DOCTORAL THESIS

The estimation and inference of complex models

Zhou, Min

Date of Award:
2017

HONG KONG BAPTIST UNIVERSITY

Doctor of Philosophy

THESIS ACCEPTANCE

DATE: August 24, 2017

STUDENT'S NAME: ZHOU Min

THESIS TITLE: The Estimation and Inference of Complex Models

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairman:	Dr. Chu Xiaowen Associate Professor, Department of Computer Science, HKBU (Designated by Dean of Faculty of Science)
Internal Members:	Prof. Liao Lizhi Professor, Department of Mathematics, HKBU (Designated by Head of Department of Mathematics)
	Prof. Zhu Lixing Chair Professor in Statistics, Department of Mathematics, HKBU
External Members:	Dr. Feng Yang Associate Professor Department of Statistics Columbia University
	Dr. Wang Junhui Associate Professor Department of Mathematics City University of Hong Kong
Proxy:	Dr. Tong Tiejun Associate Professor, Department of Mathematics, HKBU (as proxy for Dr. Feng Yang)
In-attendance:	Dr. Peng Heng Associate Professor, Department of Mathematics, HKBU

Issued by Graduate School, HKBU

The Estimation and Inference of Complex Models

ZHOU Min

A thesis submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

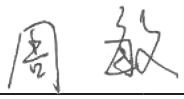
Principal Supervisor:
Dr. PENG Heng (Hong Kong Baptist University)

August 2017

DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of MPhil (or PhD as appropriate) at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University's Committee on the Use of Human & Animal Subjects in Teaching and Research (HASC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: 

Date: August 2017

Abstract

In this thesis, we investigate the estimation problem and inference problem for the complex models. Two major categories of complex models are emphasized by us, one is generalized linear models, the other is time series models. For the generalized linear models, we consider one fundamental problem about sure screening for interaction terms in ultra-high dimensional feature space; for time series models, an important model assumption about Markov property is considered by us.

The first part of this thesis illustrates the significant interaction pursuit problem for ultra-high dimensional models with two-way interaction effects. We propose a simple sure screening procedure (SSI) to detect significant interactions between the explanatory variables and the response variable in the high or ultra-high dimensional generalized linear regression models. Sure screening method is a simple, but powerful tool for the first step of feature selection or variable selection for ultra-high dimensional data. We investigate the sure screening properties of the proposal method from theoretical insight. Furthermore, we indicate that our proposed method can control the false discovery rate at a reasonable size, so the regularized variable selection methods can be easily applied to get more accurate feature selection in the following model selection procedures. Moreover, from the viewpoint of computational efficiency, we suggest a much more efficient algorithm-discretized SSI (DSSI) to realize our proposed sure screening method in practice. And we also investigate the properties of these two algorithms SSI and DSSI in simulation studies and apply them to some real data analyses for illustration.

For the second part, our concern is the testing of the Markov property in time series processes. Markovian assumption plays an extremely important role in time series analysis and is also a fundamental assumption in economic and financial models. However, few existing research mainly focused on how to test the Markov properties for the time series processes. Therefore, for the Markovian assumption, we propose a new test procedure to check if the time series with beta-mixing possesses the Markov property. Our test is based on the Conditional Distance Covariance (CDCov). We investigate the theoretical properties of the proposed method. The asymptotic dis-

tribution of the proposed test statistic under the null hypothesis is obtained, and the power of the test procedure under local alternative hypothesizes have been studied. Simulation studies are conducted to demonstrate the finite sample performance of our test.

Keywords: Variable selection; Interaction screening; Log-likelihood functions; Boolean representation; Data discretization; Markov property; Conditional independence; Conditional distance covariance; β-mixing; Complex models; Generalized linear models; Time series models.

