_0(t)$ by maximizing the profiled likelihood $L(h_0(t), \boldsymbol{\beta})$ conditioning on the estimated $\hat{\boldsymbol{\beta}}$. Substituting this profiled estimate $h_0(t)$ into (2.73) and take log-transformation, we get the resulting log-likelihood function that depends only on $\boldsymbol{\beta}$

$$\ell(\boldsymbol{\beta}) = \sum_{j}^{N}[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}], \tag{2.74}$$

which is the partial likelihood function proposed by Cox (1975). Based on (2.74), penalized maximum partial likelihood estimator can be used to select significant covariates. The penalized partial likelihood is defined as

$$Q(\boldsymbol{\beta}) = \sum_{j}^{N}[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}] - n\sum_{j=1}^{d}p_\lambda(|\beta_j|). \tag{2.75}$$

By choosing a proper tuning parameter λ, we can set many of the estimated coefficients to zero and hence achieve the objectives of variable selection.

Fan and Li (2002) apply the SCAD penalty functions to solve the variable selection problem. It shows that the SCAD penalty enjoys the oracle property. To solve the nonconvex optimization problem, Fan and Li (2002) use the LQA algorithm to get a local quadratic one. Given a good initial estimate, the modified Newton-Raphson algorithm achieves an efficient penalized partial likelihood estimator. Furthermore, the author use a sandwich formula to do statistical inference. For the determination of λ, they select the optimal λ via minimizing an approximate generalized cross validation (GCV) statistic.

Zhang and Lu (2007) apply the adaptive LASSO penalty to partial likelihood for Cox's model with noninformative censoring mechanism, which avoids the in-consistency of the LASSO and the numerical complexity of the SCAD. Weights of the adaptive LASSO penalty are determined by unpenalized estimator. Similarly, they choose GCV statistic as their tuning parameter selector.

2.3.4 Feature Screening in Ultra-High Dimensional Survival Data Analysis

Similar to the case in generalized linear model, the penalized variable selection methods for the Cox model work not that good in the case of ultrahigh dimensional survival problems. To solve this, Fan et al. (2010) extend SIS and ISIS methods in GLM to Cox model and employ maximum of the partial likelihood depending on k-th covariate as k-th marginal untility μ_k, which is defined as

$$\mu_k = \max_{\beta_k}(\sum_{i=1}^{n} \delta_i x_{ik}\beta_k - \sum_{i=1}^{n} \delta_i \log\{ \sum_{j \in R(z_i)} \exp(x_{jk}\beta_k)\}), \tag{2.76}$$

where $R(t) = \{j : Z_j \geq t\}$ and x_{jk} is the k-th predictor for subject j. Therefore, the higher the marginal utility μ_k, the more information the k-th covariate contains for the survival time. By ranking all the marginal utilities from the largest to smallest, Fan et al. (2010) choose the top $d = [n/\log(n)]$ covariates and filter unimportant variables. Denote $\hat{\mathcal{I}}_1$ as the index set of these d covariates that have been selected. Based on the selected d covariates, the variable selection technique can be applied to the selected subset of the variables $\{X_j, j \in \hat{\mathcal{I}}_1\}$ to further refine the model. Mathematically, we solve the following penalized partial likelihood problem:

$$\min_{\boldsymbol{\beta}_{\hat{\mathcal{I}}_1}}[-\sum_{i=1}^{n} \delta_i \mathbf{x}_{\hat{\mathcal{I}}_1,i}^T \boldsymbol{\beta}_{\hat{\mathcal{I}}_1} + \sum_{i=1}^{n} \delta_i \log\{ \sum_{j \in \mathcal{R}(y_i)} \exp(\mathbf{x}_{\hat{\mathcal{I}}_1,i} \boldsymbol{\beta}_{\hat{\mathcal{I}}_1})\} + \sum_{m \in \hat{\mathcal{I}}_1} p_\lambda(\beta_m)], \tag{2.77}$$

where $\boldsymbol{\beta}_{\hat{\mathcal{I}}_1}$ and $\mathbf{x}_{\hat{\mathcal{I}}_1,i}$ denotes a sub-vector of $\boldsymbol{\beta}$ and $\mathbf{x}_i$ with indices in $\hat{\mathcal{I}}_1$. Optimization problem (2.77) will lead to sparse estimate $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{I}}_1}$ with the index set of nonzero components $\hat{\mathcal{M}}_1$. $\hat{\mathcal{M}}_1$ will serve as our final estimate of $\mathcal{M}^*$ in Cox-SIS two step procedures.

As Fan and Lv (2008) point out, SIS can fail for some challenging scenarios under which we will miss the jointly related but marginally unrelated covariates.

To deal with these scenarios, Fan et al. (2010) propose iterative SIS (ISIS). Instead of employing only marginal information, ISIS tries to make more use of joint covariates' information. Based on $\hat{\mathcal{M}}_1$, Fan et al. (2010) next define the conditional utility of m-th covariate that is not in $\hat{\mathcal{M}}_1$ as follows:

$$\mu_{m|\hat{\mathcal{M}}_1} = \max_{\beta_m, \boldsymbol{\beta}_{\hat{\mathcal{M}}_1}} \sum_{i=1}^{n} \delta_i [(x_{im}\beta_m + \mathbf{x}^T_{\hat{\mathcal{M}}_1,i}\boldsymbol{\beta}_{\hat{\mathcal{M}}_1}) - \log\{\sum_{j\in\mathcal{R}(y_i)} \exp(x_{jm}\beta_m + \mathbf{x}^T_{\hat{\mathcal{M}}_1,j}\boldsymbol{\beta}_{\hat{\mathcal{M}}_1})\}], \tag{2.78}$$

This conditional utility measures the additional contribution of the m-th covariate given that all covariates in $\hat{\mathcal{M}}_1$ have been included in the model. By ranking the conditional utilities whose indices are not in $\hat{\mathcal{M}}_1$ from the largest to the smallest, we select the top ranking covariates with index set $\hat{\mathcal{I}}_2$. Based on the union index set $\hat{\mathcal{M}}_1 \cup \hat{\mathcal{I}}_2$, we apply (2.77) again and get the index set $\hat{\mathcal{M}}_2$ of sparse parameter estimates, which is our updated estimate of the true index set $\mathcal{M}^*$. Repeat the above iteration until some convergence criterion such as $\hat{\mathcal{M}}_j = \hat{\mathcal{M}}_{j-1}$ is reached.

Furthermore, Fan et al. (2010) introduce two variants of iterated Cox-SIS to reduce false selected rates(FSR). However, Cox-SIS still has two major problems. Since censoring is confounding between the covariates and the survival time, it is difficult to extend the sure screening property to Cox model. Besides that, Cox-SIS only works well when the true underlying model is a Cox model. Otherwise, its power is very limited.

The aforementioned screening procedures usually dictate the number of variables to retain. There is no principled evaluation criterion for the methods of making such a choice. To get a more theoretical solid thresholding criterion, Zhao and Li (2012) provide a new, principled method for choosing the number of screened covariates based on specifying the desired false positive rate. They solve $\hat{\beta}_k$ marginally by

$$\hat{\beta}_k = \arg\max_{\beta_k}(\sum_{i=1}^{n} \delta_i x_{ik}\beta_k - \sum_{i=1}^{n} \delta_i \log\{\sum_{j\in\mathcal{R}(z_i)} \exp(x_{jk}\beta_k)\}). \tag{2.79}$$

Denote $I_k(\beta_k)$ as the information matrix at $\hat{\beta}_k$. They rank $I_k^{\frac{1}{2}}(\beta_k)|\beta_k|$ and conclude the index set of the screened model:

$$\hat{\mathcal{M}}_\gamma = \{1 \leq k \leq p : I_k^{\frac{1}{2}}(\beta_k)|\beta_k| > \gamma\}, \tag{2.80}$$

where $\gamma = \Phi^{-1}(1 - \frac{f}{2p})$. Here $\Phi(\cdot)$ is the standard normal cumulative distribution function and f is the number of false positives that we are willing to tolerate. Thus, the expected false positive rate will be expected to be $f/p = 2(1 - \Phi(\gamma))$. This method is named as principled Coxs sure independence screening procedure (PSIS), where the cutoff γ is selected to control the false positive rate. The algorithm of PSIS is illustrated as follows:

1. Fit a marginal Cox model for each of the covariates based on (2.79) to get estimates $\hat{\beta}_k$ and their corresponding variance estimates $I_k^{-1}(\hat{\beta}_k)$.
2. Specify the number of variables f, fix the false positive rate as f/p and determine γ by $\gamma = \Phi^{-1}(1 - \frac{f}{2p})$.
3. Get the screened model with the index set $\hat{\mathcal{M}}_\gamma = \{1 \leq k \leq p : I_k^{\frac{1}{2}}(\beta_k)|\beta_k| > \gamma\}$

Furthremore, Zhao and Li (2012) give the first theoretical justifications of the sure independence screening procedure for censored data. Under the asymptotic framework where the number of predictors p can be regarded as a function of sample size N, the author show that PSIS will select all of the important variables with probability going to 1. At the same time, the false positive rate is close to the prespecified level $2(1 - \Phi(\gamma))$.

Other than just considering the marginal effect of covariates, screening method of joint effects of features is promising for ultrahigh dimensional survival data. Yang et al. (2016) propose a two-stages sure joint screening procedure for Cox model based on the constrained version of partial likelihood function:

$$\hat{\boldsymbol{\beta}}_m = \arg\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta})(= \sum_{j=1}^{N}[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\{\sum_{j \in R_j} \exp(\mathbf{x}_i\boldsymbol{\beta})\}]), \text{ subject to } ||\boldsymbol{\beta}||_0 \leq m, \tag{2.81}$$

where m is a pre-specified thresholding value assumed to be greater than the number of nonzero elements of $\boldsymbol{\beta}^*$. For high-dimensional problems, it is difficult to solve the constrained maximization problem (2.81) directly. Alternatively, Yang et al. (2016) consider a proxy of the partial likelihood function by applying Taylor expansion to $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ in a neighbor of $\boldsymbol{\gamma}$,

$$\ell_p(\boldsymbol{\gamma}) \approx \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p''(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}). \tag{2.82}$$

Here $\ell_p'(\boldsymbol{\beta}) = \partial \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell_p''(\boldsymbol{\beta}) = \partial^2 \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. When $\ell_p''(\boldsymbol{\beta})$ is invertible, the computational complexity of calculating the inverse of $\ell_p''(\boldsymbol{\beta})$ is $O(p^3)$. Therefore, the computational costs will be unacceptable when p is large. Furthermore, $\ell_p''(\boldsymbol{\beta})$ is even not invertible when $p > n$. To save computational costs and deal with singularity of the Hessian matrix, we use $u\text{diag}\{\ell_p''(\boldsymbol{\beta})\}$ to approximate $\ell_p''(\boldsymbol{\beta})$,

$$g(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T W (\boldsymbol{\gamma} - \boldsymbol{\beta}), \tag{2.83}$$

where u is a scaling constant to be specified and $W = \text{diag}\{-\ell_p''(\boldsymbol{\beta})\}$ is a diagonal matrix consisting of the diagonal elements of $-\ell_p''(\boldsymbol{\beta})$. When $\boldsymbol{\gamma} = \boldsymbol{\beta}$, we have $g(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta})$. Yang et al. (2016) point out that $g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$ under some conditions, which ensures the ascent property. Furthermore, since W is a diagonal matrix, $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j for any given $\boldsymbol{\beta}$, which enables us to have a closed form solution of the maximization problem

$$\max_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma}|\boldsymbol{\beta}), \text{ subject to } ||\boldsymbol{\gamma}||_0 \leq m, \tag{2.84}$$

for a given $\boldsymbol{\beta}$ and m. Easy to verify that the maximizer of $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + u^{-1} W^{-1} \ell_p'(\boldsymbol{\beta})$. Define $r_j = \omega_j \tilde{\gamma}_j^2$ as the j-th utility measure with ω_j the j-th diagonal element of W for $j = 1, \ldots, p$. By ranking r_j so that $|r_{(1)}| \geq |r_{(2)}| \geq \ldots \geq |r_{(p)}|$, the solution to (2.84) will be

$$\hat{\gamma}_j = \tilde{\gamma}_j I\{|r_j| \geq |r_{(m+1)}|\} \hat{=} H(\tilde{\gamma}_j; m). \tag{2.85}$$

Thus, we can effectively screen features through the following algorithm:

Step1. Set the initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$.

Step2. Set $t = 0, 1, 2, \ldots$ and iteratively conduct **Step 3** and the **Step 4** until the algorithm converges.

Step3. Calculate $\tilde{\boldsymbol{\gamma}}^{(t)} = (\tilde{\gamma}_1^{(t)}, \ldots, \tilde{\gamma}_p^{(t)})^T = \boldsymbol{\beta}^{(t)} + u_t^{-1} W^{-1}(\boldsymbol{\beta}^{(t)}) \ell_p'(\boldsymbol{\beta}^{(t)})$ and

$$\tilde{\boldsymbol{\beta}}^{(t)} = (H(\tilde{\gamma}_1^{(t)}; m), \ldots, H(\tilde{\gamma}_p^{(t)}; m))^T \hat{=} \mathbf{H}(\tilde{\boldsymbol{\gamma}}^{(t)}; m) \tag{2.86}$$

Denote $S_t = \{j : \tilde{\beta}_j^{(t)} \neq 0\}$ as the nonzero index of $\tilde{\boldsymbol{\beta}}^{(t)}$

Step4. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\beta_1^{(t+1)}, \ldots, \beta_p^{(t+1)})^T$ as follows. If $j \notin S_t$, set $\beta_j^{(t+1)} = 0$; otherwise, set $\{\beta_j^{(t+1)} : j \in S_t\}$ to be the maximum partial likelihood estimate of the submodel S_t.

2.4 Varying Coefficient Model

In our statistical toolkit, parametric models such as linear regression model and generalized linear model are the most useful tools. However, strict assumptions about the relationship between the response and the covariates sometimes impose restriction on their applications. As a result, the dynamic features existing in data from various scientific areas often cannot be appropriately modeled through parametric model. Among many attempts to understand the dynamic features, varying coefficient model allows the regression coefficients to depend on certain factors, which help to increase model flexibility and incorporate dynamics.

2.4.1 Model Structure and Penalized Least Squares

In Hastie and Tibshirani (1993), the varying coefficient model is systematically studied and defined as

$$y = \beta_0 + X_1\beta_1(R_1) + \ldots + X_p\beta_p(R_p) + \epsilon = \eta + \epsilon, \tag{2.87}$$

where y is the response with mean η, $\mathbf{x} = (X_1, \ldots, X_p)$ is a p-dimensional predictor, $\mathbf{R} = (R_1, \ldots, R_p)$ is the "effect modifiers" smoothly changing coefficient functions $\boldsymbol{\beta}(\mathbf{R}) = (\beta_1(R_1), \ldots, \beta_p(R_p))$, and ϵ is the random error with $E(\epsilon|\mathbf{x}, \boldsymbol{\beta}(\mathbf{R})) = 0$. Here, $\boldsymbol{\beta}(\mathbf{R})$ are estimated by nonparametric approach. Model (4.7) can be extended to generalized linear model framework, by assuming

$$E(y|\mathbf{x}, U) = g^{-1}(\mathbf{x}^T\boldsymbol{\beta}(\mathbf{R})), \tag{2.88}$$

where $g^{-1}(\cdot)$ is the inverse of the link function $g(\cdot)$.

Model (4.8) show the common structure of various models. Some instances of model (4.8) listed below will be familiar by imposing different constraint conditions on $\beta_j(R_j)$

1. If $\beta_j(\mathbf{R}_j) = j$(the constant function), then that term is linear in X_j. If all the terms are linear, then model (4.8) is the usual linear model or generalized linear model.

2. If $X_j = c$, then the jth term is simply, $\beta_j(\mathbf{R}_j) = j$, an unspecified function in $\mathbf{R}_j$. If all the terms have this form or are linear, then model (4.8) has the form of a generalized additive model.

3. Often the $\mathbf{R}_j$ s will be the same variable that suspect could modify the effects of $X_1, \ldots, X_p$. Suppose, for example, the data consist of repeated measurements over n time points $t \in (t_1, \ldots, t_n)$. Then we might model this as

$$\eta_t = \beta_0(t) + X_1(t)\beta_1(t) + \ldots + X_p(t)\beta_p(t), \tag{2.89}$$

4. Each R_j can be scalar or vector.

Model (4.7) as it stands is too general for most applications, in that no restrictions are imposed on the coefficient functions. For that reason we impose restrictions of one form or another on the coefficient functions. The parametric approaches such as polynomial functions do not provide enough flexibility and local adaptiveness. As a result, a set of regression spline bases with a fixed arrangement of knots is likely to be preferable. Furthermore, all the standard inferential tools can be used to evaluate sets of coefficients with this approach. However, the characteristics of the fitted curves can be quite different with minor changes in the positions of the knots. To solve that, Hastie and Tibshirani (1993) present a general nonparametric procedure for the varying-coefficients model (4.7) based on a penalized least squares criterion to consider measuring the goodness of fit and penalizing the roughness of each β_j as a whole.

Estimation in L_2: Suppose that we decide to estimate $\beta_1(\cdot), \ldots, \beta_p(\cdot)$ in model (4.7) by minimizing

$$E\{Y - \sum_{j=1}^{p} X_j \beta_j(\mathbf{R}_j)\}^2, \tag{2.90}$$

Conditioning on each R_j, a sufficient condition for the solutions is

$$E[X_j\{Y - \sum_{j=1}^{p} X_j \beta_j(\mathbf{R}_j)\}|R_j] = 0, j = 1, 2, \ldots, p. \tag{2.91}$$

To find $\beta_j(\cdot)$, we can rearrange the above equation and solve

$$\beta_j(R_j) = \frac{E[X_j^2\{Y - \sum_{k \neq j} X_k \beta_k(\mathbf{R}_k)\}/X_j|R_j]}{E(X_j^2|R_j)}, \tag{2.92}$$

Since a scatterplot smoother can be viewed as a flexible estimate of a conditional expectation, this suggests that each function $\beta_j(R_j)$ can be estimated in an iterative 'one at a time' manner by smoothing $\{Y - \sum_{k \neq j} X_k \beta_k(\mathbf{R}_k)\}/X_j$ on R_j, with weights X_j^2.

Penalized Least Squares: To solve the sensitivity issue with resepct to minor changes in the knots positions, a penality term is added to L_2 loss to penalize the roughness of each $\beta_j(R_j)$ with a fixed parameter λ. As a result, we propose to minimize the penalized least squares criterion

$$\mathbf{J}(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n}\{y_i - \sum_{j=1}^{p} x_{ij}\beta_j(r_{ij})\}^2 + \sum_{j=1}^{p} \lambda_j \int \beta_j^{''}(r_j)^2 dr_j, \tag{2.93}$$

For smoothing splines of β_j, one might consider the use of the weighted cubic smoothing spline, a locally weighted running line smoother or for time varying coefficients an exponentially weighted moving average in the iterative procedure above.

In statistical literature, two different estimation methods for $\beta_k(t)$ in (4.8) are widely used. One is kernel-local polynomial smoothing proposed in Hoover et al. (1998), Fan and Zhang (1999) and Fan and Zhang (2008), etc. The other one is polynomial spline introduced in Huang et al.(2002, 2004) and Huang and Shen (2004), etc. We briefly introduce them in the following two sections.

2.4.2 Polynomial Splines

Polynomial splines are piecewise polynomials with pieces jointing smoothly at a set of interior knots under certain continuity and derivatives conditions. The knots are denoted by $\xi_0 < \xi_1 < \ldots < \xi_L < \xi_{L+1}$ with two end points ξ_0 and ξ_{L+1} of the interval on T. A spline of degree $d = 0, 1, 2, 3$ corresponds to, respectively, a piecewise constant, linear, quadratic or cubic spline on each of the intervals $[\xi_l, \xi_{l+1})$, $0 \leq l \leq L-1$ and $[\xi_L, \xi_{L+1}]$, and globally has $d-1$ continuous derivatives for $d \geq 1$. Among them, cubic spline is the most commonly used. Therefore, the user-determined parameters include:

(a) The degree of the spline function, d;

(b) The number of knots, L;

(c) The positions of the interior knots, $\{\xi_l, l = 1, \ldots, L\}$;

(d) The number of free coefficients, i.e. degree of freedom of the spline function, $M = L + d + 1$.

Two out of (a), (b) and (d) are needed to create the spline basis.

Among different options of spline basis, we introduce Basis spline, or B-spline, which has the minimal support with respect to a given degree, smoothness and knots positions. De Boor et al. (1978) and Schumaker (1981) introduce the detailed construction and good properties of B-spline. Consider a varying-coefficient model based on a single effect modifying factor t. Let $\{B_{km}, m = 1, \ldots, M_k\}$ be a B-spline basis, then $\beta_k(t)$ can be approximated by

$$\beta_k(t) \approx \sum_{m=1}^{M_k} \gamma_{km} B_{km}(t), k = 0, 1, \ldots, p. \tag{2.94}$$

where M_k, the number of basis functions for $\beta_k(t)$, can be different for different k. Larger M_k leads to more accurate approximations of the varying coefficient functions but results in higher variance. Therefore, model (4.8) becomes, approximately, a linear regression model (Huang et al. 2004):

$$y_i(t) \approx \sum_{k=0}^{p} \sum_{m=1}^{M_k} \gamma_{km} B_{km}(t) x_{ik}(t) + \epsilon_i(t), \tag{2.95}$$

where $x_{i0}(t) \equiv 1$ for any i. Coefficients $\{\gamma_{km}, k = 1, \ldots, p; m = 1, \ldots, M_k\}$ can be estimated by minimizing the weighted least squares,

$$\sum_{i=1}^{n} \omega_i \{y_i(t) - \sum_{k=0}^{p} \sum_{m=1}^{M_k} \gamma_{km} B_{km}(t) x_{ik}(t)\}^2 = \sum_{i=1}^{n} \omega_i \{y_i(t) - \mathbf{z}_i^T \boldsymbol{\gamma}\}^2, \tag{2.96}$$

where ω_i is the weight for the i-th subject. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ with $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kM_k})^T$,

$$\mathbf{b}(t) = \begin{pmatrix} B_{01}(t) & \cdots & B_{0M_0}(t) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_{p1}(t) & \cdots & B_{pM_p}(t) \end{pmatrix}$$

$\mathbf{z}_i^T = \mathbf{x}_i^T(t)\mathbf{b}(t)$. Then, the minimizer of (2.96) becomes

$$\hat{\boldsymbol{\gamma}} = (\sum_{i=1}^{n} \omega_i \mathbf{z}_i^T \mathbf{z}_i)^{-1} \sum_{i=1}^{n} \omega_i y_i \mathbf{z}_i, \tag{2.97}$$

Hence, $\beta_k(t)$ can be estimated by $\hat{\beta}_k(t) = \sum_m \hat{\gamma}_{km} B_{km}(t)$.

We usually don't select all three parameters: the degree of splines, the number of basis functions and the locations of knots due to computational complexity. Huang et al. (2004) propose to use equally spaced knots and fixed degree, and only select the number of basis functions M by leave-one-out cross-validation (LooCV). This technique is also supported by Rice and Silverman (1991), Hansen et al. (1993) and Hoover et al. (1998). Denote $\hat{\beta}^{(-i)}(t)$ as the spline estimator obtained from all data without i-th subject. The cross-validation criterion is defined as

$$CV = \sum_{i=1}^{n} \{\omega_i (y_i(t) - \mathbf{x}_i^T(t) \hat{\boldsymbol{\beta}}^{(-i)}(t))^2\}. \tag{2.98}$$

Therefore, $\{M_k, k = 0, \ldots, p\}$ is obtained by minimizing this cross-validation score

(2.98). When the sample size N is large, one can also use the "K-fold" cross-validation to reduce the computational complexity.

2.4.3 Kernel-local Polynomial Smoothing

For convenience, assume the varying coefficient model has varying coefficients depending on time t. At each given time point, it can be considered as a linear model. Thus, it is reasonable to estimate the coefficients using data from a local neighborhood. That is the idea of kernel-local polynomial smoothing. The coefficient functions $\{\beta_k(t), k = 1, \ldots, p\}$ are approximated locally by

$$\beta_k(t) \approx \beta_k(t_0) + \beta_k^{'}(t_0)(t - t_0) \equiv a_k + b_k(t - t_0) \tag{2.99}$$

for any t in a neighborhood of t_0. a_k and b_k can be estimated by minimizing the weighted least squares

$$\sum_{i=1}^{n} [y_i(t_i) - \sum_{k=1}^{p} \{a_k + b_k(t_i - t_0)\}\mathbf{x}_i(t_i)]^2 K_h(t_i - t_0), \tag{2.100}$$

where $K_h(t) = K(t/h)/h$. $K(t)$ is a kernel function with bandwidth h. Kernel functions such as Gaussian kernel ($K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$), uniform kernel ($K(t) = I(|t| < 1/2)$), and Epanechikov kernel($K(t) = 0.75(1 - t^2)_+, t \in [-1, 1]$) are frequently used in practice. Technical details for matrix form of the solution are referred to Fan et al. (1999). Bandwidth parameter h is the only tuning parameter to control the extent of smoothing. Therefore, the choice of h is essential. The aforementioned LooCV and "K-fold" CV can be applied to choose h, but it might cause heavy computation. Another popular criterion for choosing h is the Mean Squared Error (MSE), which can be decomposed into the summation of variance and the squared bias. More details of estimation procedures for bias and variance are referred to Fan and Zhang (1999).

2.4.4 Feature Screening for Varying Coefficient Model

The aforementioned screening procedures screen predictors under the constant coefficients assumption. Liu et al. (2014) develop a kernel-regression based screening method specifically for ultrahigh dimensional varying coefficient models to reduce dimensionality. Suppose the varying-coefficients are functions of u. Thus, conditioning on u, the varying coefficient models are linear models. Therefore, it is natural to employ the conditional Pearson correlation coefficients $\rho(X_j, Y|u)$ between Y and $X_j's$ as a measure for the strength of association, where the $\rho(X_j, Y|u)$ and corresponding estimate $\hat{\rho}(X_j, Y|u)$ are defined as

$$\rho(X_j, Y|u) = \frac{\mathrm{cov}(X_j, Y|u)}{\sqrt{var(X_j|u)var(Y|u)}} \text{ , } \hat{\rho}(X_j, Y|u) = \frac{\hat{\mathrm{cov}}(X_j, Y|u)}{\sqrt{v\hat{a}r(X_j|u)v\hat{a}r(Y|u)}} \tag{2.101}$$

To estimate $\hat{\rho}(X_j, Y|u)$, Liu et al. (2014) apply kernel smoothing method to estimate the five conditional means involved: $E(X_j|u)$, $E(Y|u)$, $E(X_j^2|u)$, $E(Y^2|u)$ and $E(X_jY|u)$, which are assumed as nonparametric smoothing functions of u. Let $K(t)$ be a kernel function and $K_h(t) = K(t/h)/h$ with a bandwidth h. Then the kernel regression estimate for $E(Y|u)$ is

$$\hat{E}(Y|u) = \sum_{i=1}^{n} \frac{K_h(u_i - u)Y_i}{\sum_{i=1}^{n} K_h(u_i - u)}. \tag{2.102}$$

Estimates for the other four conditional means can be similarly defined. Thus, $c\hat{o}v(X_j, Y|u)$, $v\hat{a}r(X_j|u)$ and $v\hat{a}r(Y|u)$ can be obtained, and then $\hat{\rho}(X_j, Y|u)$. Based on the observed i.i.d. data $\{y_i, u_i, x_{ij}\}_{i=1}^{n}$, the utility measure of j-th predictors can be defined as $r_j = E[\rho^2(X_j, Y|u)]$. Its corresponding sample estimate is $\hat{r}_j = \sum_{i=1}^{n} \hat{\rho}^2(x_{ij}, y_i|u_i)/n$. Then,the screened submodel is defined as

$$\hat{\mathcal{M}} = \{j : 1 \leq j \leq p : \hat{r}_j \text{ ranks among first } d\}. \tag{2.103}$$

Fan and Lv (2008) suggest to set $d = [n/\log(n)]$. However, as we know, the effective sample size is nh in the kernel regression setting, and the optimal rate of bandwidth is $h = O(n^(- 1/5))$. Thus, Liu et al. (2014) suggest using $d = [n^{4/5}/log(n^{4/5})]$. Under certain regularity conditions, this method has both ranking consistency and sure screening properties, where the former states that the ranks of the true predictors are consistently higher than the ranks of the unimportant predictors.

Another screening method based on B-spline is developed in Fan et al. (2014), where the author extend the NIS procedure proposed in Fan et al. (2011) to a screening procedure for sparse varying coefficient models in ultrahigh dimension with the following form:

$$Y = \beta_0(u) + \sum_{j=1}^{p} \beta_j(u) X_j + \epsilon, \tag{2.104}$$

where u is some observable factors and $\{\beta_j(\cdot)\}_{j=0}^{p}$ are unknown smooth functions. Assuming that $\boldsymbol{\beta}(u) = (\beta_1(u), \ldots, \beta_p(u))^T$ is sparse, the index set of nonzero components is defined as $\mathcal{M}_0 = \{1 \leq j \leq p : E[\beta_j^2(u)] > 0\}$.

Given u, the marginal strength of each predictor can be measured by the expected conditional correlation between Y and X_j. For X_j, consider the marginal regression model:

$$\min_{\alpha_j(u), \beta_j(u) \in L_2(P)} E[(Y - \alpha_j(u) - \beta_j(u) X_j)^2 | u], \tag{2.105}$$

where P denotes the joint distribution of $(Y, u, \mathbf{X})$ and $L_2(P)$ is the class of square integrable functions under the measure P. The minimizer of (2.105) is

$$\alpha_j(u) = \frac{\text{cov}(X_j, Y|u)}{var(X_j|u)}, \beta_j(u) = E(Y|u) - \alpha_j(u) E(X_j|u) \tag{2.106}$$

Thus, the utility of X_j is defined as

$$\mu_j = E[\alpha_j(u) + \beta_j(u) X_j]^2 - [E(Y|u)]^2 = E\{\frac{[\text{cov}(X_j, Y|u)]^2}{var(X_j|u)}\}, \tag{2.107}$$

Let $E[\beta_0(u)] = E(Y|u)$. To estimate μ_j, similar technique used by Fan et al. (2011) is applied to approximate unknown coefficients $\{\alpha_j(u)\}_{j=1}^{p}$ and $\{\beta_j(u)\}_{j=0}^{p}$ by B-splines functions:

$$\alpha_j(u) \approx \sum_{k=1}^{d_n} \eta_{jk} \Psi_{jk}(u) \text{ and } \beta_j(u) \approx \sum_{k=1}^{d_n} \theta_{jk} \Psi_{jk}(u) \tag{2.108}$$

Then, $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jd_n})^T$ and $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jd_n})^T$ can be estimated by minimizing the ordinary least squares:

$$\min_{\boldsymbol{\eta}_j,\boldsymbol{\theta}_j} \frac{1}{n}\sum_{i=1}^{n}[Y_i - \boldsymbol{\Psi}_j(u_i)\boldsymbol{\eta}_j - \boldsymbol{\Psi}_j(u_i)\boldsymbol{\theta}_j]^2, \tag{2.109}$$

where $\boldsymbol{\Psi}_j(u_i) = (\Psi_1(u_i), \ldots, \Psi_{d_n}(u_i))^T$. Then a sample estimate of the marginal utility μ_j can be obtained by

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}[\hat{\alpha}_j(u_i) + \hat{\beta}_j(u_i)X_{ij}]^2 - \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_0(u_i)]^2 \tag{2.110}$$

where $\hat{\alpha}_j$, $\hat{\beta}_j$ and $\hat{\beta}_0$ are the estimates of α_j, β_j and β_0 using LSE from (2.109). Another equivalent measure of marginal strength is the residual sum of squares of the marginal regression model, which can be calculated by

$$\hat{\nu}_j = \sum_{i=1}^{n}[Y_i - \hat{\alpha}_j(u_i) - \hat{\beta}_j(u_i)X_{ij}]^2. \tag{2.111}$$

By properly choosing thresholds τ_n or ν_n, a submodel can be defined by

$$\mathcal{M}_{\tau_n,\nu_n} = \{1 \le j \le p : \hat{\mu}_j \ge \tau_n\} = \{1 \le j \le p : \hat{v}_j \ge \nu_n\} \tag{2.112}$$

Under certain regularity conditions, the sure screening property holds.