Acknowledgements

First and foremost, I would like to express my deep gratitude to my supervisor and mentor Dr. PENG Heng for all of his inspiring guidance and enthusiastic support. His superb intuition, broad knowledge, and continuous encouragement have been indispensable throughout my PH.D. study and are extraordinarily beneficial in my future research career. And I learned lots of various things from him in the fields of statistics during these years. It is my great honor to be his student. I would also like to thank my co-supervisor Dr. ZENG Tieyong, for his help and support.

I am very grateful to Dr. CHU Xiaowen, Prof. LIAO Lizhi, Prof. ZHU Lixing, Dr. FENG Yang, Dr. TONG Tiejun, Dr. WANG Junhui, who serve on my dissertation committee.

Equally, I would like to thank Dr. YANG Can and Dr. YAO Yuan for their insightful comments and help on my study. They inspired my work in numerous discussions and I greatly benefited from Dr YANG's expertise in biostatistics and computational skills in particular. And also I wish to thank Mr. DAI Mingwei and Dr. ZHU Xuehu for helping me develop some algorithms and sharing some comments in this thesis. A part of this work is done in collaboration with Dr. YANG Can, Dr. YAO Yuan and Mr. DAI Mingwei. Without their support and guidance, this work would not have been completed.

My special appreciation also goes to Prof. FANG Kaitai in Beijing Normal University-Hong Kong Baptist University United International College (UIC). He is a respectable and famous statistician in the world. I am very lucky to be a teaching assistant of Prof. FANG during five years in UIC. He brings me into the world of Statistics. His sensitivity and foresight in research strongly impresses me and inspires me.

I am very grateful to other faculty members in the mathematics department, especially, CHUI Claudia, YUM Rainbow, LAM Tammy, YEUNG C. W., HUI Vicky, LI Candy and CHEUNG Jenny for their great help. And also, I would like to thank LO Kamfai of Graduate School for his excellent help.

My thanks extend to all staffs and postgraduate students in our department for

their direct and indirect help in the last few years. Dr. TONG Tiejun, Prof. YUAN Xiaoming and Prof. ZHU Lixing have been always supportive and instructive. I am very appreciated that Prof. ZHU Lixing and Prof. YUAN Xiaoming provided wonderful lecture about the dimension reduction, asymptotic theory and optimization theory. And also, I appreciate all classmates in our department, especially, Dr. DONG Kai, Dr. HUANG Peng, Dr. LIU Peng, Dr. SHEN Chenyang, CAI Mingxuan, HU Xianghong, HU Zongliang, LIN Enxuan, LIU Ye, LUO Dehui, MING Jingsi, SIU Kawai, TAN Falong, WU Chong, XIE Chuanlong, YAN Hanjun, ZHAO Jingxin, ZENG Shangzhi, ZHU Zhaochen for their selfless help and support. And I also would like to thank LI Jiarong very much for helping me revise my thesis and giving me some comments on the language.

I also acknowledge the support and help from colleagues in Department of Statistics of UIC and staff in UIC during my PH.D. application period, especially, Prof. ZHANG Jianzhong, Prof. YE Huajun, Prof. TSANG Kang-Too, Dr. HE Ping, Dr. PENG Xiaoling, Dr. DENG Yuhui, Dr. CHEUNG Chin-Wing, Dr. Li Yafei, FU Songfeng, KE Xiao.

I wish to thank Hong Kong Baptist University for providing me the chance for my Ph.D. programme with financial aid and Department of Mathematics for providing the research atmosphere which gives rise to several interesting discussions and ideas, on topics both more and less related to my research.

Last but not least, I wish to express my gratitude to my parents for their love, understanding, support, and encouragement during these years.

Table of Contents

Chapter 1

Introduction

With the rapid advances in science and technology, data can be stored more massively and cheaply. The dimension of one data set can grow with (even faster than) the number of observations. Consequently, data scientists encounter the data with high or ultra-high dimensional data− "Big Data" in many fields such as biology, chemistry, economics, finance, genetics, neuroscience, and physics. It brings new opportunities and challenges to data scientists (See Fan et al. [2014a]). For a large scale data set, traditional methods or simple models may not work any more for data analysis. To deal with the challenges of big data, complex models are considered to meet the big data's needs. Complex models can improve our statistical inference and help us obtain useful information to estimate or predict the behaviours of some events or activities. The statistical inference is to make the inferences about a population based the samples that derive from it. Usually, three types of inference are estimation, interval estimation and hypothesis testing. This thesis is concerned primarily with the parameter estimation of complex models and hypothesis testing of one of complex model's assumptions.