2.4.5 Cox Model with Varying-Coefficients

Cox model proposed by Cox (1972) is widely used to model the relationships between the survival time and the time-invariant covariates. Under the conventional form of Cox model, we assume the coefficients are constant function of t, thus guaranteeing the hazard is proportional across time. However, the constant assumption may fail to detect some important covariates when their effects change over time. To deal with this issue, Zucker and Karr (1990) propose the Cox model with time-varying covariate effects

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp\{\boldsymbol{\beta}(t)^T\boldsymbol{x}\}. \tag{2.113}$$

Here $\boldsymbol{\beta}(t)$ are assumed to be time-varying smooth coefficient functions that need to be estimated nonparametrically. The author propose to estimate $\boldsymbol{\beta}(t)$ using penalized partial likelihood. The weak uniform consistency and pointwise asymptotic normality of the estimators are also derived in Zucker and Karr (1990) under certain regularity conditions. Furthermore, Hastie and Tibshirani (1993) develop an algorithm based on smoothing spline basis to maximize the penalized partial likelihood as an extension. Gray (1992) proposes to use smoothing splines to estimate $\boldsymbol{\beta}(t)$ and give corresponding test statistics.

In recent statistical literatures, local partial likelihood is frequently used to estimate (2.113). Cai and Sun (2003) obtain the the pointwise confidence bands and show the pointwise asymptotic properties of the kernel estimators. This procedure is used to measure the time-dependencies or departure from the Cox proportional hazard models. Tian et al. (2005) constructed confidence bands for the kernel estimators and compared them with the pointwise confidence bands in Cai and Sun (2003). Sun et al. (2009) develop empirical likelihood pointwise and use local partial likelihood smoothing to produce simultaneous confidence bands for the time varying coefficients. The author further prove that they perform better than the pointwise and simultaneous confidence bands in the previous studies (Cai and Sun, 2003; Tian et al., 2005).

2.5 Feature Selection, Interaction Screening and Forest Algorithms

Most previous mentioned methods for variable selection and feature screening are designed for selecting main effects only. However, main effects are not sufficient to describe the relationship between the response and predictors under complex situations. Models including interaction effects give us a better approximation for the response surface, and therefore improve the prediction accuracy and bring new insight on the interplay between predictors. In many social, political, economic, bioassay and epidemiology problems, interaction models are useful in identifying nontrivial interactions between variables in modeling product sales, social networks or financial market changes. Especially, in genome-wide association studies(GWAS), researchers pay more and more attention to identify the interaction(epistatic) effects of single-nucleotide polymorphisms (SNPs) because gene-gene interactions can provide important insight on the complex biological pathways related to human diseases.

Two main categories of methods are studied in the literature about searching interaction effects. The first main category is called "two-stage analysis", which is an computational efficient modification of so called "joint analysis". In the methods of "joint analysis", they usually include both main and interaction effect in (2.114) altogether and make a global search over all candidate models.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \cdots + \beta_{pp} X_p^2 + \epsilon \tag{2.114}$$

However, as **?** points out, "joint analysis" methods become infeasible when p is large due to limiting factors such as memory requirement and computational cost. Based on the "marginality principle" and "heredity condition", "two-stage analysis" methods conclude that the models will have hierarchical structure, and interaction effects will be accompanied by at least one main effect. Under these assumptions, "two-stage analysis" methods could first select important main effects, and then select interaction effects of main effects that are obtained in the first stage. Therefore, "two-stage analysis" will be feasible choices when the data dimention p is large.

The second main category methods are random forest algorithms. As a natural structure for modelling interaction effects, trees are constructed and ensembled in different ways, which becomes different algorithms of random forest for prediction and selecting important features and interaction effects. Unlike the "two-stage analysis" methods, the forest algorithms assume neither the model forms of underlining mechanisms, nor the order of interaction effects. Both category of methods are reviewed in this section.

2.5.1 Two-stage Analysis and Interaction Screening

The model setup and notations of "two-stage analysis" are given here. Given n IID observations $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, we consider a regression model with linear and second-order terms

$$Y = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}^{(1)} + \mathbf{z}_i^T \boldsymbol{\beta}^{(2)} + \epsilon_i, 1 \leq i \leq n, \tag{2.115}$$

where Y_i is a real-valued response, $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})$ is a p-dimensional vector and the vector $\mathbf{z}_i = (X_{i1}^2, X_{i1}X_{i2}, \ldots, X_{i1}X_{ip}, X_{i2}^2, X_{i2}X_{i3}, \ldots, X_{ip}^2)^T$ includes quadratic and two-way interaction terms. Also, β_0 is the intercept, $\boldsymbol{\beta}^{(1)} = \{\beta_i\}$ and $\boldsymbol{\beta}^{(2)} = \{\beta_{jk}\}$ are regression coefficients of linear effects and order-2 effects. Throughout this part, we assume that $E(X_{ij}) = 0$, $Var(X_{ij}) = 1$, $E(Y_i) = 0$ and $Var(Y_i) = 1$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Denote the index sets of linear and order-2 terms as

$$\mathcal{P}_1 = \{1, 2, \ldots, p\}, \mathcal{P}_2 = \{(k, l) : 1 \leq k \leq l \leq p\} \tag{2.116}$$

and the nonzero linear and order-2 effects as

$$\mathcal{T}_1 = \{j : \beta_j \neq 0, j \in \mathcal{P}_1\} \tag{2.117}$$

$$\mathcal{T}_2 = \{(j, k) : \beta_{jk} \neq 0, (j, k) \in \mathcal{P}_2\} \tag{2.118}$$

Therefore, the full model is $\mathcal{F} = \mathcal{P}_1 \cup \mathcal{P}_2$ and the tree model is $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$

One typical representative of "two-stage analysis" is the two-stage iFOR procedure proposed by **?**. It extends the FS solution path algorithm of Wang (2009), which is discussed in the previous section, to the interaction selection, and proves the sure screening property for interaction selection under some regularity conditions. The two-stage iFOR procedure(iFORT) is outlined as follow:

Stage1: Define $\mathcal{C} = \mathcal{P}_1$. Implement FS on $\mathcal{C}$ to get the solution path $\{\mathcal{S}_t^{(1)}, t = 1, 2, \ldots\}$ and the selected main effects set $\hat{\mathcal{M}} = \{j_1, \ldots, j_{t_1}\}$.

Stage2: Update $\mathcal{C} = \hat{\mathcal{M}} \cup \{(k, l) : k \in \hat{\mathcal{M}}$ and $l \in \hat{\mathcal{M}}\}$. Implement FS on $\mathcal{C}$ by forcing-in $\hat{\mathcal{M}}$. Denote the solution path by $\{\mathcal{S}_{t_1+t}^{(2)}, t = 1, 2, \ldots\}$.

The above iFORT algorithm separately selects main effects and order-2 effects at two stages. Furthermore, under marginality principle, **?** propose another new algorithm iFORM to select both main effects and order-2 effects altogether through a dynamic candidate set $\mathcal{C}$. The iFORM algorithm, also called "iFOR Under Marginality Principle", is as follow:

Step1: (Initialization) Set $\mathcal{S}_0 = \emptyset, \mathcal{M}_0 = \emptyset$ and $\mathcal{C}_0 = \mathcal{P}_1$.

Step2: (Selection) In the t-th step with given $\mathcal{S}_{t-1}$, $\mathcal{C}_{t-1}$ and $\mathcal{M}_{t-1}$, FS is used to select one more variable from $\mathcal{C}_{t-1} \setminus \mathcal{S}_{t-1}$ and add it into $\mathcal{S}_{t-1}$ to get $\mathcal{S}_t$. If the newly added variable a is a main effect, update $\mathcal{M}_t = \mathcal{M}_{t-1} \cup \{a\}$ and $\mathcal{C}_t = \mathcal{P}_1 \cup \{(k,l) : k, l \in \mathcal{M}_t\}$.

Step3: (Solution path) Iterating Step 2 and get a solution path $\{\mathcal{S}_t : t = 1, 2, \ldots, D\}$.

The above D is chosen as a reasonable total number of important effects to terminate the procedure. The optimal model is determined from the FS path by BIC criterion.

One of the largest potential applying fields of interaction screening is the genome-wide association studies. Genetic interactions, which is known as epistasis, is playing a pivotal role in contributing to the genetic variation of phenotypic traits. To help locating the genes participating in the epistasis, Li et al. (2014) propose a two-stage sure independence screening(TS-SIS) procedure to generate a set of candidate SNPs and interactions, which may help to explain and predict the phenotypes of a complex trait. The related model setting for the observed phenotypic value y_i of subject i in a population cohort of n subjects is

$$\begin{aligned} y_i &= \mu + \sum_{k=1}^{q} x_{k,i}\alpha_k + \sum_{j=1}^{p} \xi_{j,i} a_j + \sum_{j=1}^{p} \zeta_{j,i} d_j & (2.119) \\ &= + \sum_{j=1}^{p} \sum_{j'<j} \xi_{j,i}\xi_{j',i}\mathcal{I}^{aa}_{jj'} + \sum_{j=1}^{p} \sum_{j'=1}^{p} \xi_{j,i}\zeta_{j',i}\mathcal{I}^{ad}_{jj'} & (2.120) \\ &= + \sum_{j=1}^{p} \sum_{j'=1}^{p} \zeta_{j,i}\xi_{j',i}\mathcal{I}^{da}_{jj'} + \sum_{j=1}^{p} \sum_{j'<j} \zeta_{j,i}\zeta_{j',i}\mathcal{I}^{dd}_{jj'} + \epsilon_i, & (2.121) \end{aligned}$$

where μ is the overall mean, q is the number of nongenetic covariates, p is the number of SNPs, $x_{k,i}$ is the k-th covariate for subject i, $k = 1, \ldots, q$, $i = 1, \ldots, n$, α_k is the effect of the k-th covariate, a_j and d_j are the additive effect and dominant effect of the j-th SNP, respectively. For $j = 1, \ldots, p$, $\mathcal{I}^{aa}_{jj'}$ is the additive×additive epistatic effect between the j-th SNP and the $j^{'}$-th SNP, $\mathcal{I}^{ad}_{jj'}$, $\mathcal{I}^{da}_{jj}$ and $\mathcal{I}^{dd}_{jj'}$ are additive× dominant epistatic effect, dominant× additive epistatic effect and dominant× dominant epistatic effect. If an effect is nonzero in regression model (2.119), the corresponding covariate or interaction is considered active. For the i-th subject, $\xi_{j,i}$ and $\zeta_{j,i}$, which are the indicators of the additive and dominant effects of the j-th SNP, are defined as

$$\xi_{j,i} = \begin{cases} 1, & \text{if the genotype of SNP j is AA} \\ 0, & \text{if the genotype of SNP j is Aa} \\ -1, & \text{if the genotype of SNP j is aa} \end{cases} \tag{2.122}$$

$$\zeta_{j,i} = \begin{cases} 1, & \text{if the genotype of SNP j is Aa} \\ 0, & \text{if the genotype of SNP j is AA or aa.} \end{cases} \tag{2.123}$$

That is, the additive effect a_j measures the effect of the average phenotypic value by substituting allele A with allele a, while dominant effect d_j represents how the effect of allele A is modified by the presence of allele a.

Let $\mathcal{D}_a$ and $\mathcal{D}_d$ be two sets of indices of truly important additive effects and dominant effects, respectively. The first stage of the TS-SIS procedure will select two subsets of SNPs with potential nonzero additive effects $\hat{\mathcal{D}}_a$ and dominant effects $\hat{\mathcal{D}}_d$. Based on $\hat{\mathcal{D}}_a$, an additive× additive interaction term is formulated by taking one SNP from $\hat{\mathcal{D}}_a$ and the other from all SNPs. Therefore, the set of additive×additive interactions are denoted by $\mathcal{D}_{aa}^{(0)} = \{(j, j^{'}) : \xi_j\xi_{j^{'}}, j \in \hat{\mathcal{D}}_a, j^{'} = 1, 2, \ldots, p\}$. Similarly, $\mathcal{D}_{ad}^{(0)}$,$\mathcal{D}_{da}^{(0)}$ and $\mathcal{D}_{dd}^{(0)}$ are formulated.

In the second stage, Li et al. (2014) apply SIS again to the pairwise epistatic interactions given in $\mathcal{D}_{aa}^{(0)}$,$\mathcal{D}_{ad}^{(0)}$,$\mathcal{D}_{da}^{(0)}$ and $\mathcal{D}_{dd}^{(0)}$. The algorithm is summarized as

Step1: Apply the SIS procedure to all additive and dominant main effects of SNPs and obtain the $\hat{\mathcal{D}}_a$ and $\hat{\mathcal{D}}_d$.

Step2: Formulate pairwise interaction sets $\mathcal{D}_{aa}^{(0)}$,$\mathcal{D}_{ad}^{(0)}$,$\mathcal{D}_{da}^{(0)}$ and $\mathcal{D}_{dd}^{(0)}$.

Step3: Apply the SIS procedure again to $\mathcal{D}_{aa}^{(0)}$,$\mathcal{D}_{ad}^{(0)}$,$\mathcal{D}_{da}^{(0)}$ and $\mathcal{D}_{dd}^{(0)}$, and obtain the reduced model $\hat{\mathcal{D}}_{aa}$,$\hat{\mathcal{D}}_{ad}$,$\hat{\mathcal{D}}_{da}$ and $\hat{\mathcal{D}}_{dd}$.

Step4: Combine all reduced model in step 1 and 3 and give the final selected model $\{\hat{\mathcal{D}}_a, \hat{\mathcal{D}}_d, \hat{\mathcal{D}}_{aa}, \hat{\mathcal{D}}_{ad}, \hat{\mathcal{D}}_{da}, \hat{\mathcal{D}}_{dd}\}$

Also, Li et al. (2014) propose the rates adjusted thresholding estimation (RATE) approach to determine the size of the reduced models selected by SIS.

Other than searching interaction effects using two-stage analysis procedure, Niu et al. (2018) suggest to directly do interaction screening by partial correlation(ISPC). In this paper, they point out that the direct generalization of existing screening methods to interaction screening can be incorrect or insufficient because it may overlook the intrinsic relationship between main effects and interactions. That is, it may label an interaction, say X_1X_2, as "important" while it is actually not predictive to the response, or the vise versa, because its intrinsic relationship with X_1 and X_2. To fix this problem, Niu et al. (2018) show that the ISPC procedure is a main-effect-adjusted interaction screening procedure, and further extend the ISPC procedure to the nonparametric rank correlation cases. The ISPC procedure is outlined as

1. Calculate the standardized interaction effects $\mathbf{Z}$.

2. Calculate the sample partial correlation $\mathbb{P}$ as

$$P_{jk} = \begin{cases} p\hat{Corr}(Y, X_jX_k|X_j, X_k), & 1 \le j \le k \le p \\ p\hat{Corr}(Y, X_j^2|X_j), & 1 \le j \le p \end{cases} \tag{2.124}$$

3. Determine a threshold λ and obtain a model $\hat{\mathcal{I}}_\lambda = \{(j,k) : |P_{jk}| > \lambda\}$

Furthermore, Niu et al. (2018) show that ISPC is invariant of arbitrary linear coding transformation, while the Pearson correlation $Corr(Y, X_jX_k$ is not invariant. Therefore, ISPC is more preferable.

2.5.2 Feature Selection Using Random Forest

Random forest algorithms are widely used for prediction and feature selection in the literature. To understand the mechanism of the "Random Forest"(RF), Breiman (2001) makes a start by using "Permutation Variable Importance Measure"(PVIM) for ranking variable importances. In a Random Forest, the PVIM of the j-th

$(j = 1, \ldots, p)$ feature X_j of the t-th tree is defined as

$$PVIM_t(X_j) = \frac{\sum_{i \in B_t}(Y_i - \hat{Y}_{it}^*)^2 - \sum_{i \in B_t}(Y_i - \hat{Y}_{it})^2}{|B_t|} \tag{2.125}$$

Here B_t is the OOB sample for tree t, $t = 1, \ldots, ntree$. $\hat{Y}_{it}$ and $\hat{Y}_{it}^*$ are the predictions for observation i got from the tree t before and after permuting X_j. The final importance measure of X_j is averaged over all trees

$$PVIM(X_j) = \sum_{t=1}^{ntree} \frac{PVIM_t(X_j)}{ntree} \tag{2.126}$$

As pointed out by Breiman (2001), Random Forest can help to understand the interaction of variables that is providing the predictive accuracy.

However, it is not enough to give a more complete picture. Especially, when the number of features is huge and the percentage of truly imformative features is small, the performance of RF declines significantly both in prediction accuracy and variable selection. Furthermore, the probability of detection of the causal variant decreases much more rapidly for interactions than variants with marginal effect. This is because as the dimensionality increases, interacting variants will rarely appear in a tree together and therefore be rarely modeled. The interaction will be even rarer modeled if they don't exhibit strong marginal effects.

A very natural idea to solve this is variable selection. Díaz-Uriarte and De Andres (2006) selected genes by iteratively fitting RF and dropping a pre-specified proportion of genes with the smallest importances each round. After fitting all forests, Díaz-Uriarte and De Andres (2006) choose the solution with the smallest number of genes whose OOB error rate is within 0 or 1 standard error of the minimum error rates of all forests. As a result, this algorithm usually selects a very small set of genes. However, the genes selected in the original samples are rarely selected in more than 50% of the bootstrap samples, which means the selected gene set is not very stable.

Instead of dropping features, Amaratunga et al. (2008) proposed a feature-weighted version of RF for feature ranking and selection under the name "Enriched Random Forests". Weighting can be done by scoring each gene based on its ability to seperate the groups or improve the prediction accuracy. By pointing out the weakness of the t-test p-values as the weights, Amaratunga et al. (2008) proposed to determine the weights based on q-values that are provide false discovery rate (FDR)-adjusted measures of significance for the features.

$$q_{(i)} = \min_{k \geq 1} \{ \min((G/k) p_{(k)}, 1) \}, \tag{2.127}$$

where G is the number of predictors, $p_{(i)}$ and $q_{(i)}$ are the p-value the q-value associated with the feature with the i-th smallest p-value. By assigning weights $w_i^{'} = (1/q_i) - 1$, features with $p_i \cong 1$ and $q_i \cong 1$ will get almost zero weight. At the same time, features with high separability will get large weights.

Another way to improve the performance of Random Forest is through weighting trees instead of variables. Winham et al. (2013) proposed to up-weigh better performing trees based on some measure of predictive ability at the tree-level. In the OOB training data, define $v_{train,ij}$ as the vote for subject i in tree j and denote oob_{ij} as an indicator for the out of bag status of subject i in tree j. The tree-level prediction error for the j-th tree is defined as:

$$tPE_j = \frac{1}{\sum_{i=1}^{M_1}} \sum_{i=1}^{M_1} |v_{train,ij} - y_i| * oob_{ij}. \tag{2.128}$$

In this paper, Winham et al. (2013) utilizes weights using right skewed distributions such as $w_j = \exp(\frac{1}{tPE_j})$ or $(\frac{1}{tPE_j})^{\lambda}$ to further up-weigh weights the best performing trees.

2.5.3 Interaction Selection via Forest Algorithm

Beside feature selection, some forest alogrithms have been developed to further improve our understanding of model mechanism through detecting interaction effects. However, finding interactions between variables in large and high-dimensional datasets is a serious computational challenge. Therefore, most algorithms build up interactions incrementally by adding variables in a greedy way. As a result, some informative high-order interaction will be overlooked.

To solve this, Shah and Meinshausen (2014) proposed the “Random Intersection Trees"(RIT) procedure to discover interactions using the intersection of d randomly chosen sets of active features, whose d corresponding responses are from the same category. Consider a binary classification problem with p predictors and n observations. The data is given in the form $(Z_i, \mathcal{I}_i), i = 1, \ldots, n$. For the observation i, Z_i is a binary label and $\mathcal{I}_i \subset \{1, 2, \ldots, p\}$ is a index subset of "active features" viewed as an interaction. With these notations, prevalence of an interaction $S \subset \{1, 2, \ldots, p\}$ in the class $C \in \{0, 1\}$ is defined as

$$\mathbb{P}_n(S|Z = C) := \frac{\sum_{i=1}^{n} I[S \subseteq \mathcal{I}_i]}{\sum_{i=1}^{n} I[Z_i = C]} \tag{2.129}$$

For given thresholds $0 \leq \theta_0 < \theta_1 \leq 1$, RIT searches for interaction S satisfying

$$\mathbb{P}_n(S|Z = 1) \geq \theta_1 \ ,\mathbb{P}_n(S|Z = 0) \leq \theta_0 \tag{2.130}$$

For each class $C \in \{0, 1\}$, RIT randomly chooses D, a pre-specified integer, observations from the set of observations $\{i : Z_i = C\}$ and takes D-fold intersections $S = \mathcal{I}_{i(1)} \cap \ldots \cap \mathcal{I}_{i(D)}$ to search for non-empty interaction S satisfying equation 2.130. S is called a survived interaction if it remains non-empty after D-fold intersection operation. To reduce computational complexity, these interactions are performed in a tree-like fashion and result in a set of n_{child}^{D} potential interactions(leaf nodes), where n_{child} is the number of children of each non-leaf node. By repeating this process M times for a given class C, RIT gives a collection of survived interaction $\mathcal{S} = \cup_{m=1}^{M} \mathcal{S}_m$. Here, $\mathcal{S}_m$ is the set of survived interactions among all n_{child}^{D} leaf nodes in the m-th tree. Given these survived interactions in $\mathcal{S}$, comparison across different classes will be subsequently conduct using the prevalence defined in 2.130.

Although computational efficient, RIT could only be used in selecting interactions under the settings of binary response and features. By combining feature-weighted RF (Amaratunga et al., 2008) and "Random Intersection Trees"(Shah and Meinshausen, 2014) through proposing "Generalized RIT" algorithm, Basu et al. (2018) raised an "iterative random forest"(iRF) algorithm to discover interactions for problems with binary response, and continuous or categorical features. Furthermore, Basu et al. (2018) proposed a "stability score" based on an "outer layer" of bootstrapping to assess the stability of recovered interactions.

The "Generalized RIT" could be viewed as a combination of RIT and a pretreatment step transforming one observation with binary response and continuous or categorical features into T pairs of binary response and features. For the t-th tree ($t = 1, \ldots, T$) of the output tree ensemble of a RF, we denote all its leaf nodes by $j_t = 1, \ldots, J(t)$. Each feature-response pair $(y_i, \mathbf{x}_i)$ is represented with respect to the t-th tree by $(Z_{i_t}, \mathcal{I}_{i_t})$, where $\mathcal{I}_{i_t}$ is the set of feature indices falling on the path of the leaf node containing $(y_i, \mathbf{x}_i)$ of the t-th tree. As a result, each $(y_i, \mathbf{x}_i)$ will produce T such index set and label pairs corresponding to T trees. Basu et al. (2018) aggregate all these pairs across observations and trees as

$$\mathcal{R} = \{(Z_{i_t}, \mathcal{I}_{i_t}) : \mathbf{x}_i \text{ falls in leaf node } i_t \text{ of tree t}\} \tag{2.131}$$

and apply RIT on this transformed dataset $\mathcal{R}$ to obtain a set of interactions.

The iRF algorithm consist of following three components:

1. **Iteratively re-weighted RF:** Given a pre-specified iteration K, iRF iteratively grows K weighted Random Forests RF($\mathbf{w}^{(k)}$), $k = 1, \ldots, K$. For the first iteration($k = 1$), the initial weights start with $\mathbf{w}^{(1)} = (1/p, \ldots, 1/p)$. The mean decreases in Gini impurity of k-th iteration are stored as the importance $\mathbf{v}^{(k)}$ and used as the feature weights $\mathbf{w}^{(k+1)}$ of the $k+1$-th iteration.

2. **Generalized RIT:** Apply the generalized RIT to the last feature-weighted RF of the last iteration. This step produces a collection of interactions $\mathcal{S}$

3. **Bagged Stability Scores:** Generate D bootstrap samples of the data $\mathcal{D}_{(b)}, b = 1, \ldots, B$, fit RF($\mathbf{w}^{(K)}$) on each of them and use the generalized RIT to produce corresponding collection of interactions $\mathcal{S}_b$ in each bootstrap sample. Based on bootstrap samples, we use the proportion of times (out of B bootstrap samples) an interaction $S \in \cup_{b=1}^{B} \mathcal{S}_{(b)}$ appears to represent its stability and define the stability score as

$$sta(S) = \frac{1}{B} \sum_{b=1}^{B} I\{S \in \mathcal{S}_{(b)}\} \tag{2.132}$$

Although iRF is presented in the binary classification setting, Basu et al. (2018) briefly mentioned that it could be naturally extended to multiclass or continuous responses. In the multiclass setting, it proposed to select leaf nodes with predicted class or classes of interest as inputs to RIT. In the regression setting, they consider leaf nodes whose predictions fall within a range of interest as inputs to generalized RIT. This range could be determined in domain-specific manner or by grouping responses using some clustering method. However, simulation and real data analysis were not studied in this paper.

2.6 Measures of Social Inequality Level

2.6.1 Gini Index

In economics, the Gini index (Gini, 1912, 2005, 1921) is a measure of polarization level used to describe how the income or wealth of residents distribute within a country, and become the most popular criterion of social inequality in recent decades. Assume that all people are ranked according to their income or wealth in ascending order. Given the rank, Gini index is defined based on the Lorenz curve in Figure 2.1, where horizontal and vertical axes stand for the cumulative proportions of the people and corresponding income, respectively. As shown in Figure 2.1, the Gini index is defined as the ratio of the area A to the lower triangle area,

$$\mathbf{Gini} = \frac{S_A}{S_A + S_B} = 1 - 2S_B, \tag{2.133}$$

where S_A and S_B stand for the areas of the regions A and B, and the latter equation in (2.133) holds because $S_A + S_B$ equals to 1/2. The Lorenz Curve produces different Gini indices under different distributions. For example, in an absolute equal society where everyone receives the exactly same income, the Lorenz Curve is the 45 degrees line, and the corresponding Gini index will be 0. In the other extreme case where one person possesses the total income and the remaining people have none, the Lorenz Curve becomes the two perpendicular sides of the lower triangle, and the corresponding Gini index will be 1. However, there is little chance that the Lorenz curve meets either of the two cases above; it is usually a curve between these two extreme lines, just as illustrated in Figure 2.1. That is, the Gini index takes value in the interval $[0, 1]$. Many empirical studies show that a good Gini index usually takes value between 0.3 and 0.4. Especially, if it exceeds

0.5, there will be a risk of social instability, while people will lack motivation to create wealth if the index is less than 0.2.

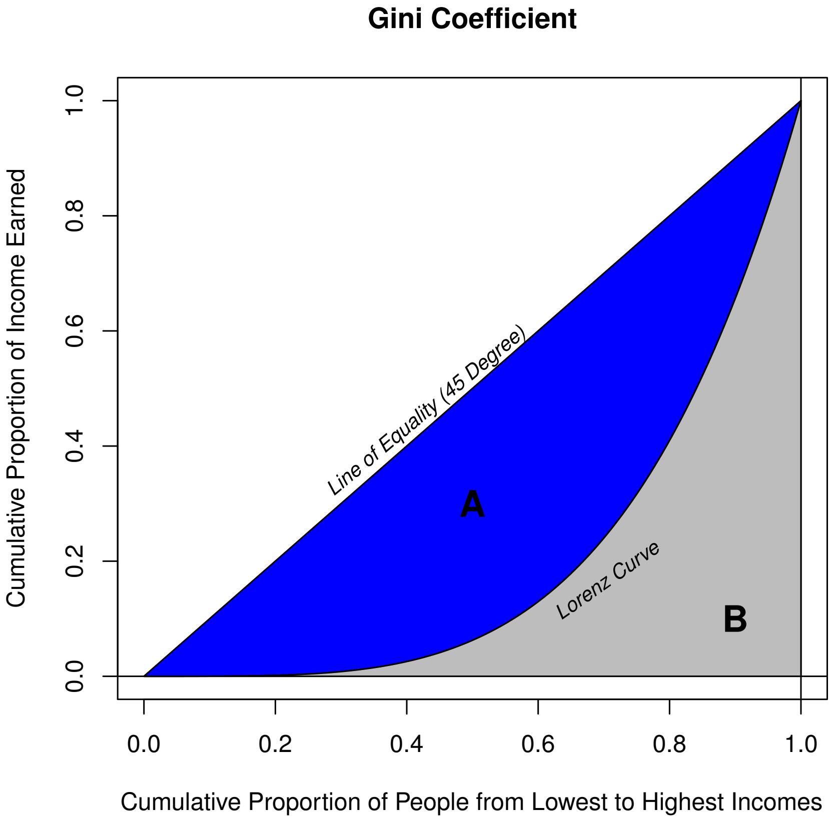

Figure 2.1. A typical Lorenz Curve and Gini index

The Gini index was first calculated based on the Riemann integral under the Lorenz curve. Let $x_i, i = 1, 2, ..., n$, denote n incomes ranked in ascending order, and $\hat{L}_i, i = 1, 2, ..., n$ denote the proportion of the first i incomes among all incomes. Thus we can obtain the estimation of S_B,

$$\hat{S}_B = \frac{1}{2n}\Sigma_{i=1}^{n}(\hat{L}_i + \hat{L}_{i-1}), \tag{2.134}$$

where $\hat{L}_i = \frac{1}{n\hat{\mu}}\Sigma_{j=1}^{i}x_j$, $\hat{\mu} = \frac{1}{n}\Sigma_{j=1}^{n}x_j$ and $L_0 = 0$.

In the following decades after Gini index was introduced, there were many other attempts to calculate it more precisely and efficiently. For instance, Kendall and Stuart (1977), Sen et al. (1997), Anand (1983) and Jasso (1979) proposed different calculation methods and showed their good properties. Pyatt (1976) and Silber (1989) improved the efficiency of the index computation with matrix methods. Until now, the Gini index has been playing a critical role in describing the inequality in social and economic issues.

2.6.2 Applications and Interpretations of the Gini Index

Along with theoretical research, more researchers have paid attention to the applications and interpretations of the Gini index. For example, the global income Gini index in 2005 had been estimated to be between 0.61 and 0.68 by the literatures (Hillebrand et al., 2009; Klugman, 2010). Moreover, some researchers have applied the Gini index to other fields as diverse as sociology, economics, health science, ecology, engineering and agriculture. For example, in social sciences and economics, Shorrocks (1978) introduced a measure based on income Gini coefficients to estimate income mobility. This measure, generalized by Maasoumi and Zandvakili (1986), is now generally referred to as the Shorrocks index. Thomas et al. (2001) have proposed an education Gini index, which can be used to discern trends in social development through educational attainment over time. Roemer (2013) and Weymark (2003) have created an opportunity Gini index. Sadras and Bongiovanni (2004) has assessed the yield inequality with paddocks using the Lorenz curve and Gini index. In addition, for a given time interval, the Gini index can also be used to compare diverse countries and different regions or groups within a country such as states, educational levels, gender and ethnic groups. Kopczuk et al. (2010) have applied social security income data for the United States since 1937 into Shorrocks indices and concluded that income mobility in the United States has a complicated history, primarily due to the mass influx of women into the American labor force after World War II and difference for men and women workers between 1937 and the 2000s.

2.6.3 More Indices about Economic and Social Inequality

At the same time, more indices are proposed to describe the economic and social inequality, which has become one focus of economic research. There are discussions how to make theoretical improvements on these indices and their applications in recent literature. Atkinson (1970) has proposed aptly-named Atkinson index to determine which end of the distribution contributed most to the observed inequality. Imedio-Olmedo et al. (2011) have compared the efficiency of a class of Bonferroni indices on measurements of inequality. Sundrum (2003) has derived the decomposition method to compute the indices of different subgroups. Firebaugh (1999) has ascribed some measures of inequality to the average deviations from the mean incomes, the greater the average deviation, the greater the inequality. Instead of the Gini index, Greselin and his co-workers(Greselin et al., 2013; Greselin, 2014) have introduced the Zenga index, $L-$functions and other statistics in measuring economic inequality and actuarial risks. All of these indices mentioned above can be viewed as the improvements of the Gini index.

Chapter 3

Feature Screening in Ultrahigh Dimensional Varying-coefficient Cox's Model

3.1 New Feature Screening Procedure for Varying Coefficient Cox's Model

Let T be the survival time, and $\mathbf{x}$ and U be p-dimensional covariate vector and univariate covariate, respectively. Throughout this paper, we consider the following varying coefficient Cox proportional hazard model:

$$h(t|\mathbf{x}, U) = h_0(t) \exp\{\mathbf{x}^\top \boldsymbol{\alpha}(U)\}, \tag{3.1}$$

where $h_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\alpha}(U) = \{\alpha_1(U), \ldots, \alpha_p(U)\}^\top$ consists of the unknown nonparametric coefficient functions. Here it is assumed that the support of U is finite and denoted by $[a, b]$. In survival data analysis, the survival time is subject to the censoring time C. Denote the observed time by $Z = \min\{T, C\}$ and the event indicator by $\delta = I(T \leq C)$. It is assumed throughout this paper that the censoring mechanism is noninformative. That is, given $\mathbf{x}$ and U, T and C are conditionally independent.

Suppose that $\{(\mathbf{x}_i, U_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ is an independently and identically distributed random sample from model (3.1). Let $t_1^0 < \ldots < t_N^0$ be the ordered observed failure times. Let (j) provide the label for the subject failing at t_j^0 so that the covariates associated with the N failures are $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$. Denote the risk set right before the time t_j^0 by $R_j = \{i : Z_i \geq t_j^0\}$. The partial likelihood function (Cox, 1975) of the random sample is

$$\ell_p\{\boldsymbol{\alpha}(U)\} = \sum_{j=1}^{N} \left\{ \mathbf{x}_j^\top \boldsymbol{\alpha}(U_j) - \log \left[\sum_{i \in R_j} \exp\{\mathbf{x}_i^\top \boldsymbol{\alpha}(U_i)\} \right] \right\}. \tag{3.2}$$

To estimate the nonparametric regression, we use B-spline basis in this paper. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$ and $\{\psi_{jk}, k = 1, \ldots, d_{nj}\}$ denote a normalized B-spline basis with $\|\psi_{jk}\|_\infty \leq 1$ and $d_{nj} = O(n^{1/5})$, where $\|\cdot\|_\infty$ is the sup norm. For any $\alpha_{nj}(U) \in \mathcal{S}_n$, we have

$$\alpha_{nj}(U) = \sum_{k=1}^{d_{nj}} \beta_{jk}\psi_{jk}(U) = \boldsymbol{\beta}_j^\top \boldsymbol{\psi}_j(U), \quad j = 1, \ldots, p, \tag{3.3}$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_{nj}}$. Here we allow d_{nj} to increase with n and be different for different j since different coefficient functions may have different smoothness. Under some conditions, the nonparametric coefficient function $\{\alpha_j(U)\}_{j=1}^{p}$ can be well approximated by functions in $\mathcal{S}_n$.

Substituting (3.3) into (3.2), the maximum partial likelihood estimate of (3.2) is to maximize

$$\begin{aligned} \ell_p(\boldsymbol{\beta}) &\hat{=} \sum_{j=1}^{N} \left\{ \mathbf{x}_j^\top \boldsymbol{\psi}_j^\top(U_j)\boldsymbol{\beta}_j - \log \left[\sum_{i \in R_j} \exp\{\mathbf{x}_i^\top \boldsymbol{\psi}_i^\top(U_i)\boldsymbol{\beta}_i\} \right] \right\}, \\ &= \sum_{j=1}^{N} \left\{ \mathbf{z}_j^\top \boldsymbol{\beta}_j - \log \left[\sum_{i \in R_j} \exp\{\mathbf{z}_i^\top \boldsymbol{\beta}_i\} \right] \right\}, \end{aligned} \tag{3.4}$$

with respect to $\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$ and $\mathbf{z}_i = (x_{i1}\boldsymbol{\psi}_1(U_i)^\top, \ldots, x_{ip}\boldsymbol{\psi}_p(U_i)^\top)^\top$. We next propose a feature screening procedure based on (3.4).

3.1.1 A New Feature Screening Procedure

Denote $\|\alpha_j(U)\|_2 = [\mathrm{E}\alpha_\mathrm{j}^2(\mathrm{U})]^{1/2}$, the L_2-norm of $\alpha_j(U)$. For ease of presentation, denote s as an arbitrary subset of $\{1, \ldots, p\}$, $\mathbf{x}_s = \{x_j, j \in s\}$ and $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$. For a set s, $\tau(s)$ stands for the cardinality of s. Suppose the effect of $\mathbf{x}$ is sparse, and the true value of $\boldsymbol{\alpha}(U)$ is $\boldsymbol{\alpha}^*(U)$, where $\boldsymbol{\beta}^*$ is the corresponding coefficients of $\boldsymbol{\alpha}^*(U)$. Denote $s^* = \{j : \|\alpha_j(U)\|_2 > 0\}$. By sparsity, we mean that $\tau(s^*)$ is much less than p. The goal of feature screening is to identify a subset s such that $s^* \subset s$ with overwhelming probability and $\tau(s)$ is also much less than p. According to (3.4), we propose screening features for the varying coefficient Cox model by the constrained partial likelihood

$$\hat{\boldsymbol{\beta}}_m = \arg\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) \quad \text{subject to} \quad \tau(\{j : \|\boldsymbol{\beta}_j\|_2 > 0\}) \leq m \tag{3.5}$$

for a pre-specified m which is assumed to be greater than the number of nonzero elements of $\boldsymbol{\beta}^*$.

For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (3.5) directly. Alternatively, we consider a proxy of the partial likelihood function. It follows by the Taylor expansion for the partial likelihood function $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ lying within a neighbor of $\boldsymbol{\gamma}$ that

$$\ell_p(\boldsymbol{\gamma}) \approx \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p''(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}),$$

where $\ell_p'(\boldsymbol{\beta}) = \partial \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell_p''(\boldsymbol{\beta}) = \partial^2 \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. Denote $P_t = \sum_{j=1}^p d_{nj}$. If $\ell_p''(\boldsymbol{\beta})$ is invertible, the computational complexity of calculating the inverse of $\ell_p''(\boldsymbol{\beta})$ is $O(P_t^3)$. For large P_t, small n problems (i.e., $P_t \gg n$), $\ell_p''(\boldsymbol{\beta})$ becomes not invertible. Low computational costs are always desirable for feature screening. To deal with singularity of the Hessian matrix and save computational costs, we propose to use the following approximation for $\ell_p''(\boldsymbol{\gamma})$

$$h(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}), \tag{3.6}$$

where u is a scaling constant to be specified and $W(\boldsymbol{\beta}) = \mathrm{diag}\{W_1(\boldsymbol{\beta}), \ldots, W_p(\boldsymbol{\beta})\}$, a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being a $d_{nj} \times d_{nj}$ matrix. Here (3.6) is the minimization of the original objective function, i.e,. $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$ under some conditions. Due to the properties of the majorization and minorization algorithm, using (3.6) we can obtain the same estimates as the original objective function. The two functions themselves, however, are not numerically equal. Here we allow $W(\boldsymbol{\beta})$ to depend on $\boldsymbol{\beta}$. This implies that we approximate $\ell_p''(\boldsymbol{\beta})$ by $-uW(\boldsymbol{\beta})$. Throughout this paper, we will use $W_j(\boldsymbol{\beta}) = -\partial^2\ell_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top$.

It can be seen that $h(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta})$, and under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details. Since $W(\boldsymbol{\beta})$ is a block diagonal matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of $\boldsymbol{\gamma}_j$ for any given $\boldsymbol{\beta}$. The additivity enables us to have a closed form solution for the following maximization problem

$$\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \qquad \text{subject to} \quad \tau(\{j : \|\boldsymbol{\gamma}_j\|_2 > 0\}) \leq m \tag{3.7}$$

for given $\boldsymbol{\beta}$ and m. Define $\tilde{\boldsymbol{\gamma}}_j = \boldsymbol{\beta}_j + u^{-1}W_j^{-1}(\boldsymbol{\beta}_j)\partial\ell_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_j$ for $j = 1, \ldots, p$, and $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1^\top, \ldots, \tilde{\boldsymbol{\gamma}}_p^\top)^\top = \boldsymbol{\beta} + u^{-1}W^{-1}(\boldsymbol{\beta})\ell_p'(\boldsymbol{\beta})$ is the maximizer of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. Denote $g_j = \tilde{\boldsymbol{\gamma}}_j^\top W_j(\boldsymbol{\beta}_j)\tilde{\boldsymbol{\gamma}}_j$ for $j = 1, \ldots, p$, and sort g_j so that $g_{(1)} \geq g_{(2)} \geq \ldots \geq g_{(p)}$. The solution of maximization problem (3.7) is the hard-thresholding rule defined below

$$\hat{\boldsymbol{\gamma}}_j = \tilde{\boldsymbol{\gamma}}_j I\{g_j > g_{(m+1)}\}.$$

This enables us to effectively screen features by using the following algorithm.

```
Feature Screening Algorithm of Varying Coefficient Cox's Models
Step 1. Set the initial value $\boldsymbol{\beta}_j^{(0)} = \mathbf{0}$, $j = 1, \cdots, p$.
Step 2. Set $t = 0, 1, 2, \cdots$, iteratively conduct Step 2a and Step 2b below
    until the algorithm converges.
    Step 2a.Calculate $\tilde{\boldsymbol{\gamma}}_j^{(t)} = \boldsymbol{\beta}_j^{(t)} + u_t^{-1}W_j^{-1}(\boldsymbol{\beta}_j)\partial\ell(\boldsymbol{\beta}^{(t)})/\partial\boldsymbol{\beta}_j$, and $g_j^{(t)} = \{\tilde{\boldsymbol{\gamma}}_j^{(t)}\}^\top W_j(\boldsymbol{\beta}^{(t)})\tilde{\boldsymbol{\gamma}}_j^{(t)}$.
        Let $g_{(1)}^{(t)} \geq g_{(2)}^{(t)} \geq \ldots \geq g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : g_j^{(t)} \geq g_{(m+1)}^{(t)}\}$,
        the nonzero index set.
    Step 2b.Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{\beta}_1^{(t+1)}, \ldots, \boldsymbol{\beta}_p^{(t+1)})^\top$ as follows. If $j \notin S_t$, set $\boldsymbol{\beta}_j^{(t+1)} = \mathbf{0}$,
        otherwise, set $\{\boldsymbol{\beta}_j^{(t+1)} : j \in S_t\}$ be the partial likelihood estimate of the submodel $S_t$.
```

Theorem 1. *Suppose that Conditions (D1)—(D4) in the Appendix hold. Let*

$\{\boldsymbol{\beta}^{(t)}\}$ be the sequence defined in Step 2b in the above algorithm. Denote

$$\rho^{(t)} = \sup_{\boldsymbol{\beta}} \left[\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell_p''(\boldsymbol{\beta})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\} \right],$$

where $\lambda_{\max}(A)$ stands for the maximal eigenvalue of a matrix A. If $u_t \geq \rho^{(t)}$, then

$$\ell_p(\boldsymbol{\beta}^{(t+1)}) \geq \ell_p(\boldsymbol{\beta}^{(t)}),$$

where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b in the above algorithm.

Theorem 1 claims the ascent property of the proposed algorithm if u_t is appropriately chosen. That is, the proposed algorithm may improve the current estimate within the feasible region (i.e. $\tau(\{j : \|\alpha_j(U)\|_2 > 0\}) \leq m$), and the resulting estimate in the current step may serve as a refinement of the last step. This theorem also provides us some insights about choosing u_t in practical implementation.

3.1.2 Sure Screening Property

For a subset s of $\{1, \ldots, p\}$ with size $\tau(s)$, recall notation $\mathbf{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$ corresponding to $\boldsymbol{\beta}_s = \{\boldsymbol{\beta}_j, j \in s\}$ with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_{n_j}})^\top$. We denote the true model by $s^* = \{j : \mathrm{E}\alpha_{\mathrm{j}}^2(\mathrm{U}) > 0, 1 \leq \mathrm{j} \leq \mathrm{p}\}$ with $\tau(s^*) = q$. The objective of feature screening is to obtain a subset $\hat{s}$ such that $s^* \subset \hat{s}$ with very high probability.

We now provide some theoretical justifications for the ultrahigh dimensional varying coefficient Cox model. The sure screening property (Fan and Lv, 2008) is referred to as

$$\mathrm{Pr}(\mathrm{s}^* \subset \hat{\mathrm{s}}) \longrightarrow 1, \quad \text{as} \quad \mathrm{n} \to \infty, \tag{3.8}$$

To establish this sure screening property for the proposed varying coefficient Cox model, we introduce some additional notations as follows. For any model s, let $\ell'(\boldsymbol{\beta}_s) = \partial\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s$ and $\ell''(\boldsymbol{\beta}_s) = \partial^2\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^\top$ be the score function and the Hessian matrix of $\ell(\cdot)$ as a function of $\boldsymbol{\beta}_s$, respectively. Assume that a screening procedure retains m out of p features such that $\tau(s^*) = q < m$. So, we define

$$S_+^m = \{s : s^* \subset s; \|s\|_0 \leq m\} \quad \text{and} \quad S_-^m = \{s : s^* \not\subset s; \|s\|_0 \leq m\}$$

as the collections of the over-fitted models and the under-fitted models, respectively. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_m$ under the scenario where p, q, m and $\boldsymbol{\beta}^*$ are allowed to depend on the sample size n. We impose the following conditions, some of which are purely technical and only serve to facilitate theoretical understanding of the proposed feature screening procedure.

(C1) The support of U is bounded on [a,b].

(C2) The function $\{\alpha_j(U)\}_{j=1}^p$ belong to a class of functions $\mathcal{F}$, whose r-th derivative $\alpha_j^{(r)}(\cdot)$ exists and is Lipschitz of order η,

$$\mathcal{F} = \left\{\alpha_j(\cdot) : |\alpha_j^{(r)}(s) - \alpha_j^{(r)}(t)| \leq K|s-t|^\eta \text{ for } s, t \in [a, b]\right\},$$

for some positive constant K, where r is a nonnegative integer and $\eta \in (0, 1]$ such that $\nu = r + \eta > 0.5$.

(C3) There exist $w_1, w_2 > 0$ and some non-negative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 1/2$ and

$$\min_{j \in s^*} \|\alpha_j(U)\|_2 \geq w_1 n^{-\tau_1} \quad \text{and} \quad q < m \leq w_2 n^{\tau_2}.$$

(C4) $\log p = O(n^\kappa)$ for some $0 \leq \kappa < 1 - 2(\tau_1 + \tau_2)$.

(C5) There exist constants $C_1, C_2 > 0$, $\delta > 0$, such that for sufficiently large n,

$$C_1 d_n^{-1} \leq \lambda_{\min}[-n^{-1}\ell_p''(\boldsymbol{\beta}_s)] \leq \lambda_{\max}[-n^{-1}\ell_p''(\boldsymbol{\beta}_s)] \leq C_2 d_n^{-1}$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leq \delta\}$ and $s \in S_+^{2m}$, where $\lambda_{\min}[\cdot]$ and $\lambda_{\max}[\cdot]$ denotes the smallest and largest eigenvalues of a matrix.

Under Conditions (C1) and (C2), the following two properties of B-splines are valid.

(a) (De Boor et al. (1978)) For $k = 1, \ldots, d_n$, $\psi_{jk}(U) \geq 0$ and $\sum_{k=1}^{d_n} \psi_{jk}(U) = 1$, $U \in [a, b]$. In addition, there exist positive constants C_3 and C_4 such that $C_3 d_n^{-1} \leq E\psi_{jk}^2(U) \leq C_4 d_n^{-1}$.

(b) (Stone (1982); Stone et al. (1985)) If $\{\alpha_j, j = 1, 2, \cdots, p\}$ is a set of functions in $\mathcal{F}$ described in condition (C2), there exists a positive constant C_5 that does not depend on $\alpha_j(U)$, then the uniform approximation error satisfies $\rho = \sup_{U\in[a,b]} \|\alpha_j(U) - \alpha_{nj}(U)\|_2 \leq C_5 d_n^{-\nu}, \forall j$, as $d_n \to \infty$.

Conditions (C1) and (C2) ensure properties (a) and (b), which are required for the B-spline approximation and establishing the sure screening properties.

Note that $\|\alpha_{nj}(U)\|_2^2 = \boldsymbol{\beta}_j^\top \mathrm{E}\{\boldsymbol{\psi}_\mathrm{j}(\mathrm{U})\boldsymbol{\psi}_\mathrm{j}(\mathrm{U})^\top\}\boldsymbol{\beta}_\mathrm{j}$, based on the properties (a), (b) and Condition (C3), we can derive that

$$\min_{j\in s^*} \|\boldsymbol{\beta}_j\|_2 \geq w_1 d_n n^{-\tau_1}.$$

Condition (C3) states a few requirements for establishing the sure screening property of the proposed procedure. The first one is the sparsity of $\boldsymbol{\alpha}^*(U)$ which makes the sure screening possible with $\tau(\hat{s}) = m > q$. Also, it requires that the minimal component in $\boldsymbol{\alpha}^*(U)$ does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Meanwhile, together with (C4), it confines an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leq m$. Condition (C5) assumes that p diverges with n at up to an exponential rate; it implies that the number of covariates can be substantially larger than the sample size.

We establish the sure screening property of the quasi-likelihood estimation by the following theorem.

Theorem 2. *Suppose that Conditions (C1)—(C5) and Conditions (D1)—(D7) in the Appendix hold. Let $\hat{s}$ be the model obtained by the (3.5) of size m. We have*

$$Pr(s^* \subset \hat{s}) \to 1, \quad as \quad n \to \infty.$$

The proof is given in the following section. The sure screening property is an appealing property of a screening procedure since it ensures that the true active predictors are retained in the model selected by the screening procedure. To be distinguished from the SIS procedure, the proposed procedure is referred to as sure joint screening (SJS) procedure.

3.2 Technical Proofs

We need the following notation to present the regularity conditions for the partial likelihood and the Cox model. Most notations are adapted from Andersen and Gill Andersen and Gill (1982), in which counting processes were introduced for the Cox model and the consistency and asymptotic normality of the partial likelihood estimate were established. Denote $\overline{N}_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $R_i(t) = \{T_i \geq t, C_i \geq t\}$. Assume that there are no two component processes $N_i(t)$ jumping at the same time. For simplicity, we shall work on the finite interval $[0, \tau]$.

In Cox's model, properties of stochastic processes, such as being a local martingale or a predictable process, are relative to a right-continuous nondecreasing family $(\mathcal{F}_t : t \in [0, \tau])$ of sub σ-algebras on a sample space $(\Omega, \mathcal{F}, \mathcal{P})$; $\mathcal{F}_t$ represents everything that happens up to time t. Throughout this section, we define $\Lambda_0(t) = \int_0^t h_0(u)\, du$.

By stating that $\overline{N}_i(t)$ has intensity process $h_i(t) \hat{=} h(t|\mathbf{x}_i)$, we mean that the processes $M_i(t)$ defined by

$$M_i(t) = \overline{N}_i(t) - \int_0^t h_i(u) du, \quad i = 1, \ldots, n,$$

are local martingales on the time interval $[0, \tau]$.

Define

$$\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} R_i(t) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \mathbf{x}_i^{\otimes k}, \quad \mathbf{s}^{(k)}(\boldsymbol{\beta}, t) = \mathrm{E}[\mathbf{S}^{(\mathrm{k})}(\boldsymbol{\beta}, \mathrm{t})] \quad \text{for} \quad \mathrm{k} = 0, 1, 2,$$

and

$$\mathrm{E}(\boldsymbol{\beta}, \mathrm{t}) = \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, \mathrm{t})}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, \mathrm{t})}, \quad \mathrm{V}(\boldsymbol{\beta}, \mathrm{t}) = \frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, \mathrm{t})}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, \mathrm{t})} - \mathrm{E}(\boldsymbol{\beta}, \mathrm{t})^{\otimes 2}.$$

where $\mathbf{x}_i^{\otimes 0} = 1$, $\mathbf{x}_i^{\otimes 1} = \mathbf{x}_i$ and $\mathbf{x}_i^{\otimes 2} = \mathbf{x}_i \mathbf{x}_i^T$. Note that $\mathbf{S}^{(0)}(\boldsymbol{\beta}, t)$ is a scalar, $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)$ and $\mathrm{E}(\boldsymbol{\beta}, \mathrm{t})$ are p-vector, and $\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)$ and $\mathrm{V}(\boldsymbol{\beta}, \mathrm{t})$ are $p \times p$ matrices.

Define

$$Q_j = \sum_{i=1}^{n} \int_0^{t_j} \left\{ \mathbf{x}_i - \frac{\sum_{i \in R_j} \mathbf{x}_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right\} dM_i.$$

Here, $E[Q_j|\mathcal{F}_{j-1}] = Q_{j-1}$ i.e. $E[Q_j - Q_{j-1}|\mathcal{F}_{j-1}] = 0$. Let $b_j = Q_j - Q_{j-1}$, then $(b_j)_{j=1,2,\ldots}$ is a sequence of bounded martingale differences on $(\Omega, \mathcal{F}, P)$. That is,

b_j is bounded almost surely (a.s.) and $E[b_j|\mathcal{F}_{j-1}] = 0$ a.s. for $j = 1, 2, \ldots$.

(D1) (Finite interval). $\Lambda_0(\tau) = \int_0^\tau h_0(t)dt < \infty$

(D2) (Asymptotic stability). There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ and scalar, vector and matrix functions $\mathbf{s}^{(0)}$,$\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ defined on $\mathcal{B} \times [0, \tau]$ such that for $k = 0, 1, 2$

$$\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}} \|\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(k)}(\boldsymbol{\beta}, t)\| \xrightarrow{p} 0.$$

(D3) (Lindeberg condition). There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i,t} |\mathbf{x}_i| R_i(t) I\{\boldsymbol{\beta}_0' \mathbf{x}_i > -\delta |\mathbf{x}_i|\} \xrightarrow{p} 0,$$

(D4) (Asymptotic regularity conditions). Let $\mathcal{B}$, $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ be as in Condition (D2) and define $e = \mathbf{s}^{(1)}/\mathbf{s}^{(0)}$ and $v = \mathbf{s}^{(2)}/\mathbf{s}^{(0)} - e^{\otimes 2}$. For all $\boldsymbol{\beta} \in \mathcal{B}$,$t \in [0, \tau]$;

$$\mathbf{s}^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{s}^{(0)}(\boldsymbol{\beta}, t), \quad \mathbf{s}^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \mathbf{s}^{(0)}(\boldsymbol{\beta}, t),$$

$\mathbf{s}^{(0)}(\cdot, t)$, $\mathbf{s}^{(1)}(\cdot, t)$ and $\mathbf{s}^{(2)}(\cdot, t)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0, \tau]$, $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ are bounded on $\mathcal{B} \times [0, \tau]$; $\mathbf{s}^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0, \tau]$, and the matrix

$$\mathbf{S} = \int_0^\tau v(\boldsymbol{\beta}_0, t) \mathbf{s}^{(0)}(\boldsymbol{\beta}_0, t) h_0(t) dt$$

is positive definite.

(D5) The function $\mathbf{S}^{(0)}(\boldsymbol{\beta}^*, t)$ and $\mathbf{s}^{(0)}(\boldsymbol{\beta}^*, t)$ are bounded away from 0 on $[0, \tau]$.

(D6) There exist constants $C_1, C_2 > 0$, such that $\max_{ij} |x_{ij}| < C_1$ and $\max_i |\mathbf{x}_i^T \boldsymbol{\beta}^*| < C_2$.

(D7) $\{b_j\}$ is a sequence of martingale differences and there exit nonnegative constants c_j such that for every real number t,

$$E\{\exp(tb_j)|\mathcal{F}_{j-1}\} \leq \exp(c_j^2 t^2/2) \quad a.s. \quad (j = 1, 2, \ldots, N)$$

For each j, the minimum of those c_j is denoted by $\eta(b_j)$.

$$|b_j| \le K_j \quad a.s. \quad \text{for} \quad j = 1, 2, \ldots, N$$

and $E\{b_{j_1}, b_{j_2}, \ldots, b_{j_k}\} = 0$ for $b_{j_1} < b_{j_2} < \cdots < b_{j_k}$;$k = 1, 2, \ldots$.

Note that the partial derivative conditions on $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ are satisfied by $\mathbf{S}^{(0)}$, $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$; and that $\mathbf{S}$ is automatically positive semidefinite. Furthermore the interval $[0, \tau]$ in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$.

Condition (D1)—(D5) is a standard condition for the proportional hazards model (Andersen and Gill, 1982), which is weaker than the one required by Bradic et al. (2011) and $\mathbf{S}^{(k)}(\boldsymbol{\beta}_0, t)$ converges uniformly to $\mathbf{s}^{(k)}(\boldsymbol{\beta}_0, t)$. Condition (D6) is a routine one, which is needed to apply the concentration inequality for general empirical processes. For example, the bounded covariate assumption is used by Huang et al. (2013) for discussing the Lasso estimator of proportional hazards models. Condition (D7) is needed for the asymptotic behavior of the score function $\ell_p'(\boldsymbol{\beta})$ of partial likelihood because the score function cannot be represented as a sum of independent random vectors, but it can be represented as sum of a sequence of martingale differences.

Proof of Theorem 1. Applying the Taylor expansion to $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \boldsymbol{\beta}$, it follows that

$$\ell_p(\boldsymbol{\gamma}) = \ell_p(\boldsymbol{\beta}) + \ell_p'(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p''(\tilde{\boldsymbol{\beta}})(\boldsymbol{\gamma} - \boldsymbol{\beta}),$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

$$\begin{aligned} (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \{-\ell_p''(\tilde{\boldsymbol{\beta}})\}(\boldsymbol{\gamma} - \boldsymbol{\beta}) &= (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W^{1/2}(\boldsymbol{\beta}) W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-/2}(\boldsymbol{\beta}) W^{1/2}(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) \\ &\le \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta})](\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}), \end{aligned}$$

where $W(\boldsymbol{\beta})$ is a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being a $d_{n_j} \times d_{n_j}$ matrix. Since $-\ell''(\boldsymbol{\beta})$ is non-negative definite, $\lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta})] \ge 0$. Thus, if

$$u > \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta})] \ge 0,$$

then

$$\ell_p(\boldsymbol{\gamma}) \geq \ell_p(\boldsymbol{\beta}) + \ell_p'(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) = h(\boldsymbol{\gamma}|\boldsymbol{\beta}).$$

Thus it follows that $\ell_p(\boldsymbol{\gamma}) \geq h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ and $\ell_p(\boldsymbol{\beta}) = h(\boldsymbol{\beta}|\boldsymbol{\beta})$ by the definition of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. The solution of $\partial h(\boldsymbol{\gamma}|\boldsymbol{\beta})/\partial\boldsymbol{\gamma} = 0$ is $\boldsymbol{\gamma} = \boldsymbol{\beta} + u^{-1}W(\boldsymbol{\beta})\ell'(\boldsymbol{\beta})$. Hence, under the conditions of Theorem 1, it follows that

$$\ell_p(\boldsymbol{\beta}^{*(t+1)}) \geq h(\boldsymbol{\beta}^{*(t+1)}|\boldsymbol{\beta}^{(t)}) \geq h(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t)}) = \ell(\boldsymbol{\beta}^{(t)}).$$

The second inequality is due to the fact that $\tau(\{j : \|\boldsymbol{\beta}_j^{*(t+1)}\|_2 > 0\}) = \tau(\{j : \|\boldsymbol{\beta}_j^{(t)}\|_2 > 0\}) = m$, and $\boldsymbol{\beta}^{*(t+1)} = \arg\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})$ subject to $\tau(\{j : \|\boldsymbol{\gamma}_j\|_2 > 0\}) \leq m$. By definition of $\boldsymbol{\beta}^{(t+1)}$, $\ell_p(\boldsymbol{\beta}^{(t+1)}) \geq \ell_p(\boldsymbol{\beta}^{*(t+1)})$ and $\tau(\{j : \|\boldsymbol{\beta}_j^{(t+1)}\|_2 > 0\}) = m$. This proves Theorem 1. □

Proof of Theorem 2. For a given model s, a subset of $\{1, \ldots, p\}$, let $\hat{\boldsymbol{\alpha}}_s(U)$ be the partial likelihood estimate of $\boldsymbol{\alpha}_s(U)$ based on the spline approximation. The theorem is implied if $\Pr\{\hat{\mathrm{s}} \in \mathrm{S}_+^{\mathrm{m}}\} \to 1$. Thus, it suffices to show that

$$\Pr\left\{\max_{\mathrm{s}\in\mathrm{S}_-^{\mathrm{m}}} \ell_{\mathrm{p}}\{\hat{\boldsymbol{\alpha}}_{\mathrm{s}}(\mathrm{U})\} \geq \min_{\mathrm{s}\in\mathrm{S}_+^{\mathrm{m}}} \ell_{\mathrm{p}}\{\hat{\boldsymbol{\alpha}}_{\mathrm{s}}(\mathrm{U})\}\right\} \to 0, \tag{3.9}$$

as $n \to \infty$.

We approximate the coefficient function $\alpha_j(U)$ by

$$\alpha_{nj}(U) = \sum_{k=1}^{d_{n_j}} \beta_{jk}\psi_{jk}(U) = \boldsymbol{\beta}_j^\top \boldsymbol{\psi}_j(U), \quad j = 1, \ldots, p, \tag{3.10}$$

where $\psi_{jk}(U)$, $k = 1, \ldots, d_n$, are basis functions and d_{n_j} is the number of basis functions, which is allowed to increase with the sample size n. For $\alpha_{nj}(U)$, define the approximation error by

$$\rho_j(U) = \alpha_j(U) - \alpha_{nj}(U) = \alpha_j(U) - \boldsymbol{\beta}_j^\top \boldsymbol{\psi}_j(U), \quad j = 1, \ldots, p.$$

Let $\mathrm{dist}(\alpha_j(U), \mathcal{S}_j) = \inf_{\alpha_{nj}(U)\in\mathcal{S}_j} \sup_{U\in[a,b]} \|\rho_j(U)\|_2$, and take $\rho = \max_{1\leq j\leq p} \mathrm{dist}(\alpha_j(U), \mathcal{S}_j)$. Let $\boldsymbol{\alpha}_n(U) = (\alpha_{n1}(U),$

$\ldots, \alpha_{np}(U))^\top$ and $\boldsymbol{\alpha}(U) = (\alpha_1(U), \ldots, \alpha_p(U))^\top$. For any s,

$$\begin{aligned} \boldsymbol{\alpha}_s(U) &= \begin{pmatrix} \boldsymbol{\psi}_1(U) & & \\ & \ddots & \\ & & \boldsymbol{\psi}_s(U) \end{pmatrix}_{s \times sd_n} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_s \end{pmatrix}_{sd_n \times 1} + \begin{pmatrix} \rho_1(U) \\ \vdots \\ \rho_s(U) \end{pmatrix} \\ &\hat{=} \Psi_s(U)\boldsymbol{\beta}_s + \rho_s(U), \end{aligned}$$

where $\Psi_s(U) = \mathrm{diag}(\boldsymbol{\psi}_1(U), \ldots, \boldsymbol{\psi}_s(U))$ with $\boldsymbol{\psi}_j(U) = (\psi_{j1}(U), \ldots, \psi_{jd_n}(U))$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_{n_j}})^\top$, $j = 1, \ldots, s$.

For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. So, we have

$$\begin{aligned} \ell_p\{\boldsymbol{\alpha}_{s'}(U)\} - \ell_p\{\boldsymbol{\alpha}^*_{s'}(U)\} &= \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}_{s'} + \rho_{s'}(U)\} - \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}^*_{s'} + \rho^*_{s'}(U)\} \\ &= \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}_{s'}\} + \ell'_p\{\Psi_{s'}(U)\tilde{\boldsymbol{\beta}}_{s'}\}\rho_{s'}(U) - \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}^*_{s'}\} - \ell'_p\{\Psi_{s'}(U) \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ and $\tilde{\boldsymbol{\beta}}^*_{s'}$ are two immediate values. Denote

$$\Delta_1 = \{\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'})\}, \quad \Delta_2 = \ell'_p(\tilde{\boldsymbol{\beta}}_{s'})\rho_{s'}(U), \quad \Delta_3 = \ell'_p(\tilde{\boldsymbol{\beta}}^*_{s'})\rho^*_{s'}(U).$$

Thus, we have

$$\ell_p\{\boldsymbol{\alpha}_{s'}(U)\} - \ell_p\{\boldsymbol{\alpha}^*_{s'}(U)\} = \Delta_1 + \Delta_2 + \Delta_3.$$

For Δ_2, by Cauchy-Schwartz inequality, we have

$$\mathrm{E}|\Delta_2| = \mathrm{E}|\ell'_\mathrm{p}(\tilde{\boldsymbol{\beta}}_{\mathrm{s}'})\rho_{\mathrm{s}'}(\mathrm{U})| \leq \sqrt{\mathrm{E}\|\ell'_\mathrm{p}(\tilde{\boldsymbol{\beta}}_{\mathrm{s}'})\|^2}\sqrt{\mathrm{E}\|\rho_{\mathrm{s}'}(\mathrm{U})\|^2}.$$

By condition (C5) and Corollary 1 in Wei et al. (2011), we can obtain $\Delta_2 = o_p(1)$. Similarly Δ_2, we can also obtain, $\Delta_3 = o_p(1)$.