In the era of big data, variable selection and hypothesis testing becomes more prevalent in a wide range of fields. Variable screening or feature screening is definitely associated with the parameter estimation in complex models. It can boost the prediction and estimation of our models. And one of the complex models is generalized linear models. Thus, the first part of this thesis pays more attention to interaction pursuit in this kind of complex models. On the other hand, the second

part focuses on the hypothesis testing of Markov property in the times series models such as AR models, ARCH models, and GARCH models, which are another one kind of complex models.

1.1 Variable Selection or Feature selection in Ultra-high Dimensional Feature Space

The word "interaction", in Oxford English Dictionary, is illustrated as the reciprocal action, action or influence of persons or things on each other. It is one kind of relationship among two or more objects, which have mutual influence upon one another. There has been a long history of interaction, which is widely used in many scientific fields. For example, in Experimental Design, the interaction of two or more factors helps us to plan a study in order to arrive at some goals of an experiment. It is defined as the statistical departure from the additive effects of two or more factors in the experiment by Fisher [1918]. Later, Cox [1984] provided a broad review of the various aspects of interactions in the experimental designs. In Physical Chemistry, the main topics are interactions about atoms and molecules. A simple example in real-world is that neither of carbon and steel has much effect on the strength but a combination of them has an awfully significant effect. And, Wigner [1934] considered the interaction of between free electrons in an electron gas. In Medicine and Pharmacology, the pharmacist needs to monitor the combination of medications used for the patients. One of the interactions between medications (drug interaction) is in pharmacodynamics [Lees et al. [2004]], which involves the actions of the two interacting drugs. Gene-gene and gene-environment are studied by researchers in Biology and Genetics. They play significant roles in complex diseases. Interaction between two different genes is also called "Epistatic interaction" or "Epistasis", which was firstly invented by Bateson [1909], who described distortions of mendelian segregation ratios about one hundred years ago. Epistasis has become a hot topic in complex genetic disease in recent years. The existence of epistasis has been considered as an important contributor to generic variation in complex diseases. For complex traits such as diabetes and hypertension, the search for susceptibility loci has been less successful

because of the effect of some complicating factors such as the growing number of contributing loci and susceptibility alleles, environmental factors [Cordell [2002]]. A most likely reason for lack of success in genetic studies of complex diseases is the existence of interactions between locus.

From a statistical point of view, Friedman and Popescu [2008] illustrated the statistical definition of interaction effect. Suppose that we have a population model with some function $G(\cdot)$,

$$Y = G(\boldsymbol{x}) + \varepsilon,$$

where $\boldsymbol{x} = (X_1, X_2, \ldots, X_p)^T$ be the vector of input variables, Y be the response, and ε is the noise term. Any two numerical variables X_j and X_k have two-way interaction effect for y if

$$E_{\boldsymbol{x}} \left[\frac{\partial G(\boldsymbol{x})}{\partial X_j \partial X_k} \right]^2 > 0.$$

For categorical variables, the left part of the above formula is changed into finite differences. If no interaction effect exists between these two variables, the function $G(\boldsymbol{x})$ can be expressed as

$$G(\boldsymbol{x}) = G_1(\boldsymbol{x}_{-j}) + G_2(\boldsymbol{x}_{-k}),$$

where $G_1(\cdot)$ is not associated with X_j and $G_2(\cdot)$ is independent of X_k, $\boldsymbol{x}_{-j}$ and $\boldsymbol{x}_{-k}$ represent all variables excluding X_j and X_k, respectively. Similarly, if the function $G(\cdot)$ is said to have an higher-order($m > 2$) interaction among the numerical random variables $\boldsymbol{x}$ if