Next, we consider term Δ_1. For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. Under (C3) condition, we consider $\boldsymbol{\beta}_{s'}$ close to $\boldsymbol{\beta}^*_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1 d_n n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\boldsymbol{\beta}_{s'}$ falls into a small neighborhood of $\boldsymbol{\beta}^*_{s'}$, so that Condition (C5) becomes applicable. Thus, it follows Condition (C5) and the Cauchy-Schwarz inequality that

$$\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) = [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^\top \ell'_p(\boldsymbol{\beta}^*_{s'}) + (1/2)[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^\top \ell''_p(\tilde{\boldsymbol{\beta}}_{s'})[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]$$

$$
\begin{aligned}
&\leq [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^\top \ell'_p(\boldsymbol{\beta}^*_{s'}) - (C_1 d_n^{-1}/2) n \|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2^2 \\
&\leq w_1 d_n n^{-\tau_1} \|\ell'_p(\boldsymbol{\beta}^*_{s'})\|_2 - (C_1 d_n/2) w_1^2 n^{1-2\tau_1}, \qquad (3.11)
\end{aligned}
$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ is an intermediate value between $\boldsymbol{\beta}_{s'}$ and $\boldsymbol{\beta}^*_{s'}$. Thus, we have

$$
\begin{aligned}
Pr\{\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) \geq 0\} &\leq Pr\{\|\ell'_p(\boldsymbol{\beta}^*_{s'})\|_2 \geq (C_1 w_1/2) n^{1-\tau_1}\} \\
&= Pr\left\{\sum_{j \in s'} [\ell'_j(\boldsymbol{\beta}^*_{s'})]^2 \geq (C_1 w_1/2)^2 n^{2-2\tau_1}\right\} \\
&\leq \sum_{j \in s'} Pr\{[\ell'_j(\boldsymbol{\beta}^*_{s'})]^2 \geq (2m)^{-1} (C_1 w_1/2)^2 n^{2-2\tau_1}\}
\end{aligned}
$$

Also, by (C3), we have $m \leq w_2 n^{\tau_2}$, and also the following probability inequality

$$
\begin{aligned}
Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \geq (2m)^{-1/2} (C_1 w_1/2) n^{1-\tau_1}\} &\leq Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \geq (2 w_2 n^{\tau_2})^{-1/2} (C_1 w_1/2) n^{1-\tau_1}\} \\
&= Pr\left\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \geq c n^{1-\tau_1-0.5\tau_2}\right\} \\
&= Pr\left\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \geq n c n^{-\tau_1-0.5\tau_2}\right\} \qquad (3.12)
\end{aligned}
$$

where $c = C_1 w_1/(2\sqrt{2w_2})$ denotes some generic positive constant. Recall (3.2), by differentiation and rearrangement of terms, it can be shown as in Andersen and Gill (1982) that the gradient of $\ell_p(\boldsymbol{\beta})$ is

$$
\ell'_p(\boldsymbol{\beta}) \equiv \frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^n \int_0^\infty [\mathbf{z}_i - \bar{\mathbf{z}}_n(\boldsymbol{\beta}, t)] \, d\overline{N}_i(t). \qquad (3.13)
$$

where $\bar{\mathbf{z}}_n(\boldsymbol{\beta}, t) = \sum_{i \in R_j} \mathbf{z}_i \exp(\mathbf{z}_i^T \boldsymbol{\beta}) / \sum_{i \in R_j} \exp(\mathbf{z}_i^T \boldsymbol{\beta})$. As a result, the partial score function $\ell'_p(\boldsymbol{\beta})$ no longer has a martingale structure, and the large deviation results for continuous time martingale in Bradic et al. (2011) and Huang et al. (2013) are not directly applicable. The martingle process associated with $\overline{N}_i(t)$ is given by $M_i(t) = \overline{N}_i(t) - \int_0^t R_i(u) \exp(\mathbf{z}^T \boldsymbol{\beta}^*) d\Lambda_0(u)$.

Let t_j be the time of the jth jump of the process $\sum_{i=1}^n \int_0^\infty R_i(t) d\overline{N}_i(t)$, $j = 1, \ldots, N$ and $t_0 = 0$. Then, t_j are stopping times. For $j = 0, 1, \ldots, N$, define

$$
Q_j = \sum_{i=1}^n \int_0^{t_j} b_i(u) d\overline{N}_i(u) = \sum_{i=1}^n \int_0^{t_j} b_i(u) dM_i(u) \qquad (3.14)
$$

where $b_i(u) = \mathbf{z}_i - \bar{\mathbf{z}}_n(\boldsymbol{\beta}, u)$, $i = 1, \ldots, n$ are predictable, under no two component processes jumping at the same time and (D6), and satisfy $|b_i(u)| \leq 1$.

Since $M_i(u)$ are martingales and $b_i(u)$ are predictable, $\{Q_j, j = 0, 1, \ldots\}$ is a martingale with the difference $|Q_j - Q_{j-1}| \leq \max_{u,i} |b_i(u)| \leq 1$. Recall definition of N in Section 2, we define $C_0^2 n \leq N$, where C_0 is a constant. So, by the martingale version of the Hoeffding's inequality Azuma (1967) and under Condition (D7), we have

$$Pr(|Q_N| > nC_0x) \leq 2\exp\{-n^2C_0^2x^2/(2N)\} \leq 2\exp(-nx^2/2)$$

By (3.14), $Q_N = n\ell_p'(\boldsymbol{\beta})$ if and only if $\sum_{i=1}^n \int_0^\infty R_i(t)d\overline{N}_i(t) \leq N$. Thus, the left-hand side of (3.15) in Lemma 3.3 of Huang et al. (2013) is no greater than $Pr(|Q_N| > nC_0x) \leq 2\exp(-nx^2/2)$.

Now (3.12) can be rewritten as follows.

$$Pr\left\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq ncn^{-\tau_1-0.5\tau_2}\right\} \leq \exp\{-0.5nn^{-2\tau_1-\tau_2}\} = \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \quad (3.15)$$

By the same arguments, we have

$$Pr\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \leq -m^{-1/2}(C_1w_1/2)n^{1-\tau_1}\} \leq \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \quad (3.16)$$

The inequalities (3.15) and (3.16) imply that,

$$Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}^*_{s'})\} \leq 4m\exp\{-0.5n^{1-2\tau_1-\tau_2}\}$$

Consequently, by Bonferroni inequality and under conditions (C3) and (C4), we have

$$\begin{aligned}
\mathrm{Pr}\left\{\max_{\mathrm{s}\in\mathrm{S}_-^{\mathrm{m}}} \ell_{\mathrm{p}}(\boldsymbol{\beta}_{\mathrm{s}'}) \geq \ell_{\mathrm{p}}(\boldsymbol{\beta}^*_{\mathrm{s}'})\right\} &\leq \sum_{s\in S_-^m} Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}^*_{s'})\} \\
&\leq 4mp^m \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4\exp\{\log m + m\log p - 0.5n^{1-2\tau_1-\tau_2}\} \\
&\leq 4\exp\{\log w_2 + \tau_2 \log n + w_2 n^{\tau_2}\tilde{c}n^{\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4w_2\exp\{\tau_2\log n + w_2\tilde{c}n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= a_1\exp\{\tau_2\log n + a_2n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\}
\end{aligned}$$

$$= \quad o(1) \quad \text{as} \quad n \to \infty \tag{3.17}$$

for some generic positive constants $a_1 = 4w_2$ and $a_2 = w_2\tilde{c}$. By Condition (C5), $\ell_p(\boldsymbol{\beta}_{s'})$ is concave in $\boldsymbol{\beta}_{s'}$, (3.17) holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1 d_n n^{-\tau_1}$.

For any $s \in S^m_-$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\hat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in s'/s^* (i.e. $s' = \{s \cup (s^*/s)\} \cup (s'/s^*)$). By Condition (C3), it is seen that $\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^* \cup (s'/s^*)} - \boldsymbol{\beta}^*_{s^* \cup (s'/s^*)}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^* \cup (s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geq \|\boldsymbol{\beta}^*_{s^* \cup (s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geq \|\boldsymbol{\beta}^*_{s'/s^*}\|_2 = w_1 d_n n^{-\tau_1}$. Consequently,

$$\Pr\left\{\max_{\mathrm{s} \in \mathrm{S}^\mathrm{m}_-} \ell_\mathrm{p}(\hat{\boldsymbol{\beta}}_\mathrm{s}) \geq \min_{\mathrm{s} \in \mathrm{S}^\mathrm{m}_+} \ell_\mathrm{p}(\hat{\boldsymbol{\beta}}_\mathrm{s})\right\} \leq \Pr\left\{\max_{\mathrm{s} \in \mathrm{S}^\mathrm{m}_-} \ell_\mathrm{p}(\breve{\boldsymbol{\beta}}_{\mathrm{s}'}) \geq \ell_\mathrm{p}(\boldsymbol{\beta}^*_{\mathrm{s}'})\right\} = \mathrm{o}(1).$$

So, we have shown that

$$\Pr\left[\max_{\mathrm{s} \in \mathrm{S}^\mathrm{m}_-} \ell\{\hat{\boldsymbol{\alpha}}_\mathrm{s}(\mathrm{U})\} \geq \min_{\mathrm{s} \in \mathrm{S}^\mathrm{m}_+} \ell\{\hat{\boldsymbol{\alpha}}_\mathrm{s}(\mathrm{U})\}\right] \longrightarrow 0,$$

as $n \to \infty$. The theorem is proved. □

3.3 Numerical Studies

In this section, we assess the finite sample performance of the proposed procedure and compare it with other existing ones via Monte Carlo simulations and illustrate the proposed procedure by an empirical analysis of a genomic data set.

3.3.1 Simulation Studies

The main purpose of simulation studies is to assess the performance of the proposed procedure by comparing with the SIS procedure (Fan et al., 2010) and SJS procedure (Yang et al., 2016) for the Cox model. The model size selected by all three methods are set to be the same for the purpose of comparison. We vary the dimension of predictors p, the sample size n and the sample correlation ρ to examine their impact on the performance of the proposed procedure. We use the success rate of active predictors being selected and computing time as our criteria to compare the performance of screening procedures.

In our simulation, the predictors $\mathbf{x}$ are generated from a p-dimensional normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$. Two commonly-used covariance structures are used in our simulation:

(S1) Σ is compound symmetric. That is, $\sigma_{ij} = \rho$ for $i \neq j$ and equal 1 for $i = j$. We choose $\rho = (0.25, 0.5, 0.75)$.

(S2) Σ has autoregressive structure with $AR(1)$. That is, $\sigma_{ij} = \rho^{|i-j|}$. We choose $\rho = (0.25, 0.5, 0.75)$.

We generate the survival time from the Cox model with $h_0(t) = 1$, and the censoring time from a uniform distribution $U[0, 10]$. Three different coefficient function settings $\boldsymbol{\alpha}(u)$s are considered:

(a1) : $\alpha_1^{(1)}(u) = 1 + 2\sin(2\pi u)$, $\alpha_2^{(1)}(u) = 1 - 2\cos(2\pi u)$, $\alpha_3^{(1)}(u) = 0.5 + 2u^2$.

(a2) : $\alpha_1^{(2)}(u) = 5\sin(2\pi u)$, $\alpha_2^{(2)}(u) = 5\cos(2\pi u)$, $\alpha_3^{(2)}(u) = 2.5 + 5u^2$.

(a3) : $\alpha_1^{(3)}(u) = e^{0.5u}$, $\alpha_2^{(3)}(u) = 2(u^3 + 1.5(u - 0.5)^2)$, $\alpha_3^{(3)}(u) = 2u$.

We set $n = 200$ and 400, and $p = 2,000$ and $5,000$. For the feature screening model size, we follow Liu et al. (2014) and set $m = [n^{0.8}/\log(n^{0.8})]$, where $[a]$ denotes the integer part of a. For each combination of different inputs, we conduct $1,000$ replications of Monte Carlo simulation.

To illustrate the performance of a statistical procedure in survival data analysis, we want the censoring rates to lie within a reasonable range. Table 3.1 depicts the censoring rates for the 18 combinations of covariance structure, sample correlation ρ and the values of $\boldsymbol{\alpha}(u)$. The censoring rates range from 22% to 37%, which is reasonable to carry out simulation studies.

Table 3.1: Censoring Rates

	$\rho = 0.25$			$\rho = 0.5$			$\rho = 0.75$		
Σ	(a1)	(a2)	(a3)	(a1)	(a2)	(a3)	(a1)	(a2)	(a3)
S1	.276	.367	.223	.277	.356	.260	.277	.340	.248
S2	.275	.365	.265	.279	.358	.283	.278	.347	.245

We compare the performance of feature screening procedures using the following two criteria: P_s: the proportion that an individual active predictor is selected, and

P_a: the proportion that all active predictors are selected. It is expected that the performance of the proposed varying-coefficient sure joint screening (VSJS) procedure depends on the following factors: the structure of the covariance matrix, the values of $\boldsymbol{\alpha}(u)$, the dimension of all candidate features p, sample correlation ρ and the sample size n.

Table 3.2: Comparison between VSJS, SIS and SJS with $\Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^T$ (n=200)

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n=200$, $p=2000$ and $\rho=.25$														
$\alpha^{(1)}$	.989	1	1	.989	74.5	.796	.747	.990	.580	9.5	.499	.419	.936	.190	3.6
$\alpha^{(2)}$	.999	.998	.999	.996	67.7	.016	.002	1	0	8.3	.018	.037	.999	.002	2.4
$\alpha^{(3)}$	1	.810	.993	.803	82.2	1	.771	.992	.763	6.0	1	.785	.996	.781	2.8
	$n=200$, $p=2000$ and $\rho=.5$														
$\alpha^{(1)}$	.970	.976	.915	.868	68.9	.621	.557	.968	.325	9.2	.392	.311	.863	.092	2.9
$\alpha^{(2)}$	.922	.922	.990	.848	66.8	.006	.003	1	0	7.8	.020	.052	.997	0	2.5
$\alpha^{(3)}$	.998	.617	.938	.581	74.8	.999	.611	.932	.573	5.3	1	.574	.932	.542	3.2
	$n=200$, $p=2000$ and $\rho=.75$														
$\alpha^{(1)}$	.628	.670	.682	.259	62.4	.357	.316	.879	.093	9.4	.247	.211	.701	.031	3.0
$\alpha^{(2)}$	.485	.535	.738	.204	67.3	.005	.001	1	0	6.8	.018	.059	.935	0	3.4
$\alpha^{(3)}$	.910	.361	.686	.247	62.5	.987	.341	.736	.250	5.3	.958	.286	.644	.181	3.4
	$n=200$, $p=5000$ and $\rho=.25$														
$\alpha^{(1)}$	1	1	.993	.993	464.0	.721	.649	.983	.456	15.4	.391	.326	.865	.097	32.9
$\alpha^{(2)}$	.996	.994	1	.990	416.3	.004	.004	1	0	18.1	.007	.016	.994	0	17.6
$\alpha^{(3)}$	1	.708	.984	.694	451.5	1	.684	.974	.667	15.2	1	.627	.980	.615	16.8
	$n=200$, $p=5000$ and $\rho=.5$														
$\alpha^{(1)}$	.925	.930	.845	.725	412.7	.496	.430	.954	.199	22.9	.281	.224	.779	.040	16.8
$\alpha^{(2)}$	.856	.876	.976	.740	423.7	.005	.002	1	0	16.1	.007	.030	.968	0	18.9
$\alpha^{(3)}$	.992	.508	.884	.446	390.4	.999	.455	.866	.38	15.2	.998	.435	.878	.383	24.0
	$n=200$, $p=5000$ and $\rho=.75$														
$\alpha^{(1)}$	.510	.501	.504	.121	398.1	.261	.218	.803	.042	15.3	.135	.140	.541	.010	20.3
$\alpha^{(2)}$	.372	.399	.625	.093	396.6	.002	0	.999	0	14.9	.006	.022	.867	0	22.2
$\alpha^{(3)}$	.892	.276	.597	.158	369.5	.977	.258	.624	.159	13.3	.909	.164	.493	.075	24.7

Tables 3.2 and 3.3 report P_s and P_a of VSJS, SIS and SJS for the active predictors under (S1). Overall, VSJS outperforms both SIS and SJS for all the three sets of $\boldsymbol{\alpha}(u)$ in terms of P_s and P_a. For (a1), VSJS achieves high success rate in detecting signals of $\alpha_1^{(1)}$ and $\alpha_2^{(1)}$, while SIS and SJS fail from time to time. We next consider the performance of VSJS under (a2). For the zero centered $\alpha_1^{(2)}$

Table 3.3: Comparison between VSJS, SIS and SJS with $\Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^T$ (n=400)

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n=400$, $p=2000$ and $\rho=.25$														
$\alpha^{(1)}$	1	1	1	1	217.7	1	.960	1	.960	8.8	.859	.805	.999	.686	5.8
$\alpha^{(2)}$	1	1	1	1	205.9	.020	.001	1	0	7.9	.010	.076	1	0	5.6
$\alpha^{(3)}$	1	1	1	1	215.3	1	.974	1	.974	8.3	1	.997	1	.997	4.9
	$n=400$, $p=2000$ and $\rho=.5$														
$\alpha^{(1)}$	1	1	1	1	190.2	.900	.871	.999	.779	8.5	.736	.607	.998	.437	4.6
$\alpha^{(2)}$	1	1	1	1	184.3	.010	.001	1	0	8.5	.023	.133	1	.002	6.3
$\alpha^{(3)}$	1	.988	1	.988	199.5	1	.918	.997	.916	8.2	1	.944	1	.944	5.1
	$n=400$, $p=2000$ and $\rho=.75$														
$\alpha^{(1)}$	.984	.991	.976	.955	169.0	.655	.566	.997	.349	8.6	.474	.356	.955	.155	6.3
$\alpha^{(2)}$	.998	.995	1	.994	162.2	.001	0	1	0	9.5	.035	.193	.999	.004	6.6
$\alpha^{(3)}$	1	.733	.982	.719	162.8	1	.676	.968	.657	8.2	1	.576	.938	.540	6.1
	$n=400$, $p=5000$ and $\rho=.25$														
$\alpha^{(1)}$	1	1	1	1	1202	.963	.957	1	.920	21.6	.963	.957	1	.920	21.6
$\alpha^{(2)}$	1	1	1	1	1164	.006	.001	1	0	20.6	.004	.038	1	.001	31.1
$\alpha^{(3)}$	1	1	1	1	1180	1	.960	1	.960	18.2	1	.993	1	.993	36.5
	$n=400$, $p=5000$ and $\rho=.5$														
$\alpha^{(1)}$	1	1	1	1	1086	.849	.798	.999	.669	21.0	.849	.798	.999	.669	21.1
$\alpha^{(2)}$	1	1	1	1	1101	.001	0	1	0	22.2	.011	.071	1	.002	32.1
$\alpha^{(3)}$	1	.975	1	.975	1071	1	.840	.998	.838	19.6	1	.872	1	.872	40.3
	$n=400$, $p=5000$ and $\rho=.75$														
$\alpha^{(1)}$	1	1	.980	.980	929.0	.562	.426	.994	.224	21.0	.336	.267	.933	.073	35.9
$\alpha^{(2)}$	.994	.992	.997	.988	936.7	.001	0	1	0	20.8	.016	.109	1	.001	35.3
$\alpha^{(3)}$	.995	.621	.926	.586	909.6	1	.580	.935	.535	18.3	.999	.446	.900	.401	46.1

and $\alpha_2^{(2)}$, VSJS successfully detects their variation signal and achieves high success rates. As comparison, SIS and SJS fail to identify $\alpha_1^{(2)}$ and $\alpha_2^{(2)}$ as active predictors completely in (a2). In general, VSJS still performs better to some extent in (a3), though SIS slightly outperforms VSJS in a few cases.

Tables 3.2 and 3.3 clearly show how performances are affected by sample correlation ρ, predictor dimension p and sample size n. When ρ increases, n decreases or p increases, all three methods perform worse under (S1). Compared to SIS and SJS, VSJS's performance is more resistant to these changes. Also, Tables 3.2 and 3.3 depict that VSJS is more computational inefficient comparing to SIS and SJS.

Tables 3.4 and 3.5 report P_s and P_a of VSJS, SIS and SJS for the active

Table 3.4: Comparison between VSJS, SIS and SJS with $\Sigma = (\rho^{|i-j|})$ (n=200)

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n = 200$, $p = 2000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	76.9	1	1	.988	.988	5.2	.856	.809	.997	.684	2.6
$\alpha^{(2)}$	1	1	1	1	70.8	.042	.116	1	.008	5.9	.047	.027	1	0	2.4
$\alpha^{(3)}$	1	1	1	1	86.6	1	1	1	1	7.1	1	.981	1	.981	3.0
	$n = 200$, $p = 2000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	73.1	1	1	.999	.999	8.2	.889	.792	.990	.690	2.5
$\alpha^{(2)}$	1	1	1	1	67.6	.166	.611	1	.145	5.8	.052	.065	1	.011	2.4
$\alpha^{(3)}$	1	1	1	1	82.8	1	1	1	1	7.7	1	.977	1	.977	3.1
	$n = 200$, $p = 2000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	75.5	1	1	1	1	5.2	.877	.768	.990	.642	3.0
$\alpha^{(2)}$	1	1	1	1	68.6	.722	.968	1	.720	5.8	.125	.417	.997	.076	2.6
$\alpha^{(3)}$	1	.997	1	.997	79.4	1	1	1	1	8.4	1	.926	.991	.917	3.1
	$n = 200$, $p = 5000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	456.4	.968	.997	1	.965	15.4	.785	.734	.989	.559	16.1
$\alpha^{(2)}$	1	1	1	1	463.8	.016	.067	1	.004	14.6	.016	.022	.999	0	14.9
$\alpha^{(3)}$	1	1	.998	.998	477.1	1	.999	1	.999	16.2	1	.967	1	.967	20.1
	$n = 200$, $p = 5000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	451.1	1	1	1	1	13.1	.799	.730	.979	.543	13.2
$\alpha^{(2)}$	1	1	1	1	439.9	.121	.501	1	.103	14.3	.030	.025	1	.003	16.0
$\alpha^{(3)}$	1	1	1	1	475.4	1	1	1	1	15.8	1	.966	.997	.963	20.3
	$n = 200$, $p = 5000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	448.2	1	1	1	1	15.4	.844	.685	.987	.538	19.0
$\alpha^{(2)}$	1	1	1	1	427.3	.627	.938	1	.626	14.8	.062	.327	1	.040	15.9
$\alpha^{(3)}$	1	.996	1	.996	453.9	1	1	1	1	14.4	1	.916	.980	.896	23.3

predictors under (S2). Overall, VSJS still outperforms SIS and SJS. It is worth noting that all the three methods have much better performance under (S2) than previous cases under (S1), especially when the correlation ρ is larger. In (a1), VSJS and SIS both perform perfectly and slightly better than SJS. When we consider (a2), SIS and SJS perform better under (S2) and successfully identify $\alpha_1^{(2)}$ and $\alpha_2^{(2)}$ from time to time. However, VSJS again outperforms them in (a2). For (a3), all three methods achieve almost 100 percent success rate for selecting active predictors. Among them, SJS miss some active predictors in a few cases.

We can conclude from Tables 3.4 and 3.5 that SIS and SJS tend to perform better when ρ increases, n increases or p decreases. For VSJS, it performs perfectly in all three setting under (S1). Similarly, Tables 3.4 and 3.5 show that VSJS is

Table 3.5: Comparison between VSJS, SIS and SJS with $\Sigma = (\rho^{|i-j|})$ $(n = 400)$

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n = 400$, $p = 2000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	229.6	1	1	1	1	8.6	.991	.979	1	.970	6.3
$\alpha^{(2)}$	1	1	1	1	223.3	.083	.251	1	.036	8.5	.047	.040	1	.001	5.2
$\alpha^{(3)}$	1	1	1	1	240.1	1	1	1	1	11.9	1	1	1	1	7.0
	$n = 400$, $p = 2000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	225.9	1	1	1	1	7.5	.992	.959	1	.951	5.3
$\alpha^{(2)}$	1	1	1	1	226.1	.387	.922	1	.382	8.8	.070	.263	1	.031	5.2
$\alpha^{(3)}$	1	1	1	1	236.8	1	1	1	1	8.5	1	1	1	1	7.3
	$n = 400$, $p = 2000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	217.9	1	1	1	1	8.9	.979	.907	1	.886	6.4
$\alpha^{(2)}$	1	1	1	1	218.4	.969	1	1	.969	9.1	.139	.598	1	.080	5.8
$\alpha^{(3)}$	1	.999	1	.999	227.8	1	1	1	1	11.9	1	.997	1	.997	7.6
	$n = 400$, $p = 5000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	1264	1	1	1	1	20.6	.988	.962	1	.952	29.5
$\alpha^{(2)}$	1	1	1	1	1265	.054	.183	1	.018	18.7	.029	.032	1	0	28.8
$\alpha^{(3)}$	1	1	1	1	1215	1	1	1	1	20.8	1	1	1	1	33.8
	$n = 400$, $p = 5000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	1274	1	1	1	1	20.5	.976	.924	1	.900	32.5
$\alpha^{(2)}$	1	1	1	1	1256	.318	.884	1	.312	19.9	.038	.162	1	.017	29.1
$\alpha^{(3)}$	1	1	1	1	1194	1	1	1	1	20.6	1	.999	1	.999	35.6
	$n = 400$, $p = 5000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	1202	1	1	1	1	20.7	.969	.902	1	.871	36.9
$\alpha^{(2)}$	1	1	1	1	1225	.954	1	1	.954	21.9	.085	.548	1	.051	29.9
$\alpha^{(3)}$	1	1	1	1	1139	1	1	1	1	29.5	1	.995	1	.995	34.6

more time consuming than SIS and SJS.

3.3.2 Real Data Analysis

We analyze The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) data on liver hepatocellular carcinoma to illustrate the application. Liver hepatocellular carcinoma is the most common form of liver cancer and the third cancer death cause worldwide. Zhang and Sun (2015) studies 17,255 patients in the SEER cancer registry and suggests that age is a prognosis factor for liver cancer. Therefore, we consider age as the univariate covariate for coefficient functions, allowing the effects of gene expression on survival time to vary with age. After removing 5 subjects whose survival time are zero, we obtain 354 subjects with gene expressions (IlluminaHiSeq RNA-seq v2 platform), age at diagnostic, and survival months. We apply $\log 2$ transformation to gene expressions and analyze 14683 genes that have more than 90% nonzero observations.

For VSJS, we use a linear combination of 5 B-spline basis to approximate the varying coefficient functions. As a result, VSJS retains 23 ($354^{0.8}/\log(354^{0.8})$) genes and the partial likelihood function value for the corresponding model is -544.9. With the same number of genes retrained, the resulting partial likelihood function values for SIS and SJS are -589.2 and -588.4, respectively. Simultaneous modeling of the screened 23 genes shows a clear advantage of VSJS in terms of higher partial likelihood value. To better understand the screening result of VSJS, we further apply the backward selection procedure to those 23 genes and obtain a more parsimonious model. Specifically, each backward elimination step removes a gene with the smallest likelihood ratio test statistic until all the genes are significant at level 0.05. Table 3.6 provides the final list of 11 genes after applying the backward elimination and Figure 3.1 depicts their varying coefficients. Our literature search finds that those 11 genes are all associated with cancer risk and some genes such as GTPBP4 (Liu et al., 2017) and SLC2A2 (Kim et al., 2017) are promising prognostic factors for hepatocellular carcinoma. To test whether those 11 genes have varying coefficients versus constant coefficients, a test of H_0: $\alpha_j(u) = \alpha_j$ for some constant α_j versus H_1: $\alpha_j(u) \not\equiv \alpha_j$ can be conducted for each j in the selected gene set. The test result is shown in Table 3.7 and all the genes except DYNC1LI1 have significant varying coefficient functions of age at the 5% level of significance. There is no evidence of their timing varying effects in the current medical literature, but our study may suggest some evidence for potential granular investigation on those genes.

Table 3.6: Genes selected by backward elimination

Gene Name	ANLN	CEP55	DYNC1LI1	GTPBP4
LRT Stat	15.869	14.137	18.171	22.658
p value	0.00723	0.0148	0.00274	< 0.001
Gene Name	SLC2A1	KIF2C	KIF20A	KPNA2
LRT Stat	18.465	26.261	15.839	14.511
p value	0.00241	< 0.001	0.00731	0.0127
Gene Name	LIMS2	TRIP13	UCK2	
LRT Stat	23.093	17.517	14.671	
p value	< 0.001	0.00361	0.0119	

Table 3.7: LRT statistics and p-values for the varying coefficients of the final selected genes

Gene Name	ANLN	CEP55	DYNC1LI1	GTPBP4
LRT Stat	15.058	10.495	8.268	19.036
p value	0.00458	0.0328	0.0822	0.000773
Gene Name	SLC2A1	KIF2C	KIF20A	KPNA2
LRT Stat	17.473	24.253	15.183	14.238
p value	0.00156	0.000071	0.00433	0.00657
Gene Name	LIMS2	TRIP13	UCK2	
LRT Stat	23.097	16.191	13.803	
p value	0.000121	0.00277	0.00795	

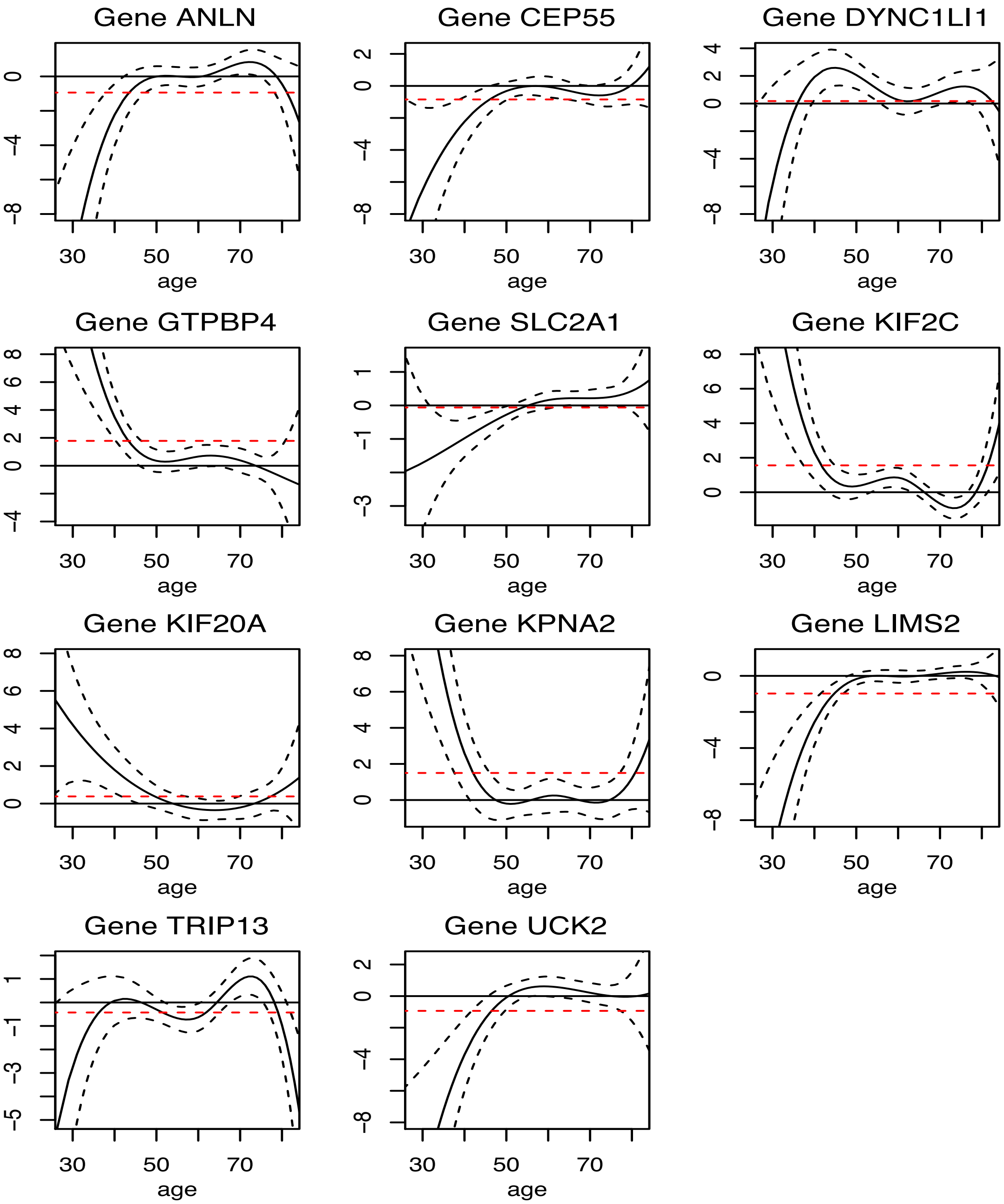

Figure 3.1. Estimated coefficient functions and the pointwise conference intervals of selected genes. The red line represents the average level of the varying coefficient functions.

Chapter 4

Feature and Interaction Selection via “Iteratively Kings’ Forests”

4.1 A Forest-based Procedure for Feature and Interaction Selection

4.1.1 Tree Structure: An All-inclusive Search Engine

The latest researches suggest that forest algorithms perform excellent in selecting interaction effects. We believe that most credit goes to the properties of tree structure. First of all, when constructing a tree, the split in every node, except for the root node, is based on the previous splits of the node’s ancesters. That makes tree structure a natural structure to model interaction effects. Moreover, the greedy algorithm used to search and split each node allows tree structure to include the variable that brings the most significant instant improvement. This improvement may result from either a marginal effect or an interaction effect. To conclude, tree structure is just like an all-inclusive search engine for important marginal effects and interaction effects of different orders. The mathematical form of a tree, say $q(\mathbf{x})$, is given as follows.

Denote the p-dimensional candidate variables as $\mathbf{x} = (x_1, \ldots, x_p)$. $q(\mathbf{x})$ could be expressed as a data-driven simple function, which is a linear combination of k indicator functions. Assuming the tree depth is d, the score a_i of an indicator function is determined by all d variables in the path from the corresponding leaf node to the root node.

$$q(\mathbf{x}) = \sum_{i=1}^{k} a_i I(\mathbf{x}_i \in \mathbf{B}_i) = \sum_{i=1}^{k} a_i I(x_{i_1} \in B_{i_1}, \ldots, x_{i_d} \in B_{i_d}). \tag{4.1}$$

Here $\mathbf{a} = (a_1, \ldots, a_k)^T$ are the scores of leaf nodes and $\mathbf{x}_i$ is the subset of $\mathbf{x}$ chosen for i-th indicator function. For any B_{i_j}, it is an interval with form $(-\infty, s_{i_j})$ or $[s_{i_j}, \infty)$, where s_{i_j} is given for spliting variable x_{i_j}.

For $i-$th indicator function, the selected variable set $\mathbf{x}_i$ may include marginal effects, all or part of an interaction effect or even useless variables due to randomicity. Given this, we want to maximize the chance to exclude random useless variables and include all effects of different orders in the forest. Furthermore, we try to separate one effect from the other, and therefore outline the hidden model form.

4.1.2 Forest Algorithms

Random forest is well known for many good characteristics. It is applicable to high dimensional data with a relative low number of observations, a large amount of noise and highly correlated variables. At the same time, random forest is less prone to over-fitting and can handle the problem of imbalanced classes. However, when the truly imformative features is very sparse among a huge number of features, the prediction performance of RF declines significantly. This requires us to do feature selection. Breiman (2001) proposed to rank the variable importance using the "Permutation Variable Importance Measure"(PVIM). PVIM of the j-th ($j = 1, \ldots, p$) feature X_j is defined as

$$PVIM_t(X_j) = \frac{\sum_{i \in B_t}(Y_i - \hat{Y}_{it}^{*})^2 - \sum_{i \in B_t}(Y_i - \hat{Y}_{it})^2}{|B_t|} \tag{4.2}$$

Here B_t is the OOB sample for tree t, $t = 1, \ldots, ntree$. $\hat{Y}_{it}$ and $\hat{Y}_{it}^{*}$ are the predictions for observation i gotton from the tree t before and after permuting X_j.

The final importance measure is averaged over all trees

$$PVIM(X_j) = \sum_{t=1}^{ntree} \frac{PVIM_t(X_j)}{ntree} \tag{4.3}$$

The main idea of PVIM is to see how much prediction accuracy will decrease if we “kill” a variable. If one feature is randomly permuted, all its modelled effects on the response will be destroyed. That idea is very powerful because it could measure both the variable’s marginal effect and interaction effects as a whole.

However, the power of PVIM will be sharply weaken if the effects related to the variable are not adequately modelled in RF because of three possible reasons. First, the important variables may only be selected in a small proportion of trees of the whole forest; Sometimes, the node in which an important variable is selected is far from the root node. Therefore, the variable could only explain a small part of data and lead to less overall prediction accuracy improvement; Moreover, the variables participating in an interaction effect may not be in the same path. That is, the interaction effect is not adequately modelled.

A intuitive solution to these issues is to iteratively reweight variables’ importance based on certain criterion. For example, mean decrease in Gini impurity(Archer and Kimes, 2008; Basu et al., 2018) is used to update variables’ weights under different scenarios. By giving important variables larger weights, we more frequently select them for tree construction and move them up in trees. Also, variables that partipate in interaction effects are more likely to be in the same path. The reweighting process also do feature selection by ranking variable importances, which also improves the model interpretability and prediction accuracy. However, main existing forest algorithms have the following drawbacks, which indicates the potential of significant improvements.

First of all, impurity decrease of node spliting is a criteria of training process, and might not be good especially when p is large. When the parameter $mtry$ is large, there will likely be a random variable giving a larger impurity decrease than the important variables among these $mtry$ selected variables. When p is much larger than $mtry$ and the truly imformative features are sparse, it is likely that no important variable is included in these $mtry$ selected variables. As a result, some random variables will have large weights and there will be a large proportion of useless trees. Therefore, we need a new criteria to identify useful trees and variables. Furthermore, existing forest algorithms do not leave room for incorporating prior knowledge of the mechanism. When the hidden model is sparse or prior knowledge is given, a proper use of prior information might be helpful for outlining the hidden model structure. Finally, we still do not have criterion to determine the degree of interaction effects an important variable participates in because two seperate interaction effects could be in the same path and treated as one high order interaction effect.

4.1.3 King's Forests

To solve the previous mentioned three drawbacks, we revise the structure of trees in the forest and propose a new forest algorithm named "King's Forests".

To start with, we define that a variable is considered to be important if and only if it has marginal effect or participates in at least one interaction effect. Given the prior knowledge that one variable is important, we treat it as a "king" and use it to split the root node of every tree in the forest. Accordingly, all other variables that participate in at least one common interaction effect of the "king" together is called the "core team" of the "king". Through this revision, we could make room for incorporating prior knowledge and furthermore fully explore interaction structures related to the "king".

The PVIM of "king" is like the king's power and used to measure the overall effect related to the "king" on the response. To fully explore the king's power, the "king" should try the best to find all members of its "core team" and put those from same interaction in the same path. To achieve that, we iteratively select the trees with positive king's PVIM, and increase the weights of variables in the tree by the magnitude of king's PVIM, which indicates the importance level of the tree. That is, instead of impurity decrease, we increase variables' weights only when they are selected in a "useful" tree. As a result, the members of "core team" will likely have larger weights and move to upper layers close to the "king".

Based on the updated weights, we could do feature screening and select the top ranked variables. If the depth of trees is setted as d, the king's forests could model at most order d interaction effects related to the "king". Based on the selected features, we fit another d “king’s forests” of depth from 1 to d. For a forest of depth $m (1 \leq m \leq d)$, it consist of $l = n_t * 2^{m-1}$ paths, or candidate order m interactions, among which there are $f \leq l$ different paths. For each of the f different paths, we get its number of repetitions, sum of king’s PVIMs of corresponding trees and sum of impurity decrease of the last node. To select the most likely order m interactions, we rank all f paths and output three short lists respectively based on these three criterion. Generally speaking, the larger the repetitions, sum of king’s PVIM or sum of impurity decrease (sum of ID) an interaction has, the more likely it is an important interaction the “king” participates in.

The detailed algorithm for constructing “king’s forests” is outlined as follow:

1. From given prior knowledge, choose an important variable, say x_{k_1}, as the “king”.

2. For x_{k_1}, construct a “king’s forest” of size n_t and depth d for it.

3. For the t-th iteration($t \geq 1$), calculate the king’s PVIM using OOB sample for each tree. Set initial weights $\mathbf{w}^{(0)}$ as a positive constant w_0. Update variables’ weights by

$$w_i^{(t)} = w_i^{(t-1)} + \sum_{j=1}^{n_t} pvim_j * I(pvim_j > 0, x_i \text{ is selected in tree j}) \tag{4.4}$$

   Keep the weights updating procedure until n_{ite} iterations or king’s PVIM stops increasing.

4. Rank the features base on updated weights, and select the top ranked variables with weights higher than the average weight.

5. Construct d “king’s forests” from depth 1 to d based on the selected features.

6. For the forest of depth m, output three ranked short lists of most likely order m interactions.

The proposed “king’s forests” algorithm gives the most likely interaction effects and “core team” related to the “king”. However, to distinguish two separate low-order interaction effects from one high-order interaction effect, we evaluate the trend of king's PVIMs as the depth m increases. For example, if the king's PVIMs significantly increases as m increases from 2 to 3, the most likely explanation is that the "king" participates in at least one third-order interactions, which is not modelled in the previous forest of depth 2. Otherwise, if its PVIMs doesn't increase as the depth increases, the "king" likely does not have a higher order interaction effect.

4.1.4 Iteratively Kings' Forests

While the "king's forests" algorithm gives a framework to discover interaction effects related to the "king", it just focuses on one important variable. To explore more for a whole picture, a natural idea is to iteratively implement the "king's forests" algorithm using different variables as the "king". Among the variables have not been chosen as "king", we choose the variable whose weight is ranked first as the next "king", which is considered the most important based on current best knowledge.

Besides, feature screening is also conducted during this iterative procedure. For i-th iteration, we select the top ranked α percent variables of i-th "king's forest" and denote the index set as I_i. We start with an index set S_0 of all p variables. For iteration $i \geq 1$, we get the intersection set $S_i = S_{i-1} \cap I_i$. We keep constructing "king's forests" until the size of S_i is smaller than a pre-specified number $ss = [n/log(n)]$. For each "king's forests", we also save the king's PVIMs and the top ranked interaction effects of different orders. For each depth m, we give PVIMs of different "kings" and three short lists of interactions with a size of a pre-specified n_{top}.

The details of iteratively "king's forests" are as follows

1. Choose an important variable x_{k_1} as the "king" based on prior knowledge. If no prior knowledge is given, just randomly choose x_{k_1}. The initial index set is $S_0 = \{1, \dots, p\}$.

2. Construct a "king's forests".

3. Select the α percent top ranked variables index set I_i, and get the intersection set $S_i = S_{i-1} \cap I_i$.

4. Save the lists of ranked interaction effects and king's PVIMs, and choose the next "king".
5. Iteratively conduct step 2 to 4 until the size of S_i is smaller than $ss = [n/log(n)]$
6. Output the intersection S, PVIMs of different "kings" and three short lists for potential interaction effects of different order.

To conclude, the "Iteratively Kings' Forests"(iKF) algorithm proposes a framework for discovering hidden model structure through kings' PVIMs, selected features index set and lists of ranked interactions. If one variable is chosen as a "king" and has positive king's PVIMs, it should be important. If the PVIMs increases with the depth of forest, it participates in some high-order interactions. If one variable is in the selected feature index set, it is an indicator that the variable is important. Furthermore, when an interaction is selected into the short lists of top ranked interaction, it could be an important interaction. Especially when its number of repetitions is large comparing to the forest size n_t, its sum of PVIM is large, or the interaction show up more than once in different order, it is very likely a very important interaction.

4.2 Two Examples

In this part, we use two examples, (a1) and (b1), to demonstrate how the proposed algorithm select important features and interactions. In these two cases, the p-dimensional predictors $\mathbf{x}$ and the random error e are independently generated from standard normal distribution. We set the sample size $n = 300$, the number of predictors $p = 200$ and the parameter scale $s = 2$. For the parameters of the "Iteratively Kings' Forests", we set the forest size $n_t = 100$, tree depth $d = 4$, the size of screened variable set $ss = [n/log(n)]$ and the length of potential interaction effect list $n_{top} = 10$. When constructing a tree, we search $mtry = [p/2]$ variables for each node and $n_{quantile} = 9$ quantile positions for each splitting. For any node, if the number of observations in it is less than $s_{node} = 3$, we stop splitting it and use it as a leaf node.

(a1): $y = s * x_1 * x_3 - s * x_5 * (x_7 < 0.2) + e$

We apply the "Iteratively Kings' Forests" method on the dataset $(y, \mathbf{x})$ generated by model (a1). Kings' PVIMs, selected index set and top ranked interactions lists are given for analyzing the hidden model mechanism.

Table 4.1: Kings' PVIMs for Model (a1) for Forests of Different Depths

King	$Depth = 1$	$Depth = 2$	$Depth = 3$	$Depth = 4$
5	4.65	5.67	5.72	6.34
7	-0.21	1.79	1.26	0.93
3	0.99	3.75	4.03	4.47
1	-0.16	3.48	5.23	4.44
10	-0.19	0.33	0.33	0.49
146	0.68	0.59	0.53	0.34
91	0.61	0.31	0.77	0.16

The method iteratively chooses seven variables as "kings" and give corresponding kings' PVIMs from depth 1 to 4 in Table 4.1. When the depth is 1, kings' PVIMs show that variables 3, 5, 91 and 146 may have marginal effects on the response, while variables 7, 1 and 10 are marginally insignificant. Among them, variable 5 has a very strong marginal effect. When the depth increases to 2, PVIMs of variables 1, 3, 5 and 7 sharply increase. This indicates that they all participate in some pairwise interaction effects. Furthermore, it seems that variable 1 might also have some third order interactions, and variables 3 and 5 might have slight forth order interactions according to the trend of their PVIMs.

Table 4.2: The Rank of Selected Important Variables for Model (a1)

5 7 3 1 110 104 193 48 146 122 49 103 47 190 64 12 65 163 72 159 118 172 196 117

Furthermore, the method also screens the features. Variables selected in Table 4.2 are ranked based on their weights. As we can see, all four variables of model (a1) are selected and ranked as top four.

Table 4.3: Top Ten Second Order Interactions Ranked by the Sum of PVIM for Model (a1)

Interactions	Sum of ID	Layer 1	Layer 2	Repetition	Sum of PVIM
1	27592	5	7	72	493
2	76970	1	3	93	434
3	52956	3	1	88	419
4	39879	1	5	72	286
5	60392	3	5	82	280
6	89495	7	5	91	202
7	4845	5	104	15	96
8	7726	5	3	18	85
9	52028	146	5	98	84
10	38692	91	5	76	81

Finally, we give the top ten possible interactions from order 2 to 4 ranked by the sum of king's PVIM. For second order interaction effects, the top ten pairs are given in Table 4.3. Among them, pairwise interactions (5, 7) and (1, 3) both show up twice in different orders. That is, when the "king" is 5, it tends to select variable 7 to split the node of next layer, and vice versa. Similar things happen to interaction (1, 3). According to all three criterion, (5, 7) and (1, 3) are selected and ranked high. Considering that we only have $n_t = 100$ trees in each forest, the repetition times of the top seven pairs show that they dorminate the corresponding kings' forests of depth 2 because they show up in at least 64% of trees. Variable 5 seems show up everywhere. However, we don't think variable 5 has second order interaction effects with variable 3 and 1 because it has strong marginal effect and only select variable 7 frequently when it is the "king". All those signals together strongly suggest that (1, 3) and (5, 7) are important second order interaction effects. Table 4.4 gives top ten third order interaction effects. Among them, we can see that third order interactions (1, 5, 3), (3, 1, 5), (3, 5, 7) and (1, 5, 7) show up most frequentyly.

Table 4.4: Top Ten Third Order Interactions Ranked by the Sum of PVIM for Model (a1)

Interactions	Sum of ID	Layer 1	Layer 2	Layer 3	Repetition	Sum of PVIM
1	29270	1	5	3	82	466
2	18888	3	1	5	70	399
3	16683	1	5	7	62	292
4	16222	3	5	7	54	204
5	5814	5	3	7	26	181
6	12830	5	3	1	34	173
7	2044	3	1	82	17	110
8	1489	1	3	22	12	94
9	2160	1	3	146	14	83
10	9701	7	5	3	42	79

Table 4.5: Comparison of Average PVIM Between Second and Third Order Interactions

Interactions	Average PVIM	Interactions	Average PVIM
(1, 5)	286/72=3.97	(1, 5, 3)	466/82=5.68
(3, 1)	419/88=4.76	(3, 1, 5)	399/70=5.70
(3, 5)	280/82=3.41	(3, 5, 7)	204/54=3.78
(1, 5)	286/72=3.97	(1, 5, 7)	292/62=4.71

However, if we calculate the average PVIM of them and compare with the average of corresponding second order interactions, we have the Table 4.5. It shows that average PVIM of third order interactions (1, 3, 5) and (1, 5, 7) are much larger than that of interaction (1, 3) and (1, 5). Therefore, the researcher may need to look into these two possible third order interactions.

Table 4.6: Top Ten Order 4 Interactions Ranked by the Sum of PVIM for Model (a1)

Interactions	Sum of ID	Layer 1	Layer 2	Layer 3	Layer 4	Repetition	Sum of PVIM
1	7299	3	5	1	7	46	242
2	3519	3	5	1	193	13	65
3	993	3	5	1	127	9	63
4	1393	5	7	104	1	9	54
5	1217	5	7	1	190	7	47
6	1720	1	3	5	157	9	47
7	3278	10	3	5	7	20	42
8	583	1	5	7	140	6	42
9	573	3	5	176	1	4	40
10	566	1	5	80	3	7	39

When we set the tree depth $d = 4$, we get top ten possible order 4 interaction effects given in Table 4.6. Among them, (1, 3, 5, 7) is obviously larger than other nine order 4 interactions. It could be either an additional order 4 interaction effect, or two separate second order interactions (1, 3) and (5, 7), which we already identify before. Considering that all four variables' PVIMs don't increase significantly when the depth increase from 3 to 4, we tend to conclude there is no additional order 4 interaction effect.

(b1): $y = s * x_1 * sign(1 + x_3) * \sin(x_5) + e$

We apply the "Iteratively Kings' Forests" method on the dataset $(y, \mathbf{x})$ generated by model (b1). Kings' PVIMs, selected index set and top ranked interactions lists are given for analyzing the hidden model mechanism.

The method iteratively chooses five variables as "king" and give corresponding king's PVIMs of depth from 1 to 4 in table 2.60. As the PVIMs of depth 1 forest show, no king variable has marginal effect on the response. When the depth increases to 2, PVIMs of "king" 1 and 5 increase a lot. We will guess that variable 1 and 5 participate in some second order interactions. When the depth increases to 3, king's PVIMs of variable 1, 3 and 5 slightly increase. This indicates that they might participate in some important third order interaction effects.

Table 4.7: PVIM of Kings for Model (b1) in Forests of Different Depth

King Variable	$Depth = 1$	$Depth = 2$	$Depth = 3$	$Depth = 4$
63	-0.03	0.15	-0.07	-0.03
1	0.20	1.11	1.33	1.54
5	-0.10	1.12	1.23	1.17
3	0.06	0.06	0.12	0.25
46	0.12	0.14	0.16	0.10

Variables selected in Table 4.8 are ranked based on their weights. All three important variables of model (b3) are selected and ranked as top three. Finally, we rank the top ten possible order 2 and 3 interaction effects. For pairwise interactions, the top ten pairs are given in Table 4.9.

Table 4.8: The Rank of Selected Important Variables for Model (b1)

1	5	3	69	114	46	98	24	10	6	189	68	131	15	182	152	89	184	144	137	78	47	49	174	198	195

Table 4.9: Top Ten Pairwise Interactions Ranked by the Sum of PVIM for Model (b1)

Interactions	Sum of ID	Layer 1	Layer 2	Repetition	Sum of PVIM
1	22816	1	5	123	189
2	23290	5	1	126	161
3	737	5	27	12	14
4	1292	1	69	11	10
5	296	1	121	5	9
6	4326	3	69	39	9
7	282	5	194	5	8
8	366	1	83	5	8
9	371	5	190	5	8
10	527	5	28	5	7

In Table 4.9, pairwise interaction (1, 5) shows up twice in different orders. Furthermore, interaction (1, 5) is selected and ranked high according to all three criterion. Considering that the forest size $n_t = 100$, the average of repetition times for both (1, 5) and (5, 1) are larger than 1. That means, when variable 1 is the “king”, both second layer nodes of some trees are selected as variable 5, and vice versa. All these suggest that (1, 5) is a dominant second order interaction effect.

Table 4.10: Top Ten Third Order Interactions Ranked by the Sum of PVIM for Model (b3)

Interactions	Sum of ID	Layer 1	Layer 2	Layer 3	Repetition	Sum of PVIM
1	3301	5	1	3	39	66
2	1291	1	5	3	17	40
3	544	1	5	184	12	24
4	1250	1	5	69	18	23
5	773	1	5	200	14	20
6	548	1	5	24	9	18
7	1608	5	128	1	18	18
8	588	5	1	69	10	16
9	355	1	5	7	6	16
10	369	1	121	5	6	15

Table 4.10 gives top ten third order interaction effects. Among them, we can see that third order interactions (1, 3, 5) show up most frequentyly in different orders. By calculating the average PVIM of them and comparing with second order average, we have Table 4.11. It shows that average PVIM of third order interactions (5, 1, 3) and (1, 5, 3) are much larger than that of interaction (5, 1) and (1, 5). Therefore, (1, 3, 5) should be an important third order interaction.

Table 4.11: Comparison of Average PVIM Between Second and Third Order Interactions

Interactions	Average PVIM	Interactions	Average PVIM
(1, 5)	189/123=1.54	(1, 5, 3)	40/17=2.35
(5, 1)	161/126=1.28	(5, 1, 3)	66/39=1.69

Similarly, we set the tree depth $d = 4$. Table 4.12 shows the top ten order 4 interactions. All of them have repetition times smaller than or equal to 6 times, which is only 6% of the forest size. Therefore, we conclude that there is no order 4 interaction effect in the model.

Table 4.12: Top Ten Order 4 Interactions Ranked by the Sum of PVIM for Model (b1)

Interactions	Sum of ID	Layer 1	Layer 2	Layer 3	Layer 4	Repetition	Sum of PVIM
1	165	1	5	12	15	5	16
2	110	5	1	181	126	6	14
3	203	5	161	1	60	4	10
4	104	1	5	3	10	4	9
5	55	5	1	181	129	4	8
6	73	5	1	28	3	3	7
7	66	1	5	29	100	2	6
8	75	1	5	29	131	2	6
9	20	1	5	6	68	2	6
10	35	1	5	6	152	2	6

4.3 Simulation Studies

The main purpose of simulation studies is to assess the performance of the "Iteratively Kings' Forests" in feature and interaction selection through comparison with feature screening procedure DC-SIS (Li et al., 2012) and the Iterative Random Forest (iRF) (Basu et al., 2018). We focus on models that all important variables affect the response through participating in some interaction effects. The proposed procedure is considered to be good if both important variables and interaction effects are selected and ranked top.

In this section, we set the sample size $n = 200$, parameter scale $s = 2$ and the number of predictors $p = 200, 500$. For the screening model size, we set $d_1 = [0.5 * n/\log(n)]$ and $d_2 = [n/\log(n)]$, where $[a]$ denotes the integer part of a. For each comparison setting, we do 100 times monto carlo simulation. We compare the performance of the proposed procedure and DC-SIS using the following three criteria:

1. $\mathcal{S}$: The minimum model size to include all active predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of 100 replications.
2. P_s: The proportion that an individual active predictor is selected for a given model size d in the 100 replications.

3. P_a: The proportion that all active predictors are selected for a given model size d in the 100 replications.

Besides the above mentioned criteria, we also compare the performance of the proposed procedure and iRF in selecting interaction effects. Two additional criteria are used for this purpose.

1. p_{inter}: The proportion that one active interaction effect is selected in the 100 replications.
2. P_{inter}: The proportion that all active interaction effects are selected in the 100 replications.

When comparing the performance of the proposed procedure and iRF, we study both the continuous and the binary response situations through different settings.

For the parameters of the proposed procedure, we set the forest size $n_t = 100$ and the length of potential interaction effect list $n_{top} = 20$. When constructing a tree, we search $mtry = [p/2]$ variables for each node and $n_{quantile} = 9$ quantile positions for each splitting. For any node, if the number of observations in it is less than $s_{node} = 3$, we stop splitting it and use it as a leaf node. When comparing with iRF in continuous response situation, we set tree depth $d = 4$; when comparing with iRF in binary response cases, we set tree depth $d = 6$.

4.3.1 Comparison with DC-SIS in Feature Selection

We assess the performance of the proposed procedure by comparing with the Distance Correlation feature screening (Li et al., 2012) procedure. Simulation settings are divided into two parts. For the part **(a)**, we compare the performance of these two procedures to find variables participating in pairwise interaction effects. Five settings listed from $(a1)$ to $(a5)$ are studied. In each setting, four variables affect the response through two pairwise interaction effects of different forms.

$$\text{(a1):}\quad y = s * x_1 * x_3 - s * x_5 * (x_7 < 0.2) + e$$

$$\text{(a2):}\quad y = 2 * s * x_1 * \sin(x_3) + 2 * s * x_5 * \cos(x_7 + \pi/2) + e$$

$$\text{(a3):}\quad y = s * \exp(x_1) * x_3/2 - s * \log(5 * |x_5|) * x_7 + e$$

(a4): $y = s * x_1 x_3^2/2 - s * sign(x_5) * x_7^2 + e$

(a5): $y = s * x_1 * x_3 + 1.5 * s * x_5 * \sin(x_7) + e$

For the part **(b)**, we compare the performance of two procedures to find variables participating in one third-order interaction effect. Five cases, $(b1)$ to $(b5)$, are studied. In each of them, three variables affect the response through one third-order interaction effect.

(b1): $y = s * x_1 * (1 + x_3)^2 * \sin(x_5) + e$

(b2): $y = s * x_1 * \log(5 * |1 + x_3|) * \sin(x_5) + e$

(b3): $y = s * x_1 * sign(1 + x_3) * \sin(x_5) + e$

(b4): $y = s * x_1 * x_3 * \sin(x_5) + e$

(b5): $y = s * x_1 * x_3 * x_5 + e$

Tables 4.13-4.15 depict the simulation results of $\mathcal{S}$, P_s and P_a for the proposed algorithm and DC-SIS.

Table 4.13 shows that the proposed procedure outperforms DC-SIS in all cases with respect to minimum model size. When $p = 200$, the proposed procedure ranks all four important variables in settings (a1)-(a4) as the top five in more than 50% repetitions. That is, we find all the important variables of the model directly in more than 50% of times. For the cases (a2), (a5), (b3) and (b4), DC-SIS gives $\mathcal{S}$ larger than 16 in more than 95% of times, which mean DC-SIS doesn't work well in keeping all important variables under these settings. When we increase p to 500, the proposed procedure still outperform DC-SIS in all settings, and works well in all cases except for (b3).

Tables 4.14-4.15 give the proportions that active predictors are selected for given model sizes. The proposed procedure again outpeforms DC-SIS in all cases. Especially, DC-SIS fails to include all active predictors in cases (a2), (a5), (b3) and (b4), while the proposed procedure still works pretty well. Furthermore, as p increases from 200 to 500, the proposed procedure has even larger advantage over DC-SIS with respect to P_a. If we consider $p = 500$ and the model size $d_2 = [n/\log(n)]$ for comparison, the proposed procedure gives P_a larger than 80%

Table 4.13: Quantiles of $\mathcal{S}$ between iKF and DC-SIS

$\mathcal{S}$	iKF					DC-SIS				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
	$n = 200$ and $p = 200$									
$(a1)$	4.00	4.00	4.00	8.00	22.30	9.00	20.00	37.00	62.00	117.00
$(a2)$	4.00	4.00	4.00	22.50	57.75	23.95	44.00	69.50	103.25	152.00
$(a3)$	4.00	4.00	5.00	8.00	23.70	6.00	15.00	29.00	52.00	110.10
$(a4)$	4.00	4.00	4.00	5.00	9.00	4.00	7.00	13.00	26.00	65.10
$(a5)$	4.00	4.00	16.00	40.00	107.00	16.00	37.00	57.00	87.00	128.40
$(b1)$	3.00	3.00	3.00	4.00	23.00	3.95	9.00	24.00	48.75	123.90
$(b2)$	3.00	3.00	5.00	11.00	38.40	4.95	10.00	20.00	35.00	91.10
$(b3)$	4.00	8.00	23.00	50.00	112.00	28.00	65.00	121.50	155.25	186.50
$(b4)$	4.00	9.00	16.00	36.00	79.50	19.80	40.00	63.50	87.25	129.05
$(b5)$	3.00	4.00	6.00	9.00	18.00	5.95	10.00	18.00	28.00	67.10
	$n = 200$ and $p = 500$									
$(a1)$	4.00	4.00	12.00	28.00	92.70	18.00	49.00	89.00	150.25	326.05
$(a2)$	4.00	4.00	18.00	80.00	196.00	49.00	114.75	177.00	263.00	383.10
$(a3)$	4.00	5.00	9.00	17.00	56.20	9.00	32.00	66.00	129.00	281.05
$(a4)$	4.00	4.00	5.00	11.00	25.30	6.00	13.00	25.00	53.00	142.10
$(a5)$	4.00	8.25	48.00	110.00	244.05	40.90	91.25	160.00	223.00	325.55
$(b1)$	3.00	3.00	3.00	6.00	31.05	5.00	21.50	62.00	140.00	258.70
$(b2)$	3.00	3.75	10.50	26.50	109.40	6.00	19.75	38.50	74.25	161.15
$(b3)$	30.85	95.50	193.50	323.25	411.90	53.85	146.50	240.00	391.00	471.35
$(b4)$	7.55	19.75	49.00	87.50	180.30	41.00	70.75	109.00	165.75	305.15
$(b5)$	3.00	5.00	12.00	19.00	42.30	7.00	21.00	36.50	63.25	175.05

in most cases, while DC-SIS gives P_a less than 60% in all situations. Therefore, the proposed procedure performs much better than DC-SIS when the active variables have no marginal effect, but affect the response through interaction effects.

4.3.2 Comparison with Iterative Random Forest in Interaction Detection

In this part, we assess the performance of the proposed procedure in both continuous response settings and binary response settings. In the continuous response situation, we still use the settings **(a)** and **(b)** to compare the proposed procedure and iRF. We use the three Boolean rule settings (OR, AND, and XOR) used by Basu et al. (2018) for comparison in the binary response situation, in which setting details will be given correspondingly.

When calculating the proportion of selecting active interaction effects p_{inter}

Table 4.14: Comparison between iKF and DC-SIS for (**a**) Settings

		iKF					DC-SIS				
		P_s				P_a	P_s				P_a
Model	Size	X_1	X_3	X_5	X_7	all	X_1	X_3	X_5	X_7	all
$n = 200$ and $p = 200$											
$(a1)$	d_1	0.98	0.97	1.00	0.96	0.92	0.68	0.67	1.00	0.46	0.22
	d_2	0.99	1.00	1.00	0.99	0.98	0.87	0.87	1.00	0.67	0.50
$(a2)$	d_1	0.97	0.86	0.93	0.87	0.71	0.70	0.25	0.68	0.26	0.02
	d_2	0.99	0.94	0.99	0.92	0.86	0.87	0.47	0.88	0.49	0.18
$(a3)$	d_1	0.99	1.00	0.94	1.00	0.93	0.78	1.00	0.39	1.00	0.27
	d_2	0.99	1.00	0.99	1.00	0.98	0.89	1.00	0.63	1.00	0.56
$(a4)$	d_1	1.00	0.98	1.00	1.00	0.98	1.00	0.61	1.00	0.99	0.60
	d_2	1.00	0.99	1.00	1.00	0.99	1.00	0.84	1.00	1.00	0.84
$(a5)$	d_1	0.87	0.89	0.87	0.69	0.53	0.50	0.55	0.64	0.27	0.03
	d_2	0.94	0.97	0.95	0.82	0.73	0.78	0.81	0.91	0.43	0.21
$n = 200$ and $p = 500$											
$(a1)$	d_1	0.83	0.83	1.00	0.83	0.62	0.38	0.41	1.00	0.27	0.05
	d_2	0.92	0.95	1.00	0.95	0.83	0.60	0.62	1.00	0.43	0.16
$(a2)$	d_1	0.82	0.64	0.87	0.77	0.51	0.36	0.10	0.34	0.11	0.01
	d_2	0.87	0.75	0.95	0.84	0.64	0.58	0.19	0.53	0.21	0.03
$(a3)$	d_1	0.97	1.00	0.80	1.00	0.77	0.67	1.00	0.17	1.00	0.14
	d_2	0.99	1.00	0.90	1.00	0.89	0.77	1.00	0.31	1.00	0.26
$(a4)$	d_1	1.00	0.89	1.00	1.00	0.89	1.00	0.40	1.00	0.93	0.38
	d_2	1.00	0.97	1.00	1.00	0.97	1.00	0.59	1.00	0.99	0.59
$(a5)$	d_1	0.67	0.73	0.72	0.54	0.31	0.29	0.22	0.52	0.12	0.00
	d_2	0.82	0.80	0.88	0.61	0.43	0.48	0.42	0.78	0.21	0.02

and P_{inter}, we consider one interaction is selected by iRF if it is selected in the interaction list given by the R package "iRF", in which all interactions with non-zero stability scores are included. For the proposed procedure, we consider one interaction is selected if it is selected in at least one of the three shortlists with length $n_{top} = 20$.