$$E_{\boldsymbol{x}} \left[\frac{\partial G(\boldsymbol{x})}{\partial X_{i_1} \cdots \partial X_{i_m}} \right]^2 > 0,$$

and an analogous expression involves finite differences for categorical variables, where $\{i_1 < i_2 < \cdots < i_m\}$ is one subset of $\{1, 2, \cdots, p\}$. If the higher-order interaction effect does not appear among these m variables, the function $G(\boldsymbol{x})$ can be expressed as

$$G(\boldsymbol{x}) = G_1(\boldsymbol{x}_{-i_1}) + \cdots + G_m(\boldsymbol{x}_{-i_m}),$$

where $G_j(\cdot)$ is not associated with X_{i_j} and $\boldsymbol{x}_{-i_j}$ represents all variables excluding X_{i_j}, $j = 1, \ldots, m$. Generally speaking, two-way interaction effect is more easily understood

and interpreted, and the higher-order interaction effect is more complex and confusing in practice, hence, our emphasis will be put on two-way interaction effect in the generalized linear model, that is,

$$E(Y|X=\boldsymbol{x}) = b'(\theta(\boldsymbol{x})) = g^{-1}\left(\beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_iX_j\right) \tag{1.1}$$

for some link function $g(\cdot)$, where X_iX_j is the interaction term, β_i is the coefficient of main effect and β_{ij} is the interaction's coefficient. And here, we focus on the canonical link function, hence $b' = g^{-1}$ and the canonical natural parameter $\theta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_iX_j$. The conditional distribution of the random variable Y given the predictor vector $\boldsymbol{x}$ belongs to an exponential family, whose probability density function has the canonical form

$$f_{Y|\boldsymbol{x}}(y|\boldsymbol{x}) = \exp\{y\theta(\boldsymbol{x}) - b(\theta(\boldsymbol{x})) + c(y)\} \tag{1.2}$$

where $b(\cdot)$ and $c(\cdot)$ are some known functions and $\theta(\boldsymbol{x})$ is a canonical natural parameter. Here we ignore the dispersion parameter ϕ in (1.2), since we only concentrate on the estimation of mean regression function. It is well known that the distributions in the exponential family include the Binomial, Gaussian, Gamma, Inverse-Gaussian, Poisson distributions.

Because of the importance of the interaction effect, a lot of researchers have developed various methods or algorithms to detect or to test the existence of the interaction terms in different fields, especially in statistics and genetics. For example, ANOVA F-test is always applied in the detection of interaction in Experimental Design [See Cox [1984]]. With the advent of the era of "Big Data", many other techniques such as variable selection have been developed by researchers over the past few decades. They are beneficial for us to filter out the awfully unimportant interaction terms.

1.1.1 Variable Selection and Feature Selection

For any statistical issue, statisticians are usually concerned about three important things: statistical accuracy, algorithm's efficiency, and model interpretability. In the traditional studies of statistics, the sample size n of data is much larger than the number of variables p. Subset selection is our best choice to achieve our expected results in the conventional problems, such as stepwise regression, forward regression, backward regression and "best" subset method. Based on some criteria, we are able to make the variable selection according to apply these methods.

Assume that the observed data $(\boldsymbol{x}_1^T, y_1)$, $(\boldsymbol{x}_2^T, y_2), \ldots, (\boldsymbol{x}_n^T, y_n)$ are independent copies of the population $(\boldsymbol{x}^T, Y)$, and they are randomly sampled from the linear model with dimensionality p

$$Y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon. \tag{1.3}$$

Denote that the response vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$ is a $n \times p$ data matrix. Now, we introduce some traditional criteria for predictor selection. Theil [1961] proposed the adjusted R^2. Denote that RSS_d is the residual sum of squares of the subset with d variables, where $d \leq p$, and SST is the total sum of squares. The adjusted R^2 is defined by