The tuning parameters of random intersection tree part were set to the default levels. That is, $M = 100$ random intersection trees of depth 5 were grown with $nchild = 2$. $B = 20$ bootstrap replicates were taken to determine the stability scores of recovered interactions.

Table 4.15: Comparison between iKF and DC-SIS for (**b**) Settings

		iKF				DC-SIS			
		P_s			P_a	P_s			P_a
Model	Size	X_1	X_3	X_5	all	X_1	X_3	X_5	all
$n = 200$ and $p = 200$									
$(b1)$	d_1	1.00	1.00	0.93	0.93	0.92	1.00	0.45	0.44
	d_2	1.00	1.00	0.97	0.97	0.99	1.00	0.66	0.66
$(b2)$	d_1	1.00	0.87	1.00	0.87	0.99	0.76	0.62	0.49
	d_2	1.00	0.95	1.00	0.95	1.00	0.90	0.87	0.79
$(b3)$	d_1	1.00	0.47	1.00	0.47	0.94	0.06	0.37	0.02
	d_2	1.00	0.65	1.00	0.65	1.00	0.16	0.56	0.11
$(b4)$	d_1	0.96	0.97	0.63	0.60	0.62	0.62	0.22	0.05
	d_2	0.98	0.99	0.79	0.77	0.85	0.77	0.38	0.20
$(b5)$	d_1	0.99	0.98	0.99	0.96	0.80	0.76	0.79	0.51
	d_2	1.00	0.99	1.00	0.99	0.94	0.90	0.94	0.82
$n = 200$ and $p = 500$									
$(b1)$	d_1	0.94	1.00	0.92	0.90	0.77	1.00	0.26	0.24
	d_2	0.97	1.00	0.97	0.96	0.91	1.00	0.42	0.39
$(b2)$	d_1	1.00	0.67	1.00	0.67	0.93	0.67	0.35	0.22
	d_2	1.00	0.81	1.00	0.81	0.99	0.86	0.56	0.48
$(b3)$	d_1	1.00	0.02	0.88	0.02	0.73	0.04	0.10	0.00
	d_2	1.00	0.08	0.94	0.08	0.94	0.05	0.30	0.04
$(b4)$	d_1	0.83	0.88	0.29	0.23	0.47	0.43	0.08	0.01
	d_2	0.96	0.94	0.54	0.46	0.64	0.61	0.14	0.04
$(b5)$	d_1	0.91	0.93	0.86	0.72	0.61	0.59	0.54	0.21
	d_2	0.97	1.00	0.97	0.93	0.86	0.77	0.77	0.51

4.3.2.1 Comparisons among Continuous Response Settings

Similar to the comparison with the feature screening procedure DC-SIS before, we compare the proposed procedure with the forest algorithm iRF under settings **(a)** and **(b)**. We compare not only their performance in selecting important variables through $\mathcal{S}$, P_s and P_a, but also their accuracy in detecting interaction effects through p_{inter} and P_{inter}.

Table 4.16 shows that the proposed procedure outperforms iRF in most cases with respect to minimum model size. However, its advantages over iRF is not as large as the advantages over DC-SIS. Especially, when $p = 500$, iRF outperforms the proposed procedure a little bit in the setting (a4). Among all ten settings, (b3) is the most difficult for both algorithms to discover. When we increase p to 500, the proposed procedure still ranks all important variables as the top 20 in more

Table 4.16: Quantiles of $\mathcal{S}$ for Continuous Response Settings

$\mathcal{S}$	iKF					iRF				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
	$n = 200$ and $p = 200$									
$(a1)$	4.00	4.00	4.00	8.00	22.30	4.00	5.00	6.00	8.25	27.25
$(a2)$	4.00	4.00	4.00	22.50	57.75	8.00	13.00	22.50	43.25	83.60
$(a3)$	4.00	4.00	5.00	8.00	23.70	4.00	5.00	7.00	10.00	21.15
$(a4)$	4.00	4.00	4.00	5.00	9.00	4.00	4.00	4.00	5.00	11.05
$(a5)$	4.00	4.00	16.00	40.00	107.00	8.90	14.00	25.00	53.50	141.15
$(b1)$	3.00	3.00	3.00	4.00	23.00	10.95	18.00	34.50	58.25	120.10
$(b2)$	3.00	3.00	5.00	11.00	38.40	5.00	7.75	13.50	22.25	52.30
$(b3)$	4.00	8.00	23.00	50.00	112.00	13.00	60.75	122.50	149.25	157.00
$(b4)$	4.00	9.00	16.00	36.00	79.50	6.00	20.00	39.00	61.00	126.45
$(b5)$	3.00	4.00	6.00	9.00	18.00	4.95	8.00	12.50	21.25	53.10
	$n = 200$ and $p = 500$									
$(a1)$	4.00	4.00	12.00	28.00	92.70	4.95	8.00	12.00	26.75	120.15
$(a2)$	4.00	4.00	18.00	80.00	196.00	15.85	48.00	84.50	138.25	266.05
$(a3)$	4.00	5.00	9.00	17.00	56.20	4.00	7.00	12.00	20.50	67.85
$(a4)$	4.00	4.00	5.00	11.00	25.30	4.00	4.00	5.00	7.00	16.05
$(a5)$	4.00	8.25	48.00	110.00	244.05	9.95	29.75	61.00	104.25	260.10
$(b1)$	3.00	3.00	3.00	6.00	31.05	14.90	42.00	75.50	139.75	225.00
$(b2)$	3.00	3.75	10.50	26.50	109.40	10.00	20.00	38.50	73.25	234.50
$(b3)$	30.85	95.50	193.50	323.25	411.90	37.00	141.25	248.00	257.00	266.00
$(b4)$	7.55	19.75	49.00	87.50	180.30	14.95	33.75	77.00	143.00	255.10
$(b5)$	3.00	5.00	12.00	19.00	42.30	6.00	14.00	26.00	52.00	122.90

than 50% repetitions for the cases (a1)-(a4), (b1), (b2) and (b5). At the same time, iRF could only achieve it in the cases (a1), (a3) and (a4).

Tables 4.17-4.18 give the proportions that active predictors are selected for given model sizes d_1 and d_2. The proposed procedure outpeforms iRF in all **(b)** settings. For the settings of **(a)**, the proposed procedure wins the cases (a2) and (a5), while iRF performs slightly better in (a1) and (a4). In general, iRF performs much better than DC-SIS in finding active variables that participate in interaction effects, but still not as good as the proposed procedure.

Tables 4.19 gives the successful selecting rate of interactions for both procedures. The proposed procedure outperforms iRF for all cases. Especially for **(b)** settings, iRF totally fails to find the third order interaction, while the proposed procedure has more than 50% probability to discover them in cases (b1)-(b3) when $p = 200$. For **(a)** settings, the proposed procedure also has a large advantage over iRF. Since iRF gives a list of all candidate interaction effects with non-zero stabil-

Table 4.17: Comparison between iKF and iRF for **a** Settings

		iKF					iRF				
		P_s				P_a	P_s				P_a
Model	Size	X_1	X_3	X_5	X_7	all	X_1	X_3	X_5	X_7	all
$n = 200$ and $p = 200$											
$(a1)$	d_1	0.98	0.97	1.00	0.96	0.92	0.98	0.98	1.00	0.94	0.91
	d_2	0.99	1.00	1.00	0.99	0.98	0.99	1.00	1.00	0.98	0.97
$(a2)$	d_1	0.97	0.86	0.93	0.87	0.71	0.96	0.59	0.99	0.71	0.40
	d_2	0.99	0.94	0.99	0.92	0.86	1.00	0.81	1.00	0.83	0.68
$(a3)$	d_1	0.99	1.00	0.94	1.00	0.93	0.99	1.00	0.94	1.00	0.93
	d_2	0.99	1.00	0.99	1.00	0.98	1.00	1.00	0.99	1.00	0.99
$(a4)$	d_1	1.00	0.98	1.00	1.00	0.98	1.00	0.97	1.00	1.00	0.97
	d_2	1.00	0.99	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.99
$(a5)$	d_1	0.87	0.89	0.87	0.69	0.53	0.87	0.83	0.95	0.50	0.37
	d_2	0.94	0.97	0.95	0.82	0.73	0.96	0.94	0.99	0.69	0.60
$n = 200$ and $p = 500$											
$(a1)$	d_1	0.83	0.83	1.00	0.83	0.62	0.91	0.90	1.00	0.87	0.70
	d_2	0.92	0.95	1.00	0.95	0.83	0.96	0.97	1.00	0.90	0.84
$(a2)$	d_1	0.82	0.64	0.87	0.77	0.51	0.89	0.30	0.80	0.25	0.07
	d_2	0.87	0.75	0.95	0.84	0.64	0.97	0.49	0.94	0.41	0.17
$(a3)$	d_1	0.97	1.00	0.80	1.00	0.77	1.00	1.00	0.72	1.00	0.72
	d_2	0.99	1.00	0.90	1.00	0.89	1.00	1.00	0.88	1.00	0.88
$(a4)$	d_1	1.00	0.89	1.00	1.00	0.89	1.00	0.96	1.00	1.00	0.96
	d_2	1.00	0.97	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
$(a5)$	d_1	0.67	0.73	0.72	0.54	0.31	0.64	0.74	0.83	0.25	0.11
	d_2	0.82	0.80	0.88	0.61	0.43	0.80	0.85	0.94	0.49	0.33

ity score, the list is much longer than that given by the proposed procedure in all settings. Therefore, the proposed procedure is not only more accurate, but also more efficient than iRF in discovering important interaction effects.

To conclude, the proposed procedure outperforms iRF in discovering both the interaction effects and the marginal unimportant variables that participates in these interaction effects.

4.3.2.2 Comparisons among Binary Response Settings

In this part, we follow Basu et al. (2018) to conduct the simulation using the Boolean-type rules settings, and compare the performance of the proposed procedure with iRF. Details of basu's settings are given as follow.

Instead of normal distribution, we sampled features $\mathbf{x} = (x_1, \ldots, x_p)^T$ from

Table 4.18: Comparison between iKF and iRF for **b** Settings

		iKF				iRF			
		P_s			P_a	P_s			P_a
Model	Size	X_1	X_3	X_5	all	X_1	X_3	X_5	all
$n = 200$ and $p = 200$									
$(b1)$	d_1	1.00	1.00	0.93	0.93	0.80	0.99	0.29	0.26
	d_2	1.00	1.00	0.97	0.97	0.95	1.00	0.58	0.57
$(b2)$	d_1	1.00	0.87	1.00	0.87	1.00	0.73	0.88	0.64
	d_2	1.00	0.95	1.00	0.95	1.00	0.93	0.97	0.90
$(b3)$	d_1	1.00	0.47	1.00	0.47	1.00	0.07	0.98	0.07
	d_2	1.00	0.65	1.00	0.65	1.00	0.12	1.00	0.12
$(b4)$	d_1	0.96	0.97	0.63	0.60	0.88	0.90	0.29	0.22
	d_2	0.98	0.99	0.79	0.77	0.94	0.97	0.51	0.47
$(b5)$	d_1	0.99	0.98	0.99	0.96	0.85	0.89	0.90	0.68
	d_2	1.00	0.99	1.00	0.99	0.97	0.96	0.96	0.89
$n = 200$ and $p = 500$									
$(b1)$	d_1	0.94	1.00	0.92	0.90	0.62	0.98	0.09	0.07
	d_2	0.97	1.00	0.97	0.96	0.79	1.00	0.28	0.24
$(b2)$	d_1	1.00	0.67	1.00	0.67	1.00	0.41	0.48	0.24
	d_2	1.00	0.81	1.00	0.81	1.00	0.60	0.73	0.48
$(b3)$	d_1	1.00	0.02	0.88	0.02	1.00	0.04	0.74	0.02
	d_2	1.00	0.08	0.94	0.08	1.00	0.06	0.89	0.06
$(b4)$	d_1	0.83	0.88	0.29	0.23	0.75	0.83	0.14	0.07
	d_2	0.96	0.94	0.54	0.46	0.90	0.95	0.32	0.28
$(b5)$	d_1	0.91	0.93	0.86	0.72	0.74	0.72	0.71	0.35
	d_2	0.97	1.00	0.97	0.93	0.87	0.86	0.86	0.62

independent, standard Cauchy distributions to reflect heavy-tailed data. The binary responses are generated from three Boolean-type rule settings (OR, AND, and XOR) as follows:

$$y^{(OR)} = I\{x_1 > t_{OR} | x_3 > t_{OR} | x_5 > t_{OR} | x_7 > t_{OR}\} \tag{4.5}$$

$$y^{(AND)} = \prod_{i=1}^{4} I\{x_{2i-1} > t_{AND}\} \tag{4.6}$$

$$y^{(XOR)} = I\{\sum_{i=1}^{4} (x_{2i-1} > t_{XOR}) \equiv 1 \pmod 2)\} \tag{4.7}$$

To introduce noise, we swap the labels for 10% of the observations selected at random. Therefore, the rules in equations (4.5)-(4.7) give rise to non-additive

Table 4.19: Successful Selecting Rate of Interactions for Continuous Response Settings

iKF						iRF					
Model	p_{inter}		P_{inter}	Model	P_{inter}	Model	p_{inter}		P_{inter}	Model	P_{inter}
$n = 200$ and $p = 200$											
$(a1)$	0.77	0.92	0.71	$(b1)$	0.87	$(a1)$	0.34	0.94	0.32	$(b1)$	0.08
$(a2)$	0.76	0.75	0.58	$(b2)$	0.67	$(a2)$	0.46	0.48	0.16	$(b2)$	0.01
$(a3)$	0.98	0.76	0.74	$(b3)$	0.53	$(a3)$	0.65	0.90	0.56	$(b3)$	0
$(a4)$	0.93	0.99	0.92	$(b4)$	0.15	$(a4)$	0.82	0.99	0.81	$(b4)$	0
$(a5)$	0.64	0.57	0.34	$(b5)$	0.21	$(a5)$	0.53	0.33	0.13	$(b5)$	0.09
$n = 200$ and $p = 500$											
$(a1)$	0.73	0.82	0.60	$(b1)$	0.68	$(a1)$	0.12	0.84	0.09	$(b1)$	0
$(a2)$	0.53	0.73	0.41	$(b2)$	0.48	$(a2)$	0.19	0.15	0.02	$(b2)$	0
$(a3)$	0.92	0.54	0.51	$(b3)$	0.16	$(a3)$	0.62	0.76	0.45	$(b3)$	0
$(a4)$	0.80	0.97	0.77	$(b4)$	0.03	$(a4)$	0.65	0.96	0.61	$(b4)$	0
$(a5)$	0.57	0.45	0.28	$(b5)$	0.05	$(a5)$	0.28	0.14	0.04	$(b5)$	0

main effects that can be represented as an order-4 interaction between the active features x_1, x_3, x_5 and x_7. For the AND and OR models, we set $t_{OR} = 3.2$ and $t_{AND} = -1$ to ensure reasonable class balance ($\sim$ 1/3 class 1 observations). We set $t_{XOR} = 1$ both for class balance ($\sim$ 1/2 class 1 observations) and to ensure that some active features were marginally important relative to inactive features.

Similarly, we compare their performance in selecting important variables and interaction effects through $\mathcal{S}$, P_s, P_a and P_{inter} in binary response situation.

Table 4.20: Quantiles of $\mathcal{S}$ for Binary Response Settings

$\mathcal{S}$	iKF					iRF				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$n = 200$ and $p = 200$										
Or4	4.00	4.00	4.00	4.00	6.00	4.00	4.00	4.00	4.00	4.00
And4	4.00	4.00	9.00	21.00	76.75	4.00	4.00	4.00	4.00	4.15
Xor4	27.00	98.00	140.00	171.00	197.55	51.35	112.50	140.00	152.75	160.00
$n = 200$ and $p = 500$										
Or4	4.00	4.00	4.00	4.00	5.00	4.00	4.00	4.00	4.00	4.00
And4	4.00	4.00	16.00	63.00	301.00	4.00	4.00	4.00	5.00	8.00
Xor4	134.00	278.00	352.00	440.00	482.40	150.60	214.00	228.00	235.00	239.60

Different from the previous simulation results, Table 4.20 shows that iRF outperforms the proposed procedure in general with respect to minimum model size. For the setting $Or4$, both procedures works very well for both $p = 200$ and $p = 500$. iRF works perfectly in the $And4$ setting, while the proposed procedure has relatively worse performance. Both procedures don't work very well in the $Xor4$ setting.

Table 4.21: Comparison between iKF and iRF for Binary Response Settings

		iKF						iRF					
Model	Size	X_1	X_3	X_5	X_7	P_a	P_{inter}	X_1	X_3	X_5	X_7	P_a	P_{inter}
$n = 200$ and $p = 200$													
Or4	d_1	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99
	d_2	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	1.00	
And4	d_1	0.87	0.85	0.84	0.91	0.70	0.67	1.00	1.00	1.00	1.00	1.00	0.22
	d_2	0.93	0.93	0.94	0.95	0.87		1.00	1.00	1.00	1.00	1.00	
Xor4	d_1	0.22	0.22	0.26	0.21	0.02	0.05	0.25	0.24	0.23	0.22	0.01	0.00
	d_2	0.34	0.33	0.41	0.38	0.07		0.35	0.36	0.34	0.35	0.03	
$n = 200$ and $p = 500$													
Or4	d_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89
	d_2	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	1.00	
And4	d_1	0.69	0.76	0.72	0.69	0.52	0.55	0.97	1.00	1.00	1.00	0.97	0.07
	d_2	0.76	0.83	0.76	0.72	0.55		1.00	1.00	1.00	1.00	1.00	
Xor4	d_1	0.18	0.10	0.02	0.08	0.00	0.02	0.16	0.08	0.06	0.10	0.00	0.00
	d_2	0.22	0.12	0.08	0.14	0.00		0.20	0.14	0.14	0.20	0.00	

Table 4.21 gives the results of P_s, P_a and P_{inter} for both procedures. With respect to P_s and P_a, iRF shows much better performance in the $And4$ setting. Both procedures work well under the $Or4$ setting, and don't work well for the $Xor4$ case. However, when we consider their performance based on the successful selecting rate of the interaction effect, the proposed procedure still outperforms iRF more or less in all three settings.

4.4 Real Data Analysis

In this section, we apply the proposed iKF procedure to the early Drosophila embryo data (https://zenodo.org/record/885529#.XMOIVOtKjVo) investigated by Basu et al. (2018), and compare its performance with iRF in identifying interaction effects. In this data, 7809 genomic sequences are given for evaluating their enhancer activity. For the response enhancer status, sequences that drive patterned expression in blastoderm (stage 5) embryos were labeled as positive. In our results, iKF not only identifies most of pairwise TF interactions recovered in Basu et al. (2018), but also verifies several other important interactions studied in related biological literatures(Harrison et al., 2011; Nien et al., 2011; Morán and Jiménez, 2006). Furthermore, through evaluation based on interactions' PVIM and times of repetition, our algorithm identifies the interactions' structure, and categorizes them as either dominant TFs or assistant TFs that are nested in dominant TFs. A best example to demonstrate this is that iKF rediscovers the basic and crucial role of early regulatory factor Zelda (Zld) in timing zygotic gene activation and promoting robust expression, and identify about ten interactions between Zld and other TFs nested in Zld. Also, we identify twi as a dominant variable, whose importance is also discussed in Stathopoulos et al. (2002), Markstein et al. (2004), Zeitlinger et al. (2007) and many other literatures.

4.4.1 Background of the Early Drosophila Embryogenesis

Precisely regulated spatio-temporal gene expression is crucial for the development in multi-cellular organisms. In this process, enhancers play a critical role through coordinating combinatorial transcription factor(TF) binding. Their integrated activities lead to patterned gene expression during embryogenesis(Levine, 2010). One of the best-studied developmental embryogenesis cases is the Drosophila embryo in which TF hierarchies act to pattern and subdivide the embryo along the anteroposterior (AP) and dorsoventral (DV) body axes. In the early Drosophila embryo, about 40 TFs drive patterning(Rivera-Pomar and Jackle, 1996), and therefore provide a valuable test case for evaluating the performance of iKF in modelling the relationship between TF binding and enhancer status.

The zinc-finger protein Zelda (Zld) plays a key role as an early regulatory factor in timing zygotic gene activation and promoting robust expression. When lacking maternal expression of Zld in early embryos, expression profiling studies revealed that about 70% of the genes normally activated between 12 hrs of development were strongly down-regulated and never recover, including many genes related to cellularization, sex determination, and dorsal patterning(Liang et al., 2008). That is, these genes are variables nested in Zld. However, in the absence of Zld, some genes involved in ventral patterning, for example twi, was just temporally delayed, but later recovered by nuclear cycle (nc) 14. To be more specific, 44% of these genes were down-regulated in lower level zld and 19% of them were bound by Zld, indicating that Zld activates many of the newly transcribed genes, both directly and indirectly.

Zld regulates the expression through binding target genes of the key patterning factors destined to be expressed in the blastoderm embryo. Nien et al. (2011) points out that Zld binds to 72% of the Bcd targets, 70% of the Cad targets, and 80% of the tll targets. About 50% overlap was observed between Zld targets and gap gene (Hb, Gt, Kr, Kni) targets. Also, Zld regulates the expression through orchestrating the timing within the segmentation gene network. For example, transcripts of gap genes gt and tll were detected at nc 10 in wild-type embryos, while transcripts of gap genes kni, Kr and hb were not observed until nc 1112. In lack of zld, embryos initial transcription of all five gap genes was delayed by 12 nc. In addition, their patterns were significantly disrupted, which can be explained in part by miscued gap gene interactions.

The TF twi is also widely studied in the literatures. At the core of the dorsoventral patterning network are Dorsal and the products encoded by two of its earliest target genes, Twist(twi) and Snail. Twist functions as a basic helix-loophelix (bHLH) activator, and is very essential for specifying the ventral neurogenic ectoderm. At least half of the tissue-specific enhancers that are regulated by different levels of Dorsal also contain binding sites for twi (Stathopoulos et al., 2002; Markstein et al., 2004).

4.4.2 Analysis Results and Comparison

The proposed iKF procedure is conducted to analyze the early Drosophila embryo data for selecting interactions. In this analysis, we set the tree depth $d = 4$, forest size $n_t = 200$ and the length of output interaction lists $n_{top} = 50$. When constructing a tree, we search $mtry = [p/3]$ variables for each node, $n_{quantile} = 7$ quantile positions for splitting each variable and stop splitting a node if the number of observations in it is less than $s_{node} = 3$. In the corresponding analysis results of iRF, Basu et al. (2018) gave the top 20 pairwise TF interactions, among which 16 interactions are already verified in related biological literatures. In our analysis, we successfully identify 14 pairwise interactions out of the 16 verified interactions, and 2 pairwise interactions out of the 4 unverified interactions. All overlapped pairwise interactions are shown in the Table 4.22. This significant overlap (80%) illustrates the powerful capability of iKF in identifying influential interaction effects.

Table 4.22: Pairwise TF Interactions Recovered by Both iRF and iKF

Interaction (S)	References
(Gt, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Twi, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Gt, Kr)	Kraut and Levine (1991); Struhl et al. (1992); Capovilla et al. (1992); Schulz and Tautz (1994)
(Gt, Twi)	Li et al. (2008)
(Kr, Twi)	Li et al. (2008)
(Kr, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Bcd, Gt)	Kraut and Levine (1991); Eldon and Pirrotta (1991)
(Bcd, Twi)	Li et al. (2008)
(Hb, Twi)	Zeitlinger et al. (2007)
(Med, Twi)	Nguyen and Xu (1998)
(Med, Zld)	Harrison et al. (2011)
(Hb, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Bcd, Kr)	Hoch et al. (1991); Hoch et al. (1990)
(Bcd, Zld)	Harrison et al. (2011); Nien et al. (2011)
(D, twi)	-
(Gt, Med)	-

Furthermore, iKF also identifies some other interaction effects that are not recovered in iRF. Four of them that are verified by the biological literatures are given in the Table 4.23. Among them, Harrison et al. (2011) and Nien et al. (2011) showed that Kni, Ftz and Cad are TFs regulated by Zld, and therefore form three pairwise interactions together with Zld. The forth selected interaction, (Zld, Kni, Tll), is a third-order interaction. As Morán and Jiménez (2006) points out, tll is a strong repressor of gap genes and becomes less expressed in the lack of Zld, hence the ectopic expression of kni can likewise be explained by the delay in tll expression in Zld^{-}.

After identifying pairwise interactions, iKF can further identify the interactions' structure, and categorize the interactions' participating variables as either dominant TFs or assistant TFs that are nested in dominant TFs. As shown in the Table 4.24, pairwise interactions between five gap gene expression TFs and Zld are all selected in both direction under the similar pattern. That is, if Zld is

Table 4.23: TF Interactions Recovered only by iKF

Interaction (S)	References
(Kni, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Ftz, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Cad, Zld)	Harrison et al. (2011); Nien et al. (2011)
(Zld, Kni, Tll)	Nien et al. (2011); Morán and Jiménez (2006)

Table 4.24: Interaction and Nested Effects between Zld and Other TFs

Layer 1	Layer 2	Repetition	Sum of PVIM	Average PVIM
Gap Gene Expression TFs Nested in Zld				
Zld	Kni	69	3.166	0.0459
Kni	Zld	15	3.011e-15	2e-16
Zld	Bcd	6	0.2142	0.0357
Bcd	Zld	48	5.024e-15	1.047e-16
Zld	Gt	5	0.1022	0.0204
Gt	Zld	27	3.011e-15	1.115e-16
Zld	Kr	11	0.1165	0.0106
Kr	Zld	29	2.970e-15	1.024e-16
Zld	Tll	4	0.0522	0.0131
Tll	Zld	29	2.359e-15	8.134e-17
Anteroposterior Patterning TFs Nested in Zld				
Zld	Ftz	29	0.9998	0.0345
Ftz	Zld	27	3.081e-15	1.141e-16
Zld	Cad	5	0.1030	0.0206
Cad	Zld	14	1.346e-15	9.614e-17
Interaction between Zld and Twi (Expressed in Ventral-most Region)				
Zld	Twi	22	0.7098	0.0323
Twi	Zld	33	0.2961	0.0090

selected as the “King” variable and splitted in the root node, the average PVIM of all five interactions are larger than 0.01. Considering the response is binary and the corresponding evaluation criterion is misclassification rate, 0.01 is a very large importance measure. At the same time, if any of the gap gene expression TFs, say Bcd, is selected as the “King”, its interaction with Zld will still stand out because of the high repetition. However, the corresponding average PVIM is amost 0. That pattern means, if Zld is not splitted before the gap gene expression TFs, their interactions are rarely modelled in the tree. Therefore, we can conclude that Zld is a dominant variable interactions, while all five gap gene expression TFs are assistant variables nested in Zld. Table 4.24 also shows that the anteroposterior patterning TFs, Cad and Ftz, also have the similar interaction pattern with Zld. The discovery of this interaction pattern surrounding Zld is already verified in the previous section, in which we briefly review how Zld regulates the expression through binding target genes of the key patterning factors destined to be expressed in the blastoderm embryo.

Also, from the Table 4.24 we discover that the TF Twi has a different interaction pattern with Zld. The average PVIM in both dirrection are significantly larger than 0. Therefore, they are not nested in each other. That discovery is also partly discussed in the previous part that twi was just temporally delayed, but later fully recovered by nuclear cycle (nc) 14 in the absence of Zld.

Table 4.25: Kings, Kings' Average PVIM and Their Roles in the Mechanism

King	Depth=1	Depth=2	Depth=3	Depth=4	Role
Kni	-4.774e-17	-3.525e-17	9.714e-19	2.220e-18	Assistant Variable
Twi	-2.318e-17	2.170e-02	5.783e-03	5.891e-03	Dominant Variable
Zld	-4.163e-19	2.959e-02	1.176e-02	1.554e-02	Dominant Variable
Bcd	2.359e-18	3.747e-18	-1.776e-17	1.207e-17	Assistant Variable
Cad	1.388e-18	1.110e-17	-6.106e-18	-2.082e-18	Assistant Variable
Gt	-4.163e-18	-4.580e-18	7.910e-18	4.996e-18	Assistant Variable
Kr	1.568e-17	-4.996e-18	-1.429e-17	3.303e-17	Assistant Variable
Tll	-5.829e-18	2.109e-17	-1.388e-18	-1.284e-17	Assistant Variable
Ftz	-4.857e-18	-3.747e-18	1.207e-17	-2.193e-17	Assistant Variable

Table 4.25 lists all the 9 TFs that are selected as "King" and their average PVIM during the "Iteratively Kings' Forest" proceduce. When depth is 1, all of them have average PVIM around 0, which means they all have no marginal effect on the response. When the depth increases to 2, the average PVIMs of both Zld and Twi increase to more than 0.02, which means they are dominant variables participating in some pairwise interactions. At the same time, the other 7 TFs still have average PVIM around 0. That means they are not only nested in Zld, but also play no role as dominant variables in other potential interactions. When the depth increases to 3 or 4, the average PVIMs of Zld and Twi don't increase any more, which means there are no very important interactions with order larger than

or equal to 3. As pointed out in the supporting information appendix of Basu et al. (2018), the high-order interactions related to the Drosophila embryo embryogenesis have only been studied in a small number of select cases, most notably eve stripe 2(Levine, 2013). These limited cases are not sufficient to conduct a comprehensive analysis of the high-order interactions.

Chapter 5

Multidimensional Economic Dispersion Index and Application

The Gini index depicts the degree of social inequality with respect to income or wealth, which is definitely an important aspect of social resources. However, social inequality is not equivalent to the income inequality. It is about inequalities of many aspects such as asset, welfare, public service and education expenditure. When considering a more comprehensive inequality measurement, the Gini index is no longer applicable because its definition is constrained to a single dimension. To evaluate multi-dimensional data, we should first use some algorithms(for example principal component analysis) to compress them into one dimension, then the Gini index can be calculated. When there is more than one dimension, the main difficulty of a direct generalization for Gini Index is the non-existence of the inverse function if we treat the cumulative proportion of resources as a function of the cumulative proportion of people. In this chapter, we solve this problem by treating the cumulative proportion of people as a function of the cumulative proportion of resources, and propose multi-dimensional economic dispersion index (MEDI), a natural extension of Gini Index for multi-dimensional resources. The Gini index is equivalent to the MDEI for one-dimensional resource. Since MEDI could comprehensively evaluate the social inequality level with respect to diverse economic aspects as a whole, it could be a more suitable reference index when studying some complex economic issues. Furthermore, based on this multi-dimensional definition, some other measures, for example Boferroni, may also be generalized to multi-dimensional cases.

5.1 Multidimensional Economic Dispersion Index

5.1.1 The Re-formulation of Gini Index

The initial Gini index is defined as shown in Fig 2.1. The horizonal and vertical axes are the cumulative proportions of population and corresponding income, respectively. Denote the income as random variable X. Let $F(t)$ be the cumulative distribution function (cdf) for X, and $L(t)$ be the cumulative income proportion, which means the proportion in the total income concerning the people whose incomes are less than or equal to t,

$$L(t) = \frac{1}{\mu}\int_0^t x dF(x) = \frac{1}{\mu}\int_0^t x f(x) dx, \tag{5.1}$$

where $\mu = \int_0^\infty t f(t) dt$. Both $F(t)$ and $L(t)$ take values in the interval $[0, 1]$. The Lorenz curve is formed by these points $(F(t), L(t)), t \in [0, \infty]$ in the unit square $[0, 1] \otimes [0, 1]$ and can be represented as an implicit function $\mathcal{L}(p)$,

$$y = \mathcal{L}(p) : \begin{cases} p & = F(t) \\ y & = L(t) \end{cases}, \tag{5.2}$$

where p and y are the cumulative proportions for people and income, respectively.