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-d}\frac{\mathrm{RSS}_d}{\mathrm{SST}}.$$

Obviously, we will choose the subset of variables whose R^2_{adj} is very close to 1. Later, Allen [1971] considered the mean square error of prediction (MSEP) as a criterion for selecting variables. Mallows [1973] provided another statistics C_p, which is given by $C_p = \mathrm{RSS}_d/s^2 + 2d - n$, where s^2 is the mean squared error of the full model. If some subset model satisfies that $C_p \approx p$, we may select this model as our best model. Another credible criterion is the PRESS statistic, that is, the prediction sum of squares. Allen [1974] gave its definition:

$$\mathrm{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2,$$

where $\hat{y}_{i,-i}$ is the predicted value of $E(y_i|\boldsymbol{x})$ except the ith observation. In other

words, the $n-1$ observations are used to fit the model and then predict the remaining observation $E(y_i|\boldsymbol{x})$. Actually, it is a form of cross validation. Usually, we choose the model with one subset of all variables that arrive at a minimum value of PRESS. AIC was founded by Akaike [1973, 1974] in information theory. Akaike [1973, 1974] suggested selecting one model that attains the minimum of the information loss. It can be measured by the Kullback-Leibler (KL) divergence of the fitted model from the true model. This results in the AIC value, which is defined as

$$\text{AIC} = 2 * l(\hat{\theta}) + 2 * d,$$

where $l(\cdot)$ is the negative log-likelihood function for the model with the parameter θ and d variables, $\hat{\theta}$ is the estimated value of θ that minimize the function $l(\cdot)$. The Bayesian information criterion (BIC) is similar to AIC, which was proposed by Schwarz [1978] for model selection. In this paper, he gave a Bayesian argument about it. The formula of BIC is

$$\text{BIC} = 2 * l(\hat{\theta}) + \log(n) * d,$$

where $l(\cdot)$ is defined same as before and n is the sample size. And also, we still choose one model with the smallest AIC or BIC. Nishii [1984] developed another criterion-generalized information criterion (GIC) in the normal linear model, that is,

$$\text{GIC} = n * \log(\hat{\sigma}^2) + a_n * d,$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2 in the model with d variables, a_n is a positive sequence such that $\lim_{n\to\infty} n^{-1}a_n = 0$ and $\lim_{n\to\infty} a_n = \infty$. If $a_n = \log(n)$, GIC is same as BIC in the normal linear model. And if a_n is taken as 2, GIC is equivalent to AIC. Zhang et al. [2010c] and Fan and Tang [2013] further discussed the properties of GIC and apply it to select tuning parameters or regularization parameters. Other classical methods certainly exist in the literature of past years. Here, we do not illustrate one by one.

When we take into account the large number of the covariates and the complex-

ity of model space, especially, the number of predictors is comparable to or much larger than the number of observations n, the traditional methods or criteria may not perform well in the large model space. Chen and Chen [2008] extended the BIC criterion to EBIC such that extremely beneficial for variable selection in the model with a moderate sample size n but with large dimensionality p. Denote that $\mathcal{S}$ is the whole model space and $\{\mathcal{S}_1, \ldots, \mathcal{S}_p\}$ is a partition of the whole model space, such that, models within $\mathcal{S}_j$ have a same amount of predictors. Let $N(\mathcal{S}_j)$ is the cardinality of each set $\mathcal{S}_j$ and $v(s)$ is the number of covariates in the model $s \in \mathcal{S}_j$. For any $s \in \mathcal{S}_j$ and $0 \leq \gamma \leq 1$, the extended BIC is expressed as

$$\text{EBIC}_\gamma = 2 * l(\hat{\theta}(s)) + \log(n) * v(s) + 2\gamma \log(N(\mathcal{S}_j)),$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ in the model s. It is obvious that BIC is a special case of EBIC.

Besides the above mentioned methods, to deal with the challenge of big data and obtain the more appropriate models to do some statistical inference, other variable selection techniques have been developed by statisticians over the past few decades. They are basically grouped into two categories: penalization or regularization methods and screening methods.