If the random variable is discrete with the probability mass function $f(t_i), i = 1, 2, ..., m$ and finite mean $\mu = \Sigma t_i f(t_i) < \infty$, the index can be computed as (5.3), as shown in (Gini, 1912, 2005, 1921),

$$\mathbf{Gini} = \frac{1}{2\mu}\Sigma_{i=1}^m \Sigma_{j=1}^m f(t_i) f(t_j) |t_i - t_j|. \tag{5.3}$$

When the population is represented by a continuous probability density function (pdf) $f(t)$ with cdf $F(t)$ and finite mean $\mu < \infty$, the Gini index can be computed by (5.4),

$$\mathbf{Gini} = \int_0^1 \mathcal{L}(p)dp = \frac{2}{\mu}\int_{-\infty}^{\infty} t[F(t) - \frac{1}{2}]f(t)dt. \tag{5.4}$$

The Gini index can be formulated as the integral shown in (5.4). This definition is suitable for just one dimension because we can obtain the implicit function $\mathcal{L}(p)$ when dimension is one. Now consider the multidimensional case. We first consider the two-dimensional case, let t_1, t_2 and $F(t_1, t_2)$ denote the income, expenditure and their joint cdf, respectively; it is difficult to obtain the implicity function because there are two independent variables. Hence we cannot expand the index definition to 2 or greater dimensions with the framework of the Gini index. Therefore, an alternative formulation must be taken to achieve the generation. In order to introduce the idea of this new framework, we first re-formulated the Gini index. In the Figure 5.1, we exchange the vertical and horizontal axes of Figure 2.1 and let $\mathcal{F}(y)$ denote the implicity population proportion function (IPPF) determined by the point $(L(t), F(t))$,

$$\pi = \mathcal{F}(y) : \begin{cases} \pi = & F(t) \\ y = & L(t) \end{cases} . \tag{5.5}$$

$\mathcal{F}(y)$ describes the cumulative population proportion p associated with the cumulative income percentage y, i.e., the proportion of the people whose cumulative income amounts the y percent in the total income. $\mathcal{F}(y) = F(t)$ because both of them represent the same cumulative probability with different domains on $[0, 1]$ and $\mathbb{R}$. Actually, since $F(t)$ and $L(t)$ are both strictly ascending functions, t one-to-one corresponds to y in (5.1).

Note that S_B in Figure 5.1 is now located in the upper-left part, so the Gini index can also be derived as,

$$\mathbf{Gini} = 1 - 2S_B = 1 - 2(1 - \int_0^1 \mathcal{F}(y)dy) = 2\int_0^1 \mathcal{F}(y)dy - 1. \tag{5.6}$$

The equation (5.6) is equivalent to (5.4), which has been interpreted in the Figures 2.1 and 5.1. However, this equation provides an alternative formulation of Gini index.

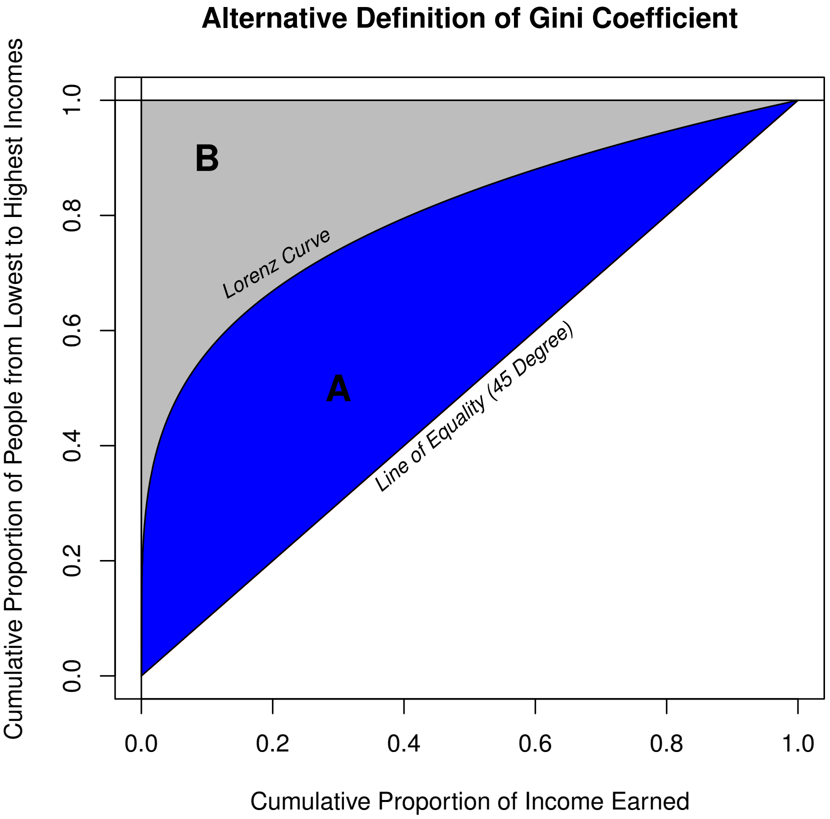

Figure 5.1. An alternative formulation of Lorenz Curve and Gini index

5.1.2 The Definition of Two Dimensional MEDI

In this section, we generalize the definition of the Gini index into MEDI with the same framework. For simplicity, we first consider the two-dimensional case. Let $F(t_1, t_2)$ denote the joint cdf of the random variables income and expenditure. The joint IPPF $\mathcal{F}(y_1, y_2)$ associated with the cumulative income percentage $y_1 = L_1(t_1)$ and expenditure percentage $y_2 = L_2(t_2)$ can be derived. The joint population proportion means the population proportion simultaneously satisfies two conditions; the income and expenditure equal to the percentages of y_1 and y_2 in the total income and expenditure, respectively. The joint IPPF can be viewed as,

$$\pi = \mathcal{F}(y_1, y_2) : \begin{cases} \pi = F(t_1, t_2), \\ y_1 = L_1(t_1) = \frac{1}{\mu_1} \int_{-\infty}^{t_1} \int_{-\infty}^{\infty} x_1 dF(x_1, x_2), \\ y_2 = L_2(t_2) = \frac{1}{\mu_2} \int_{-\infty}^{t_2} \int_{-\infty}^{\infty} x_2 dF(x_1, x_2), \end{cases} \tag{5.7}$$

where μ_1 and μ_2 are the means of the t_1 and t_2, respectively. In this framework, the percentage L_1 is proportional to the function on t_1, so does L_2 on t_2. The function $\mathcal{F}(y_1, y_2)$ shares the same value with $F(t_1, t_2)$ as t_1 and t_2 correspond to the percentages y_1 and y_2, respectively.

The two dimensional $\mathbf{MEDI}^{(2)}$ can be defined using the following equation (5.8),

$$\mathbf{MEDI}^{(2)} = \int_0^1 \int_0^1 \mathcal{F}(y_1, y_2) dy_1 dy_2, \tag{5.8}$$

With this definition, we cannot obtain the close formulation as shown in (5.4) directly related to the (t_1, t_2) because the implicity function $\mathcal{F}(y_1, y_2)$ as well as the functions $L(t_1)$ and $G(t_2)$ may be complex with respect to different joint cdfs. We can only transform the joint cdf into the joint IPPF with the percentages y_1 and y_2, then use the original multidimensional Lorenz Curve to complete the generation of MEDI to multi dimension.

From the geometric interpretation of the Gini index, MEDI^2 can be viewed as the volume which is under the curve of $\mathcal{F}(y_1, y_2)$ in the unit cube. MEDI is obviously the extension of the Gini index. When it is constrained to one dimension, MEDI, according to (5.6), is equivalent to the Gini index with the following transformation,

$$\mathbf{Gini} = 2 * \mathbf{MEDI} - 1. \tag{5.9}$$

In this paper, such transformations as (5.9) will not be introduced any more because both indices can equivalently measure the degrees of the economic inequalities in income, consumption, wealth, etc. Only the definition (5.8) directly formulated with integration will be used for the simplicity of notation and derivations.

5.1.3 Multidimensional MEDI

We can also introduce the definition of MEDI to multivariate cases. Considering a multiple economic vector $\mathbf{X} = (X_1, X_2, ..., X_p)^T$, if the joint cdf of the vector is given by $F(t_1, t_2, ..., t_p)$ and the marginal cdf for the jth component is $F_j(t_j)$ accordingly, we can obtain the joint IPPF by (5.10)

$$\pi = \mathcal{F}(y_1, y_2, ..., y_p) : \begin{cases} \pi = F(t_1, t_2, ..., t_p), \\ y_j = L_j(t_j) = \frac{1}{\mu_j} \int_{-\infty}^{t_j} x_j dF_j(x_j), j = 1, 2, ..., p, \end{cases} \tag{5.10}$$

where μ_j represents the mean of the jth component and y_j represents the cumulative percentage at the point t_j in the sum for jth component. The multivariate definition of the MEDI can be obtained by (5.11),

$$\mathbf{MEDI}^{(p)} = \int_{[0,1]^{\bigotimes p}} \mathcal{F}(y_1, y_2, ..., y_p) dy_1 dy_2 ... dy_p, \tag{5.11}$$

where $[0,1]^{\bigotimes p}$ means the Cartesian Product of p intervals $[0, 1]$ in $\mathbb{R}^p$.

Instead of solely focusing on income or wealth of the Gini index, MEDI can capture more information when it is applied to measured the social inequality. It integrates diverse resources and considers the relations between the resources. Therefore, MEDI is more suitable for investigating complex economic issues, especially those containing numerous variables. Now we will investigate some properties of MEDI with some conditions.

Remark 1: If the social resource vector $\mathbf{x}$ can be divided into m sub-vectors $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)^T$, each $\mathbf{x}_l$ has k_l components $l = 1, 2, ..., m$, $\sum_{l=1}^m k_l = p$, and these m sub-vectors are mutually independent, then the MEDI for the social resource vector can be composed of the product of the MEDIs for all the sub-vectors, i.e.,

$$\mathbf{MEDI}^{(p)} = \Pi_{l=1}^m \mathbf{MEDI}_l^{(k_l)}. \tag{5.12}$$

Especially, if the p components of the vector are mutually independent, the equation (5.12) will be rewritten as follows,

$$\mathbf{MEDI}^{(p)} = \Pi_{j=1}^p \mathbf{MEDI}_j. \tag{5.13}$$

This property is easy to interpret and understand. Since the cdf for the vector can be rewritten as the product of the cdfs for the sub-vectors, the marginal IPPF for each sub-vector can be defined as (5.14),

$$\pi_l = \mathcal{F}_l(\mathbf{y}_l) : \begin{cases} \pi_l = F_l(\mathbf{t}_l), l = 1, 2, ..., m, \\ y_j = \frac{1}{\mu_j}\int_{-\infty}^{t_j} x_j dF_j(x_j), j = k_{l-1}+1, ..., k_l. \end{cases} \tag{5.14}$$

where $\mathbf{y}_l = (y_{k_{l-1}+1}, y_{k_{l-1}+2}, ..., y_{k_l})'$, $\mathbf{t}_l = (t_{k_{l-1}+1}, t_{k_{l-1}+2}, ..., t_{k_l})'$ and $k_0 = 0$. Hence the IPPF and MEDI can also be rewritten accordingly,

$$\mathcal{F}(y_1, y_2, ..., y_p) = \prod_{l=1}^{m} \mathcal{F}_l(\mathbf{y}_l), \tag{5.15}$$

$$\mathbf{MEDI}^{(p)} = \prod_{l=1}^{m} \int_{[0,1]^{\bigotimes k_l}} \mathcal{F}_l(\mathbf{y}_l) d\mathbf{y}_l = \prod_{l=1}^{m} \mathbf{MEDI}_l^{(k_l)}. \tag{5.16}$$

Remark 1 discloses the property of MEDI with independence. In addition, the following remarks explore some extreme cases of MEDI.

Remark 2: If the components are identical and independent(i.i.d) with the same marginal cdf $F_M(t)$, we can derive the IPPF $\mathcal{F}_M(y)$ by (5.5) and MEDI can be simplified as,

$$\mathbf{MEDI}^{(p)} = (\mathbf{MEDI})^p. \tag{5.17}$$

Especially, if each social resource is in the absolutely equal condition, $\mathcal{F}_M(y_j) = y_j$ and the MEDI will be $(\frac{1}{2})^p$ which is the lower bound of MEDI.

Remark 3: If the resources are identical and definitely correlated with the marginal cdf $F_M(t)$, the joint cdf will be $\min_{j=1,2,...,p} F_M(t_j)$. We can infer from (5.10) that the joint IPPF is consistent with the smallest value among the marginal IPPF $\mathcal{F}_M(y_j), j = 1, 2, ..., p$, i.e.,

$$\mathcal{F}(y_1, y_2, ..., y_p) = \min_{j=1,2,...,p} \{\mathcal{F}_M(y_j)\}. \tag{5.18}$$

Thus, the MEDI will be

$$\mathbf{MEDI}^{(p)} = \int_{[0,1]^{\bigotimes p}} \min_{j=1,2,...,p} \{\mathcal{F}_M(y_j)\} dy_1...dy_p. \tag{5.19}$$

Especially, if each resource lies in absolutely equal condition, MEDI reaches the lower bound at $\frac{1}{p+1}$.

$$\mathbf{MEDI}^{(p)} = \int_{[0,1]^{\bigotimes p}} \min_{j=1,2,\ldots,p} \{y_j\} dy_1 \ldots dy_p = \frac{1}{p+1}. \tag{5.20}$$

From the formulations (5.14)-(5.20), we can derive that MEDI takes value in the interval $[(\frac{1}{2})^p, 1]$ instead of the interval $[0, 1]$ of the Gini index. MEDI takes the value $(\frac{1}{2})^p$ with the mutual independence and absolute equality on each social resource, and MEDI takes 1 if all the resources are processed by only one person.

5.2 The Empirical MEDI and Its Statistical Properties

5.2.1 The Definition of Empirical MEDI

In the previous section, the definition of MEDI has been theoretically put forward and explained. In this section, we mainly discuss the empirical MEDI with statistical samples.

Suppose the n samples of the economic vector $\mathbf{X}$ have been obtained and denoted by the sample data matrix X,

$$X = \begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1p} \\ x_{21}, x_{22}, \ldots, x_{2p} \\ \ldots\ldots \\ x_{n1}, x_{n2}, \ldots, x_{np} \end{pmatrix}. \tag{5.21}$$

The empirical joint cdf of $\mathbf{X}$ is given by (5.22),

$$\hat{\pi} = \hat{F}_n(t_1, t_2, \ldots, t_p) = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \mathbf{1}_{(-\infty, t_j)}(x_{ij}) \\ 0, \quad otherwise \end{cases} \tag{5.22}$$

where $\mathbf{1}_A$ is the indicator of event A. For a fixed t_j, the indicator for $\mathbf{1}_{(-\infty, t_j)}(x_{ij})$

equals 1 if the ith sample value on jth component $x_{ij} \leq t_j$, hence the product $\prod_{j=1}^{P} \mathbf{1}_{(-\infty,t_j)}(x_{ij})$ equals 1 if the conditions $x_{i1} \leq t_1, x_{i2} \leq t_2, ..., x_{ip} \leq t_p$ are simultaneously satisfied.

The empirical IPPF $\hat{\mathcal{F}}_n(y_{n,1}, y_{n,2}, ..., y_{n,p})$ is defined by the equation (5.23) with respect to the sample marginal component percentages $\hat{y}_j$ of the vector,

$$\hat{\pi} = \hat{\mathcal{F}}_n(\hat{y}_1, \hat{y}_2, ..., \hat{y}_p) : \begin{cases} \hat{\pi} = \hat{F}_n(t_1, t_2, ..., t_n) \\ \hat{y}_j = \hat{L}_{nj}(t_j) = \frac{1}{n\hat{\mu}_j} \sum_{i=1}^{n} x_{ij} \mathbf{1}_{(-\infty,t_j)}(x_{ij}), \\ j = 1, 2, ..., p, \end{cases} \tag{5.23}$$

where $\hat{\mu}_j$ is the sample mean of the jth component of the vector,

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, j = 1, 2, ..., p. \tag{5.24}$$

In order to calculate the value of empirical MEDI with samples, we firstly determine the values of the web nodes formed by the samples. For the jth component, let $x_{(i)j}, i = 1, 2, ..., n$ denote the ascending order sample values of the data matrix X,

$$x_{(1)j} \leq x_{(2)j} \leq \ldots \leq x_{(n)j}, j = 1, 2, ..., p.$$

All these ascending samples formed the web nodes in $\mathbb{R}^p$. We can obtain the empirical percentage $\hat{y}_{i,j}$ on each node with the predefined equation (5.23),

$$\begin{aligned} \hat{y}_{i,j} &= \hat{L}_{nj}(x_{(i)j}) \\ \Delta y_{i,j} &= \hat{y}_{i,j} - \hat{y}_{i-1,j}, i = 1, 2, ..., n, \end{aligned} \tag{5.25}$$

where $y_{0,j} = 0$ for $j = 1, 2, ..., p$. Hence, we can get a grid formed by np nodes $(\hat{y}_{i_1,1}, \hat{y}_{i_2,2}, ..., \hat{y}_{i_p,p})$ in p-dimensional cubic space $[0, 1]^{\bigotimes p}$.

The empirical IPPF values on the nodes $\hat{\mathcal{F}}_n(\hat{y}_{1,j_1}, \hat{y}_{2,j_2}, ..., \hat{y}_{p,j_p})$ can be used to calculate the empirical MEDI $\hat{\mathbf{M}}$. Substitute the equations (5.22)-(5.25) into (5.11) and we can propose an algorithm to calculate the empirical p dimensional

$\hat{\mathbf{M}}^{(p)}$ as shown in (5.26),

$$\hat{\mathbf{M}}^{(p)} = \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \hat{\mathcal{F}}_n(\hat{y}_{i_1,1}, \hat{y}_{i_2,2}, ..., \hat{y}_{i_p,p}) \Delta y_{i_1,1} \cdots \Delta y_{i_p,p}. \tag{5.26}$$

5.2.2 The Simplification of Empirical MEDI

As shown in equations (5.12)-(5.20), the algorithms to achieve $\hat{\mathbf{M}}^{(p)}$ can also be simplified with different conditions of data,

Case 1: If the social resource vector $\mathbf{x}$ can be divided into m sub-vectors $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)^T$ and these m sub-vectors are mutually independent, the empirical MEDI can be composed of the product of the empirical MEDIs for the sub-vectors,

$$\hat{\mathbf{M}}^{(p)} = \prod_{l=1}^{m} \hat{\mathbf{M}}_l^{(k_l)}, \tag{5.27}$$

where $\hat{\mathbf{M}}_l^{(k_l)}$ is the empirical MEDI for the lth sub-vector with k_l components.

$$\hat{\mathbf{M}}_l^{(k_l)} = \sum_{i_{k_{l-1}+1}=1}^{n} \cdots \sum_{i_{k_l}=1}^{n} \hat{\mathcal{F}}_{n,l}(\hat{\mathbf{y}}_l) \Delta \hat{y}_{i_{k_{l-1}+1}, k_{l-1}+1} \cdots \Delta \hat{y}_{i_{k_l}, k_l}, l = 1, 2, ..., m \tag{5.28}$$

where $k_0 = 0$, $\hat{\mathbf{y}}_l = (\hat{y}_{i_{k_{l-1}+1}, k_{l-1}+1}, \hat{y}_{i_{k_{l-1}+2}, k_{l-1}+2}, \hat{y}_{i_{k_l}, k_l})'$ and $\hat{\mathcal{F}}_{n,l}(\cdot)$ is the marginal empirical IPPF with the marginal empirical cdf $\hat{F}_{n,l}(\mathbf{t}_l)$ for the lth sub-vector,

$$\hat{\mathcal{F}}_{n,l}(\hat{\mathbf{y}}_l) = \begin{cases} \hat{\mathcal{F}}_{n,l}(\mathbf{y}_l) = \hat{F}_{n,l}(\mathbf{t}_l), \\ \hat{y}_j = \hat{L}_{n,j}(t_j) = \frac{1}{n\hat{\mu}_j} \sum_{i=1}^{n} x_{ij} \mathbf{1}_{(-\infty, t_j)}(x_{ij}) \\ j = k_{l-1} + 1, ..., k_l, l = 1, 2, ..., m. \end{cases} \tag{5.29}$$

Especially if the components of the vector are mutually independent, the empirical MEDI can be obtained by (5.30),

$$\hat{\mathbf{M}}^{(p)} = \prod_{j=1}^{p} \hat{\mathbf{M}}_j = \prod_{j=1}^{p} \sum_{i=1}^{n} \hat{\mathcal{F}}_{nj}(\hat{y}_{i,j}) \Delta y_{i,j}, \tag{5.30}$$

where $\hat{\mathcal{F}}_{nj}(\cdot)$ is the marginal empirical IPPF for the jth component, $j = 1, 2, ..., p$.

Case 2: If all the components of the vector are identical and independent, denote the marginal empirical cdf and IPPF with $\hat{F}_{nM}(t)$ and $\hat{\mathcal{F}}_{nM}(y)$, respectively, and then the p-dimensional MEDI can be simplified as,

$$\hat{\mathbf{M}}^{(p)} = (\sum_{i=1}^{n} \hat{\mathcal{F}}_{nM}(\hat{y}_{i,j})\Delta y_{i,j})^p. \tag{5.31}$$

From (5.12), (5.15) and (5.20), we can conclude that the MEDI for the social resource vector can be formulated as the product of the MEDIs for the sub-vector or components with the condition of independence. This property provides a possibility to reduce the complexity of MEDI. If we can ensure that the resources are independent, we can calculate the MEDI for each resource separately and combine them to form the MEDI for the resource vector.

5.2.3 The Statistical Consistence of MEDI

The MEDI enjoys good statistical properties. We can also prove that the empirical $\hat{\mathbf{M}}$ converges to the MEDI under some wild assumptions.

Theorem 1 Suppose the joint cdf $F(x_1, x_2, ..., x_p)$ is a continuous function on each component variable and the mean μ_j for each component is finite, the empirical $\hat{\mathbf{M}}$ defined by (5.26) will converge to the MEDI, almost surely.

Denote $\Theta(\mathcal{F}) = \int_{[0,1]^{\otimes p}} \mathcal{F}(\mathbf{y}) dy_1 dy_2 ... dy_p$ and apply the mean value theorem and we would have equation (5.32),

$$\hat{\mathbf{M}}^{(p)} - \mathbf{MEDI}^{(p)} = \Theta(\hat{\mathcal{F}}_n) - \Theta(\mathcal{F}) = \hat{\mathcal{F}}_n(\mathbf{t}) - \mathcal{F}(\mathbf{t}), \tag{5.32}$$

where $\mathbf{t}$ is a fixed vector in $(0,1)^{\otimes p}$. According to the central limit theorem, we have

$$\frac{\hat{\mathcal{F}}_n(\mathbf{t}) - \mathcal{F}(\mathbf{t})}{\sqrt{\hat{\mathcal{F}}_n(\mathbf{t})(1-\hat{\mathcal{F}}_n(\mathbf{t}))/n}} \to N(0,1), \text{ as } n \to \infty, \tag{5.33}$$

For a specified significance level α, we can obtain,

$$P(|\hat{\mathbf{M}}^{(p)} - \mathbf{MEDI}^{(p)}| \leq Z_{\frac{\alpha}{2}} \delta_n) = 1 - \alpha, \tag{5.34}$$

where $Z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ right quantile for the standard Gaussian distribution and δ_n is the standard error of $\hat{\mathbf{M}}^{(p)}$,

$$\delta_n = \sqrt{\frac{\hat{\mathcal{F}}_n(\mathbf{t})(1-\hat{\mathcal{F}}_n(\mathbf{t}))}{n}} = \sqrt{\frac{\hat{\mathbf{M}}^{(p)}(1-\hat{\mathbf{M}}^{(p)})}{n}}. \tag{5.35}$$

Therefore we can give the $(1-\alpha)$ level confidence interval of MEDI with $\hat{\mathbf{M}}^{(p)}$, $(\hat{\mathbf{M}}^{(p)} - Z_{\frac{\alpha}{2}}\delta_n, \hat{\mathbf{M}}^{(p)} + Z_{\frac{\alpha}{2}}\delta_n)$,

5.2.4 Proof of Theorem1

Proof: Because the joint empirical IPPF $\hat{\mathcal{F}}_n(\hat{y}_1, \hat{y}_2, ..., \hat{y}_p)$ and joint empirical cdf $\hat{F}_n(t_1, t_2, ..., t_p)$ have the same statistical properties except for the different domains, we mainly consider the convergence of the joint empirical cdf $\hat{F}_n(t_1, t_2, ..., t_p)$.

The proof of the theorem can use the Glivenko-Cantelli theorem. For simplicity, we consider the case of continuous random for each vector component. Fix $-\infty = T_{i,0} < T_{i,1} < \cdots < T_{i,m_i} = \infty$ such that,

$$F(t_1, ..., t_{i-1}, T_{i,j}, ..., t_p) - F(t_1, ..., t_{i-1}, T_{i,j-1}, ..., t_p) = \frac{1}{m_i}, \tag{5.36}$$

for $i = 1, 2, .., p$ and $j = 1, 2, ..., m_i$. Now for any $t_i \in \mathbb{R}$, there exists $j \leq m_i$ such that $t_i \in [T_{i,j-1}, T_{i,j}]$. Note that

$$\begin{cases} \hat{F}_n(t_1, ..., t_i, ..., t_p) - F(t_1, ..., t_i, ..., t_p) \leq \\ \hat{F}_n(t_1, ..., T_{i,j}, ..., t_p) - F(t_1, ..., T_{i,j}, ..., t_p) + \frac{1}{m_i} \\ \hat{F}_n(t_1, ..., t_i, ..., t_p) - F(t_1, ..., t_i, ..., t_p) \geq \\ \hat{F}_n(t_1, ..., T_{i,j-1}, ..., t_p) - F(t_1, ..., T_{i,j-1}, ..., t_p) - \frac{1}{m_i} \end{cases}. \tag{5.37}$$

Therefore, almost surely,

$$\begin{aligned} \| \hat{F}_n - F \|_\infty = \sup\nolimits_{\mathbf{t}\in\mathbb{R}^p} |\hat{F}_n(\mathbf{t}) - F(\mathbf{t})| \leq \textstyle\sum_{i=1}^p \frac{1}{m_i} + \\ \textstyle\sum_{i=1}^p \max_{j_i \in \{1,...,m_i\}} |\hat{F}_n(T_{1,j_1}, ..., T_{i,j_i}, ..., T_{p,j_p}) - F(T_{1,j_1}, ..., T_{i,j_i}, ..., T_{p,j_p})|. \end{aligned} \tag{5.38}$$

Since $\max_{j\in\{1,2,...,m_i\}} |\hat{F}_n(t_1, ..., T_{i,j}, ..., t_p) - F(t_1, ..., T_{i,j}, ..., t_p)| \to 0$ a.s. by the strong law of large numbers, for any $\mathbf{m} = (m_1, m_2, ..., m_p)^T$ we can find N such that for all $n > N$,

$$\| \hat{F}_n - F \|_\infty \leq \sum_{i=1}^{p} \frac{1}{m_i} \quad a.s., \tag{5.39}$$

which guarantees the almost sure convergence of $\hat{F}_n$.

On one hand, since $\hat{F}_n$ converges to F almost surely, the empirical joint IPPF $\hat{\mathcal{F}}_n(\hat{\mathbf{y}})$ is the strictly monotone map of $\hat{F}_n$, hence also converges to $\mathcal{F}(\mathbf{y}), a.s.$. On the other hand, the empirical MEDI is the Riemann sum of the empirical IPPF, which will almost surely converges to the integration of the IPPF as the sample size increase to infinity.

$$\hat{\mathbf{M}}^p \rightarrow \mathbf{MEDI}^p \quad a.s.. \tag{5.40}$$

5.3 Simulation Studies

In this section, some simulations of MEDI will be shown. The main purpose of the simulation studies is to explore the consistency as well as the influential factors of the empirical MEDI. For computational simplicity, we only consider the $p = 2$ case. The simulation could be generalized to the p-dimensional case easily. We vary the distribution types, distribution skewness, distribution standard error σ and the correlation ρ between the two resources to examine their impact.

Let $\mathbf{x}$ and $\mathbf{y}$ denote the first and second resources, respectively. In our simulation, we set the sample size $n = 1000$ and consider different distribution assumptions. In each simulation, we conduct $1,000$ Monte Carlo replications to obtain the average MEDIs as well as the standard errors listed in the brackets.

Simulation 1: We conduct this simulation under four different settings of two resources: (1) Both resources are from uniform distribution: $\mathbf{x} \sim U(x_0 - r, x_0 + r)$, $\mathbf{y} \sim U(y_0 - r, y_0 + r)$, where $x_0 = 5$ and $y_0 = 3$ are given constants, and r is the radius of the interval. The correlation coefficient ρ of two resources is set as 0 or 1. (2) The first distribution is log-normal $\log(\mathbf{x}) \sim N(\mu_0, \sigma^2)$ and the second distribution is uniform, $\mathbf{y} \sim U(0, 2)$, where $\mu_0 = 2$ and $\sigma_0 = 1$. $\mathbf{x}$ and $\mathbf{y}$ are assumed to be independent. (3) $\log(\mathbf{x}) \sim N(\mu_0, \sigma^2)$ and $\log(\mathbf{y}) \sim N(\mu_0, \sigma^2)$, where $\mu_0 = 2$ and $\sigma_0 = 1$. (4) Introduce the Box-Cox transformation $\mathbf{a}^{(\lambda)} = \frac{\mathbf{a}^\lambda}{\lambda}$ related with five different values of positive λ on $\log(\mathbf{x}^{(\lambda)}) \sim N(\mu_0, \sigma_0^2)$ and $\log(\mathbf{y}^{(\lambda)}) \sim N(\mu_0, \sigma_0^2)$, where $\mu_0 = 2$, $\sigma_0 = 1$.

In these settings, we examine the consistency of the empirical MEDI under different conditions such as the extreme cases, different distributions and skewness degrees. In the first setting, the extreme cases of absolutely equality situations with independent and linear dependent variables are explored. For the second and third settings, empirical MEDIs under different distributions and variances are computed. Furthermore, the distributions with high skewness are also examined in forth setting.

Table 5.1: The MEDIs for simulation 1

(1)	r	0.001	0.01	0.05	0.25	1
Uniforms	$\mathbf{MEDI}^{(2)}$	0.2500	0.2504	0.2522	0.2612	0.2963
$\rho = 0$	$\hat{\mathbf{M}}^{(2)}$	0.2504	0.2506	0.2526	0.2611	0.2968
	(S.E.)	(0.0017)	(0.0026)	(0.0026)	(0.0027)	(0.0028)
Uniforms	$\mathbf{MEDI}^{(2)}$	0.3334	0.3338	0.3356	0.3446	0.3800
$\rho = 1$	$\hat{\mathbf{M}}^{(2)}$	0.3335	0.3340	0.3355	0.3450	0.3801
	(S.E.)	(0.000006)	(0.000007)	(0.00003)	(0.0002)	(0.0007)
(2)	σ	$\sigma_0/2$	$\sigma_0/\sqrt{2}$	σ_0	$\sqrt{2} * \sigma_0$	$2 * \sigma_0$
Uniform &	$\mathbf{MEDI}^{(2)}$	0.4254	0.4610	0.5068	0.5609	0.6142
Log-normal	$\hat{\mathbf{M}}^{(2)}$	0.4259	0.4619	0.5070	0.5609	0.6131
	(S.E.)	(0.0041)	(0.0049)	(0.0058)	(0.0072)	(0.0088)
(3)	σ	$\sigma_0/2$	$\sigma_0/\sqrt{2}$	σ_0	$\sqrt{2} * \sigma_0$	$2 * \sigma_0$
Log-normal &	$\mathbf{MEDI}^{(2)}$	0.4072	0.4781	0.5780	0.7079	0.8488
Log-normal	$\hat{\mathbf{M}}^{(2)}$	0.4077	0.4782	0.5783	0.7064	0.8454
	(S.E.)	(0.0037)	(0.0049)	(0.0073)	(0.0113)	(0.0161)
(4)	λ	0.1	0.3	0.5	1.5	2
Box-Cox	$\mathbf{MEDI}^{(2)}$	0.2789	0.3411	0.4073	0.7320	0.8488
Transformed	$\hat{\mathbf{M}}^{(2)}$	0.2797	0.3411	0.4077	0.7290	0.8452
Log-normals	(S.E.)	(0.0025)	(0.0026)	(0.0035)	(0.0099)	(0.0129)

The results of simulation 1 are listed in Table 5.1. It shows that the $\hat{\mathbf{M}}^{(2)}$ approaches theoretical $\mathbf{MEDI}^{(2)}$ in all given situations including different distributions, variance and skewness. That is, the empirical MEDI is an efficient estimator of real MEDI, therefore it is an effective evaluator to the inequality of the social resources.