Penalization method is one important tool in the variable selection for high dimensional statistical modelling. It has been widely applied in statical inferences and other aspects such as machine learning and deep learning, and is a moderate scale learning technique. The general form of the penalization problem is

$$n^{-1} l_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.4}$$

where $n^{-1} l_n(\boldsymbol{\beta})$ is the loss function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is an unknown parameter vector, and $p_\lambda(\cdot)$ is a penalty function with tuning parameter $\lambda \geq 0$. By minimizing the loss function (1.4), we hope that we can simultaneously select the important variables and estimate the associated coefficients, and other unimportant variables can be deleted automatically. In most cases, $l_n(\boldsymbol{\beta})$ is denoted as the negative log-likelihood function. Consider the linear model (1.3), if ε follows the normal distribution with

mean 0 and variance σ^2, the penalized problem (1.4) is equivalent to the penalized least squares (PLS) problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $\|\cdot\|$ is the L_2-norm.

There are a mass of papers devoted to studying the penalized problems in the past decades. The bridge regression is a broad class of the penalty regression, in which, the penalty function is L_r penalty function, i.e., $p_\lambda(|\theta|) = \lambda|\theta|^r$ with $r > 0$. It was introduced by Frank and Friedman [1993] and includes two special cases: ridge (L_2 penalty) and LASSO (L_1 penalty). Hoerl and Kennard [1970] proposed the ridge regression by utilizing the L_2 penalty to achieve the better prediction in the linear regression model. Tibshirani [1996] discussed the least absolute shrinkage and selection operator (LASSO) in the ordinary regression model. Furthermore, if $p_\lambda(|\theta|) = I(\theta \neq 0)$, where $I(\cdot)$ is the indicator function, the penalized problem becomes best subset selection (L_0 penalty). It is well known that L_r penalty function can take the variable selection when $0 \leq r \leq 1$ and shrinkage the coefficients of variables in the model when $r > 1$. Fan and Li [2001] claimed that good penalty functions for model selection should generate the estimators with three properties: sparsity, unbiasedness, and continuity. For the L_r penalty, if $r > 1$, they are convex but do not satisfy the sparsity condition; if $0 \leq r < 1$, they are strictly non-convex and the continuity of the estimators is not satisfied; whereas L_1 penalty function does not meet the unbiasedness condition. As a result, Fan [1997] and Fan and Li [2001] advocated the smoothly clipped absolute deviation (SCAD) and its derivative is defined as

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\},$$

for some $t > 0$ and $a > 0$, where $p_\lambda(0) = 0$. By a Bayesian argument, $a = 3.7$ is used by default. They showed that SCAD satisfies the above three properties. Fan and Li [2001] not only proposed the SCAD penalty function but also studied the oracle property of nonconcave penalized likelihood estimators in the general penalization problem with finite dimensionality. Fan and Peng [2004] extended this result to more

general case that the model has diverging number p_n of covariates with sample size n, in other words, $\lim_{n\to\infty} p_n = \infty$ and $\lim_{n\to\infty} n^{-1}p_n = 0$. Based on the same consideration of the above three properties, the minimax concave penalty (MCP) was developed by Zhang [2010], whose derivative is given by

$$p'_\lambda(t) = \frac{(a\lambda - t)_+}{a}$$

for some $t > 0$ and $a > 0$. Nikolova [2000] discussed the transformed L_1 penalty for the regularization estimator i.e.,

$$p_\lambda(t) = \frac{\lambda a t}{1 + at}$$

for some $t \geq 0$ and $a > 0$. Zou and Hastie [2005] proposed the elastic net penalty for encouraging a grouping effect. In fact, the elastic net is a linear mixture of L_1 and L_2 penalty, that is,

$$p_\lambda(t) = \lambda[(1-\alpha)t + \alpha t^2],$$

where, $0 \leq \alpha \leq 1$ and $t > 0$. Friedman [2012] proposed the generalize elastic net family , that is,