In the following simulation, the empirical MEDI will be used to investigate the statistically influential factors of the social inequalities.

Simulation 2: We conduct this simulation under different assumptions of the distributions of the resources: (1) Both the distributions of the first and second resources are uniform $\mathbf{x} \sim U(x_0 - r, x_0 + r)$, $\mathbf{y} \sim U(y_0 - r, y_0 + r)$, where $x_0 = 5$ and $y_0 = 3$. The correlation coefficient ρ of the two resources is either 0 or 1. (2) The first distribution is log-normal $\log(\mathbf{x}) \sim N(\mu_0, \sigma^2)$ and the second distribution is uniform, $\mathbf{y} \sim U(0, 2)$, where $\mu_0 = 2$ and $\sigma_0 = 1$. $\mathbf{x}$ and $\mathbf{y}$ are independent. The variance of the second distribution varies. (3) $\log(\mathbf{x}) \sim N(\mu_0, \sigma_0^2)$ and $\log(\mathbf{y}) \sim N(\mu_0, \sigma_0^2)$, where $\mu_0 = 2$ and $\sigma_0 = 1$, ρ varies from -1 to 1. (4) $\log(\mathbf{x}) \sim N(\mu_0, \sigma^2)$ and $\log(\mathbf{y}) \sim N(\mu_0, \sigma^2)$, where $\mu_0 = 2$. The standard errors σ of both resources vary. (5) Box-Cox transformation related with λ is introduced to $\log(\mathbf{x}^{(\lambda)}) \sim N(\mu_0, \sigma_0^2)$ and $\log(\mathbf{y}^{(\lambda)}) \sim N(\mu_0, \sigma_0^2)$, where $\mu_0 = 2$, $\sigma_0 = 1$ and $\mathbf{a}^{(\lambda)} = \frac{\mathbf{a}^{\lambda}}{\lambda}$.

We examine the extreme cases under the first setting because both resources are almost distributed in the absolutely equal situations. Resources for every one just vary around a fixed value's neighborhood. We vary the radius r to examine its effect on the MEDI in condition of the correlations $\rho = 1$ and $\rho = 0$. Besides, we explore the effect of varying the standard error of one resource on MEDI under the second assumption. In this case, the distribution of $\mathbf{y}$ is fixed, and the standard error σ of $\mathbf{x}$ varies from $\sigma_0/4$ to $4 * \sigma_0$. After that, we further investigate the effect of the correlation between the resources. Ten different correlation coefficients are inspected. For the last situation, we study the impacts of the variance and skewness on social inequality. Five variances and nine values of positive λ are examined in order to detect the influences of the variance and skewness.

All the results are listed in Table 5.2, which shows that, in condition of the perfect equal social resources, MEDI approximates its theoretical lower bound at $\frac{1}{2^2} = 0.25$ and $\frac{1}{3}$ with $\rho = 0$ and $\rho = 1$, respectively. MEDI increases with r which discloses the inequality of the resource. MEDI will approach the theoretical upper bound 1 when social resources reach the other extreme case i.e. definitely unequal resources.

MEDI can interpret the whole inequality for all social resources. It will not be very large even if one of the social resources is extremely unequal, e.g. the MEDI increases from 0.3718 to 0.6555 as the inequality of the first variable ascends to a high level. However, the MEDI is not very high because the second social variable has comparably lower inequality. Although the standard error σ has been proven

Table 5.2: The estimated MEDIs for simulation 2

(1)	r	0.001	0.01	0.05	0.2	0.75
Uniforms	$\hat{\mathbf{M}}^{(2)}$	0.2504	0.2506	0.2526	0.2611	0.2968
$\rho = 0$	S.E.	(0.0017)	(0.0026)	(0.0026)	(0.0027)	(0.0028)
Uniforms	$\hat{\mathbf{M}}^{(2)}$	0.3335	0.3340	0.3355	0.3450	0.3801
$\rho = 1$	S.E.	(0.000006)	(0.000007)	(0.00003)	(0.0002)	(0.0007)
(2)	σ	$\sigma_0/4$	$\sigma_0/2$	σ_0	$2 * \sigma_0$	$4 * \sigma_0$
Uniform &	$\hat{\mathbf{M}}^{(2)}$	0.3806	0.4259	0.5070	0.6131	0.6622
log-normal	S.E.	(0.0036)	(0.0041)	(0.0058)	(0.0088)	(0.0049)
(3)	ρ	0.25	0.5	0.75	0.9	1
Log-normals	$\hat{\mathbf{M}}^{(2)}$	0.5904	0.6035	0.6184	0.6271	0.6338
	S.E.	(0.0074)	(0.0079)	(0.0083)	(0.0083)	(0.0083)
	ρ	-0.25	-0.5	-0.75	-0.9	-1
Log-normals	$\hat{\mathbf{M}}^{(2)}$	0.5666	0.5554	0.5451	0.5401	0.5362
	S.E.	(0.0073)	(0.0081)	(0.0084)	(0.0091)	(0.0093)
(4)	σ	$\sigma_0/4$	$\sigma_0/2$	σ_0	$2 * \sigma_0$	$4 * \sigma_0$
Log-normals	$\hat{\mathbf{M}}^{(2)}$	0.3256	0.4077	0.5783	0.8454	0.9848
	S.E.	(0.0029)	(0.0037)	(0.0073)	(0.0161)	(0.0063)
(5)	λ	0.1	0.2	0.3	0.5	1.5
Log-normals	$\hat{\mathbf{M}}^{(2)}$	0.2794	0.3099	0.3415	0.4077	0.7309
	S.E.	(0.0026)	(0.0027)	(0.0029)	(0.0036)	(0.0135)
	λ	2	3	5	10	
Log-normals	$\hat{\mathbf{M}}^{(2)}$	0.8439	0.9549	0.9935	0.9990	
	S.E.	(0.0156)	(0.0120)	(0.0036)	(0.0008)	

to have a large influence on the skewness of Log-normal distribution and on the social inequality, MEDI will be large only if all the social resources reach high unequal levels. Since we just vary the σ of one distribution and keep the other fixed, it does not change dramatically as expected.

The correlation takes less effect on the social inequality compared with skewness. However, positive correlation between resources will result in a larger inequality, while negative correlation will result in a smaller inequality.

Differing from the results of (2) in Table 5.2, the MEDI dramatically increases from 0.3106 to 0.9657 in (4) when both of the resources become unequal. It is even close to the upper bound with large standard error $\sigma = 4\sigma_0$, which verifies our judgement that the social inequality reaches the upper bound when both social resources are unequally distributed. (5) also shows the effect of skewness. The

MEDI dramatically decreases to 0.2741 when $\lambda = 0.1$ and almost reaches 1 when $\lambda = 10$. It demonstrates that the highly right skewed distribution will result in severe social inequality.

To summarize above simulations, MEDI can illustrate the inequality of multiple social resources. The entire inequalities of all resources take greater effect on MEDI than any individual resource. If the specific features of resource distribution are considered, the skewness influences MEDI most while the negative correlation can reduce the MEDI.

5.4 The Empirical Analysis in China Provincial Capital Cities

In this section, we apply the proposed MEDI to evaluate the inequalities of provincial capital cities in China. Our empirical analysis is based on a dataset concerning people's livelihood survey in these cities. The dataset is from a nationwide survey about people's livelihood in provincial capital cities held by Southwestern University of Finance and Economics. The research report of the survey was published in 2017. In this dataset, we have three important social resources: income, educational expenditure and asset. In order to clarify the inequalities of different crowd, we also choose three categorical covariates: job type, age and education level. As a comparison, we compute three indices, the original Gini index on income (abbreviated as GI), the MEDI based on income & educational expenditure (abbreviated as MEDI1) and the MEDI based on income & asset (abbreviated as MEDI2). Furthermore, to study the detailed inequality structures of the society, we give the indices and their 95% confidence intervals (abbreviated as 95%CI), and analyze the results according to each covariate. After removing the observations with missing values in income, education expenditure or asset, we have $n_1 = 4646$ and $n_2 = 5863$ valid observations for estimating $\hat{\mathbf{M}}_1^{(2)}$ and $\hat{\mathbf{M}}_2^{(2)}$ of MEDI1 and MEDI2, respectively.

The estimated overall MEDI1 is 0.4955 with 95% confidence interval $(0.4811, 0.5099)$, while the estimated overall MEDI2 is 0.5280 with 95% confidence interval $(0.5152, 0.5408)$. The latter is much larger than the former, which means the inequality level regarding of asset is more severe than that of educational expenditure.

According to the covariates, the samples are divided into six job types including public institution leaders (PIL), civil servants (CS), public institutions employees (PIE), enterprisers managers (EM), enterprisers employees (EE) and enterprisers R&D staffs (ERDS), six age groups from the 16-29 group to 70-79 one and six different educational levels such as primary school, junior high school, master & above and so on.

We study the detailed structures of inequality levels based on the estimated MEDIs as well as the confidence intervals.

Table 5.3: The Gini indices and MEDIs for job groups

Job Type	PIL	CS	PIE
$\hat{GI}$	0.4116	0.4597	0.4645
95%CI	(0.3795, 0.4436)	(0.4378, 0.4817)	(0.4372, 0.4919)
$\hat{\mathbf{M}}_1^{(2)}$	0.4665	0.4895	0.4693
95%CI	(0.4295, 0.5035)	(0.4635, 0.5154)	(0.4393, 0.4992)
$\hat{\mathbf{M}}_2^{(2)}$	0.5006	0.5329	0.5229
95%CI	(0.4680, 0.5332)	(0.5109, 0.5548)	(0.4955, 0.5503)

Job Type	EM	EE	ERDS
$\hat{GI}$	0.4747	0.4305	0.3826
95%CI	(0.4482, 0.5012)	(0.3712, 0.4898)	(0.2586, 0.5066)
$\hat{\mathbf{M}}_1^{(2)}$	0.5267	0.4652	0.4438
95%CI	(0.4984, 0.5550)	(0.3974, 0.5330)	(0.3018, 0.5859)
$\hat{\mathbf{M}}_2^{(2)}$	0.5289	0.5036	0.4607
95%CI	(0.5025, 0.5555)	(0.4437, 0.5634)	(0.3335, 0.5879)

Table 5.3 shows that the estimated values. $\hat{\mathbf{M}}_2^{(2)}$s are greater than $\hat{\mathbf{M}}_1^{(2)}$s almost in all the job types, which means in Chinese cities, the inequality in assets is more severe than that in expenditure. The results provide that MEDI can reliably capture more social inequality than Gini index. The indies will vary if different resources are taken into consideration.

The results also illustrate that there are significant differences among groups. Enterpriser managers have the largest Gini index and MEDI1 while civil servants have the largest MEDI2 and second largest MEDI1. These two groups also have inner inequality values where both the MEDIs higher than the average levels. Moreover, the inequality for enterpriser managers on income and expenditure is significantly larger than the average level because the MEDI1 value of enterpriser managers exceeds the 95% confident interval. Enterpriser R&D staff have the lowest MEDI1 and MEDI2 which means there is significantly lower inequality on income, expenditure and assets, for the values of MEDI1 and MEDI2 exceed the confident intervals. Both MEDI1 and MEDI2 of public institution leaders and enterpriser employees are also significantly lower than the average level.

We can infer that the inequalities on income, expenditure and assets are significantly different for different job types in China. First of all, enterpriser managers present high inequality on their income because the corresponding Gini index is the largest. Secondly, since the MEDI2 of the CS group is the largest, civil staffs may have high inequality regarding their asset, as do the public institution employees. The reason for this difference likely lies within the regional gap of the asset, especially the price gap in housing which is the most share of families' wealth for civil staff and public institution employees. In addition, enterpriser R&D staffs and public institution leaders are in low inequality because they have comparable income and expenditure. Finally, enterpriser employees get the relatively small Gini index, MEDI1 and MEDI2 as they have less income and assets than the other job types.

Table 5.4: The Gini indices and MEDIs for age groups

Age	$16-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70-79$
$\hat{GI}$	0.4423	0.5005	0.4345	0.4302	0.4658	0.4340
95%CI	(0.4188, 0.4659)	(0.4737, 0.5274)	(0.4071, 0.4620)	(0.3982, 0.4622)	(0.4192, 0.5124)	(0.3663, 0.5016)
$\hat{\mathbf{M}}_1^{(2)}$	0.4769	0.5332	0.4753	0.4767	0.4872	0.4996
95%CI	(0.4469, 0.5068)	(0.5065, 0.5599)	(0.4491, 0.5015)	(0.4309, 0.5225)	(0.4292, 0.5453)	(0.3980, 0.6012)
$\hat{\mathbf{M}}_2^{(2)}$	0.5234	0.5511	0.4983	0.4599	0.5207	0.5067
95%CI	(0.4998, 0.5471)	(0.5244, 0.5778)	(0.4706, 0.5260)	(0.4277, 0.4921)	(0.4740, 0.5673)	(0.4384, 0.5749)

Table 5.4 depicts that the MEDI concerning two resources can capture more inequality than the Gini index. The age groups of $40-49$, $50-59$ and $70-79$ have the similar values on the Gini indices. However, the MEDI1 and MEDI2 of these groups are quite different when new resources are included. The $70-79$ group has significantly higher MEDIs than the other groups. The main reason lies in the great inequalities of educational expenditure and assets. In China, the parents of age $30-49$ often have to pay the expense of their children's education as well as the cost of personal training. The grandparents of age $60-79$ sometimes partially cover the costs of their grandchildren's educations. Among these educational expenditures, there is a large gap between urban and rural regions, between the eastern developed and western developing areas. Therefore the inequality in educational expenditure is now revealed by MEDI1 which the Gini index did not capture. Additionally, the age group of $30-39$ has the highest MEDI2 among all the groups and is significantly higher than the average level. The age group of $50-59$ comes with the lowest MEDI2 and is significantly lower than the overall average value. The inequality of income and expenditure does not differ a lot except for the age group $30-39$ while the inequality of income and asset starts high when people are younger than 39, remains low during 40-59 and becomes high again with increasing age .

We can conclude from Table 5.4 that there exist different inequalities in age groups. The income and expenditure of young people is comparatively equal but the assets are quite unequal possibly because of their parents. The middle-aged people have the biggest inequalities on income, expenditure and asset, which is partly due to the differences in ability, opportunity, region, consumption and wealth. Some of the old people spend much on the education on their grandchildren and themselves, which results in greater MEDI1.

Table 5.5: The Gini indices and MEDIs for education groups

Education	Primary School	Junior High School	High School	Junior College	College	Master & above
$\hat{GI}$	0.4162	0.4110	0.3746	0.3903	0.4210	0.5205
95%CI	(0.3713, 0.4611)	(0.3826, 0.4393)	(0.3502, 0.3991)	(0.3617, 0.4189)	(0.3952, 0.4468)	(0.4539, 0.5871)
$\hat{\mathbf{M}}_1^{(2)}$	0.4841	0.4695	0.4543	0.4533	0.4439	0.5393
95%CI	(0.4310, 0.5371)	(0.4383, 0.5007)	(0.4272 0.4814)	(0.4190, 0.4876)	(0.4139, 0.4738)	(0.4625, 0.6160)
$\hat{\mathbf{M}}_2^{(2)}$	0.4961	0.4869	0.4832	0.4920	0.5108	0.5367
95%CI	(0.4505, 0.5416)	(0.4581, 0.5157)	(0.4579, 0.5084)	(0.4627, 0.5213)	(0.4847, 0.5370)	(0.4702, 0.6032)

Table 5.5 shows that people with master's degree have very polarized income and education expenditure, where even the lower bound of the confidence interval is much larger than the overall level (0.4955). All the other five groups of people have MEDI smaller than the overall level. The people with master's degree have inequality level larger than the overall level. All the other five groups of people have MEDI smaller than the overall level.

We can infer that in Chinese cities, inequalities may become more severe with regard to the three social resources when the education level is developed. We also notice that the inequality of primary school is a little larger than the groups of junior school, high school and junior college, which partly depends on the advance of technology. Some of the primary students have become skilled workers with high income, educational expenditure and assets while other students may still remain in a low social resources' level, which is the reason why primary students differ a lot in the social resources.

Chapter 6

Conclusion and Future Research

6.1 Conclusion

In chapter 3, we propose a sure joint screening procedure for the varying coefficient Cox model with ultrahigh dimensional covariates based on partial likelihood. The proposed SJS is distinguished from the existing SIS procedure in that the proposed procedure is based on the joint likelihood of potential candidate features. We propose an effective algorithm to carry out the feature screening procedure, and show that the proposed algorithm possesses an ascent property. We study the sampling property of SJS, and establish the sure screening property for SJS. Theorem 1 ensures the ascent property of the proposed algorithm under certain conditions, but it does not implies that the proposed algorithm converges to the global optimizer. If the proposed algorithm converges to a global maximizer of (3.5), then Theorem 2 shows that such a solution enjoys the sure screen property.

In chapter 4, we propose a two-layer framework, “Iteratively Kings’ Forests”, to select important features and interaction effects in classfication and regresssion problems. In the first layer, assume that one variable is important, we treat it as a “King” and construct an iteratively weighted forest with the “King” as the root node of every tree. Through this iterative process, variables participating in the same interactions with the “King” will gradually gain larger weights, and therefore be likely to line in the same path of different trees. In the second layer, we iteratively choose the next “King” and construct “King’s Forest” for each of them. We keep doing this until some convergence criterion are met. This two-layer procedure outlines the hidden model structure by selecting important features and interactions through iteratively constructing forests for important variables. Based on it, a thorough and in-depth exploration is conducted to unveil the hidden mechanism for us.

In chapter 5, we extend the definition of the Gini index to multiple dimensional cases and show its statistical consistence. The new defined inequality index, multidimensional economic dispersion index (MEDI), is more suitable for measuring and interpreting the statistical dispersion in complex economic data analysis. It could be widely used when we consider evaluating the inequality level using more than one social resource. We also propose the algorithm to compute sample-empirical MEDI which converges to the MEDI in some conditions. The MEDI can also be applied in other sciential and social fields as the Gini index did. The simulations and empirical analysis show that the MEDI can summarize multiple variables to form an index. Therefore, the new index is actually a more general and advanced method to explain the inequality in multivariate economic data.

6.2 Future Research

For the "Iteratively Kings' Forests" algorithm proposed in chapter 4, there are a number of potential future improvements. For example, we use a simple function as the mathematical form of a tree structure. Based on this, we may explore the theoretical properties of the forest and try to prove that the "Iteratively Kings' Forests" is good theoretically. Moreover, as we point before, the greedy algorithm used to search and split each node allows tree structure to include the variable that brings the most significant instant improvement, which could be either a marginal effect or a member of an interaction effect. That is, variables in one path of a tree will be a mixture of marginal effects and interaction effects. Therefore, another potential improvement is to create an algorithm or criterion to effectively separate one effect from another.

For the MEDI proposed in chapter 5, there are also some valuable questions lfet for deeper investigation. We already generalize the Gini index to the multivariate case. For the next step, we may examine how it works in high-dimensional cases. Moreover, the algorithm is a common formulation currently based on the joint empirical cdf. Maybe it can be simplified under some assumptions of data distribution.

Bibliography

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

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics 16*(1), 125–127.

Amaratunga, D., J. Cabrera, and Y.-S. Lee (2008). Enriched random forests. *Bioinformatics 24*(18), 2010–2014.

Anand, S. (1983). *Inequality and poverty in Malaysia: Measurement and decomposition.* The World Bank.

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

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association 96*(455), 939–967.

Archer, K. J. and R. V. Kimes (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis 52*(4), 2249–2260.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of economic theory 2*(3), 244–263.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series 19*(3), 357–367.

Basu, S., K. Kumbier, J. B. Brown, and B. Yu (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences 115*(8), 1943–1948.

Bradic, J., J. Fan, and J. Jiang (2011). Regularization for coxs proportional hazards model with np-dimensionality. *Annals of statistics 39*(6), 3092–3120.

Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics 5*(1), 232–253.

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

Cai, Z. and Y. Sun (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics 30*(1), 93–111.

Capovilla, M., E. D. Eldon, and V. Pirrotta (1992). The giant gene of drosophila encodes a b-zip dna-binding protein that regulates the expression of other segmentation gap genes. *Development 114*(1), 99–112.

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

Cheng, M.-Y., T. Honda, and J.-T. Zhang (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association 111*(515), 1209–1221.

Chu, W., R. Li, and M. Reimherr (2016). Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *The annals of applied statistics 10*(2), 596–617.

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

De Boor, C., C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor (1978). *A practical guide to splines*, Volume 27. Springer-Verlag New York.

Díaz-Uriarte, R. and S. A. De Andres (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics 7*(1), 1–13.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Eldon, E. D. and V. Pirrotta (1991). Interactions of the drosophila gap gene giant with maternal and zygotic pattern-forming genes. *Development 111*(2), 367–378.

Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association 106*(494), 544–557.

Fan, J., Y. Feng, Y. Wu, et al. (2010). High-dimensional variable selection for coxs proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pp. 70–86. Institute of Mathematical Statistics.

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

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

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

Fan, J., Y. Ma, and W. Dai (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association 109*(507), 1270–1284.

Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics 38*(6), 3567–3604.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics 42*(3), 819–849.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 1491–1518.

Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and its Interface 1*(1), 179–195.

Firebaugh, G. (1999). Empirics of world income inequality. *American Journal of Sociology 104*(6), 1597–1630.

Foster, D. P. and E. I. George (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 1947–1975.

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

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

Gini, C. (1912). Variability a and borrowing a. contribution to the study of distributions and statistical relations. *Typography of Cuppini*.

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal 31*(121), 124–126.

Gini, C. (2005). On the measurement of concentration and variability of characters. *METRON-International Journal of Statistics 63*(1), 1–38.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association 87*(420), 942–951.

Greselin, F. (2014). More equal and poorer, or richer but more unequal? *Economic Quality Control 29*(2), 99–117.

Greselin, F., L. Pasquazzi, and R. Zitikis (2013). Contrasting the gini and zenga indices of economic inequality. *Journal of Applied Statistics 40*(2), 282–297.

Hall, P. and H. Miller (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics 18*(3), 533–550.

Hansen, J., M. Bentley, H. W. Van Der Hart, M. Landtman, G. Lister, Y. Shen, and N. Vaeck (1993). The introduction of b-spline basis sets in atomic structure calculations. *Physica Scripta 1993*(T47), 7.

Harrison, M. M., X.-Y. Li, T. Kaplan, M. R. Botchan, and M. B. Eisen (2011). Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS genetics 7*(10), e1002266.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hillebrand, E. et al. (2009). Poverty, growth, and inequality over the next 50 years. In *How to feed the World in 2050. Proceedings of a technical meeting of experts, Rome, Italy, 24-26 June 2009*, pp. 1–23. Food and Agriculture Organization of the United Nations (FAO).

Hoch, M., C. Schröder, E. Seifert, and H. Jäckle (1990). cis-acting control elements for krüppel expression in the drosophila embryo. *The EMBO journal 9*(8), 2587–2595.

Hoch, M., E. Seifert, and H. Jäckle (1991). Gene expression mediated by cis-acting sequences of the krüppel gene in response to the drosophila morphogens bicoid and hunchback. *The EMBO journal 10*(8), 2267–2278.

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

Huang, J., T. Sun, Z. Ying, Y. Yu, and C.-H. Zhang (2013). Oracle inequalities for the lasso in the cox model. *Annals of statistics 41*(3), 1142–1165.

Huang, J. Z. and H. Shen (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian journal of statistics 31*(4), 515–534.

Imedio-Olmedo, L. J., E. Bárcena-Martín, and E. M. Parrado-Gallardo (2011). A class of bonferroni inequality indices. *Journal of Public Economic Theory 13*(1), 97–124.

Jasso, G. (1979). On gini's mean difference and gini's index of concentration. *American Sociological Review 44*(5), 867–870.

Kendall, M. and A. Stuart (1977). The advanced theory of statistics. vol. 1: Distribution theory. *London: Griffin, 1977, 4th ed.*.

Kim, Y. H., D. C. Jeong, K. Pak, M.-E. Han, J.-Y. Kim, L. Liangwen, H. J. Kim, T. W. Kim, T. H. Kim, D. W. Hyun, et al. (2017). Slc2a2 (glut2) as a novel prognostic factor for hepatocellular carcinoma. *Oncotarget 8*(40), 68381–68392.

Klugman, J. (2010). Human development report 2010–20th anniversary edition. the real wealth of nations: Pathways to human development.

Kopczuk, W., E. Saez, and J. Song (2010). Earnings inequality and mobility in the united states: evidence from social security data since 1937. *The Quarterly Journal of Economics 125*(1), 91–128.

Kraut, R. and M. Levine (1991). Spatial regulation of the gap gene giant during drosophila development. *Development 111*(2), 601–609.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology 20*(17), R754–R763.

Levine, M. (2013). Development: Computing away the magic? *eLife 2*, e01135.

Li, J., W. Zhong, R. Li, and R. Wu (2014). A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *The annals of applied statistics 8*(4), 2292–2318.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Li, X.-y., S. MacArthur, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. L. Hendriks, et al. (2008). Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS biology 6*(2), e27.

Liang, H.-L., C.-Y. Nien, H.-Y. Liu, M. M. Metzstein, N. Kirov, and C. Rushlow (2008). The zinc-finger protein zelda is a key activator of the early zygotic genome in drosophila. *Nature 456*(7220), 400.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

Liu, W.-B., W.-D. Jia, J.-L. Ma, G.-L. Xu, H.-C. Zhou, Y. Peng, and W. Wang (2017). Knockdown of gtpbp4 inhibits cell growth and survival in human hepatocellular carcinoma and its prognostic significance. *Oncotarget 8*(55), 93984–93997.

Maasoumi, E. and S. Zandvakili (1986). A class of generalized measures of mobility with applications. *Economics Letters 22*(1), 97–102.

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

Markstein, M., R. Zinzen, P. Markstein, K.-P. Yee, A. Erives, A. Stathopoulos, and M. Levine (2004). A regulatory code for neurogenic gene expression in the drosophila embryo. *Development 131*(10), 2387–2394.

Miller, A. (2002). *Subset selection in regression.* CRC Press.

Morán, É. and G. Jiménez (2006). The tailless nuclear receptor acts as a dedicated repressor in the early drosophila embryo. *Molecular and Cellular Biology 26*(9), 3446–3454.

Nguyen, H. T. and X. Xu (1998). Drosophila mef2expression during mesoderm development is controlled by a complex array ofcis-acting regulatory modules. *Developmental biology 204*(2), 550–566.

Nien, C.-Y., H.-L. Liang, S. Butcher, Y. Sun, S. Fu, T. Gocha, N. Kirov, J. R. Manak, and C. Rushlow (2011). Temporal coordination of gene networks by zelda in the early drosophila embryo. *PLoS genetics 7*(10), e1002339.

Niu, Y. S., N. Hao, and H. H. Zhang (2018). Interaction screening by partial correlation. *Statistics and Its Interface 11*(2), 317–325.

Pyatt, G. (1976). On the interpretation and disaggregation of gini coefficients. *The Economic Journal 86*(342), 243–255.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 233–243.

Rivera-Pomar, R. and H. Jackle (1996). From gradients to stripes in drosophila embryogenesis: filling in the gaps. *Trends in Genetics 12*(11), 478–483.

Roemer, J. E. (2013). Economic development as opportunity equalization. *The World Bank Economic Review 28*(2), 189–209.

Sadras, V. and R. Bongiovanni (2004). Use of lorenz curves and gini coefficients to assess yield inequality within paddocks. *Field Crops Research 90*(2-3), 303–310.

Schulz, C. and D. Tautz (1994). Autonomous concentration-dependent activation and repression of kruppel by hunchback in the drosophila embryo. *Development 120*(10), 3043–3049.

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

Sen, A., M. A. Sen, S. Amartya, J. E. Foster, J. E. Foster, et al. (1997). *On economic inequality.* Oxford University Press.

Shah, R. D. and N. Meinshausen (2014). Random intersection trees. *The Journal of Machine Learning Research 15*(1), 629–654.

Shorrocks, A. F. (1978). The measurement of mobility. *Econometrica: Journal of the Econometric Society*, 1013–1024.

Silber, J. (1989). Factor components, population subgroups and the computation of the gini index of inequality. *The Review of Economics and Statistics*, 107–115.

Song, R., F. Yi, and H. Zou (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica 24*(4), 1735–1752.

Stathopoulos, A., M. Van Drenth, A. Erives, M. Markstein, and M. Levine (2002). Whole-genome analysis of dorsal-ventral patterning in the drosophila embryo. *Cell 111*(5), 687–701.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, 1040–1053.

Stone, C. J. et al. (1985). Additive regression and other nonparametric models. *The annals of Statistics 13*(2), 689–705.

Struhl, G., P. Johnston, and P. A. Lawrence (1992). Control of drosophila body pattern by the hunchback morphogen gradient. *Cell 69*(2), 237–249.

Sun, Y., R. Sundaram, and Y. Zhao (2009). Empirical likelihood inference for the cox model with time-dependent coefficients via local partial likelihood. *Scandinavian Journal of Statistics 36*(3), 444–462.

Sundrum, R. M. (2003). *Income distribution in less developed countries.* Routledge.

Székely, G. J., M. L. Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics 35*(6), 2769–2794.

Thomas, V., Y. Wang, and X. Fan (2001). Measuring education inequality: Gini coefficients of education. policy research working paper.

Tian, L., D. Zucker, and L. Wei (2005). On the cox model with time-varying regression coefficients. *Journal of the American statistical Association 100*(469), 172–183.

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

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

Wang, L., Y. Kim, and R. Li (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics 41*(5), 2505–2536.

Wei, F., J. Huang, and H. Li (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica 21*(4), 1515–1540.

Weymark, J. A. (2003). Generalized gini indices of equality of opportunity. *The Journal of Economic Inequality 1*(1), 5–24.

Winham, S. J., R. R. Freimuth, and J. M. Biernacka (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal 6*(6), 496–505.

Xia, X., H. Yang, and J. Li (2016). Feature screening for generalized varying coefficient models with application to dichotomous responses. *Computational Statistics & Data Analysis 102*, 85–97.

Xu, C. and J. Chen (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association 109*(507), 1257–1269.

Xu, P., L. Zhu, and Y. Li (2014). Ultrahigh dimensional time course feature selection. *Biometrics 70*(2), 356–365.

Yang, G., Y. Yu, R. Li, and A. Buu (2016). Feature screening in ultrahigh dimensional cox's model. *Statistica Sinica 26*, 881–901.

Zeitlinger, J., R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young, and M. Levine (2007). Whole-genome chip–chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo. *Genes & development 21*(4), 385–390.

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

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

Zhang, W. and B. Sun (2015). Impact of age on the survival of patients with liver cancer: an analysis of 27,255 patients in the seer database. *Oncotarget 6*(2), 633–641.

Zhao, S. D. and Y. Li (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis 105*(1), 397–411.

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

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics 36*(4), 1509–1533.

Zucker, D. M. and A. F. Karr (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics 18*(1), 329–353.

Vita

Ling Zhang

Education

- **Ph.D., Statistics** Pennsylvania State University, University Park, PA, 2019 Dissertation: Procedures for Feature Screening and Interaction Identification in High-dimensional Data Modelling Advisors: Dr. Runze Li
- **M.S., Statistics** Auburn University, Auburn, AL, 2014
- **B.S., Statistics** Beijing Normal University, Beijing, China, 2012

Research Expertise

- **Main Research Area:** High Dimensional Variable Selection; Ultra-High Dimensional Feature Screening; Random Forest Algorithms
- **Familiar with:** Statistical Learning; Statistical Computing;

Publication

- **Ling Zhang**, Runze Li, A New Forest-based Screening Procedure for Nonlinear and Interaction Effects, to submit
- Yifan Xia, **Ling Zhang**, Iris L. Li, Multidimensional Economic Dispersion Index and Application, Journal of Business & Economic Statistics, under review
- Guangren Yang, **Ling Zhang**, Runze Li, Feature Screening in Ultrahigh Dimensional Varying-coefficient Cox's Model, Journal of Multivariate Analysis, Volume 171, Pages 284-297

Internship Experience

- **Sr. Algorithms Engineer, Alibaba** Beijing, China, 2017