$$p_\lambda(t) = \lambda((a-1)t^2/2 + (2-a)t), \quad 1 \leq a \leq 2.$$

An extension of this family to non-convex members is

$$p_\lambda(t) = \lambda(\log((1-a)t + a), \quad 0 < a < 1.$$

Later, the log-penalty was called by Mazumder et al. [2011], which is a generalization of the elastic net family to cover the nonconvex penalties from LASSO down to the best subset. Its definition is given by

$$p_\lambda(t) = \frac{\lambda}{\log(a+1)} \log(at+1), \quad a > 0.$$

For model selection and sparse recovery, Lv and Fan [2009] studied the smooth inte-

gration of counting and absolute deviation (SICA) penalties, which is defined by

$$p_\lambda(t) = \lambda\rho_a(t) = \lambda\left\{\frac{(a+1)t}{a+t}\right\} = \lambda\left\{\left(\frac{t}{a+t}\right)I(t \neq 0) + \left(\frac{a}{a+t}\right)t\right\}, \quad \text{for } a,\ t \geq 0.$$

Another one penalty called exponential-type penalty (ETP) was shown in Gao et al. [2011], whose derivative is expressed as

$$p'_\lambda(t) = \frac{a\lambda}{1-\exp(-a)}\exp(-at), \quad \text{for } a,\ t > 0.$$

So far, we have reviewed many different kinds of penalization methods with various penalty functions for variable selection or model selection in high dimensional model space. However, penalization may not perform well in ultra-high dimensional. For instance, in genetics, the genome-wide association studies (GWAS) focus on the associations among the single-nucleotide polymorphisms (SNPs) in the complex diseases. The number of SNPs, p is usually on the order of tens of millions while the number of the observations n is much smaller than dimension p, that is on the order of tens ([See, e.g. Klein et al., 2005; Wan et al., 2010a,b]). A natural idea is to dwindle the dimensionality p from a huge scale p ($\log p = O(n^{\xi_1})$, for some $\xi_1 > 0$) to a moderate scale d ($d = O(n^{\xi_2})$, for some $\xi_2 > 0$) by a fast and efficient method so that penalization method can be utilized to the reduced space. Fan and Lv [2008] published one brilliant paper and proposed one efficient method-sure independent screening (SIS) for variable selection in ultra-high dimensional feature space and simulated plenty of further researches. By considering the marginal correlations, SIS ranks the features and the selected variables include all important covariates in the reduced model with probability tending to 1.

More precisely, consider the linear regression model (1.3), and assume that the random design matrix $\boldsymbol{X}$ is standardized column-wise with mean 0 and variance 1. Denote that

$$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_p)^T = \boldsymbol{X}^T\boldsymbol{y}$$

is a p-vector about Pearson correlation obtained by the marginal regression. For any

given d, which can be taken as $n-1$ or $n/\log(n)$,

$$\widehat{M}_d = \{1 \le j \le p : |\omega_j| \text{ is among the first } d \text{ largest of all}\}.$$

Hence, SIS is able to reduce the original model with size p to the submodel with size d. However, SIS may not work when some important covariate X_j is jointly correlated but marginally uncorrelated with the response Y. To deal with this problem, Fan and Lv [2008] and Fan et al. [2009] proposed another procedure-iterative sure independent screening (ISIS) based on the conditional regression. Sometimes, the marginal nonlinearity exists between the predictor and the response, Hall and Miller [2009] proposed the generalized correlation between the j-th explanatory variable X_j and the response Y to measure the marginal linear dependence, whose formula is

$$\rho_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_j), Y\}}{\sqrt{\text{var}\{h(X_j)\}\text{var}(Y)}}, \quad 1 \le j \le p,$$

where $\mathcal{H}$ is a vector space of functions including all of linear functions. Li et al. [2012a] considered an alternative robust way to measure the marginal relationship between the predictor X_j and the response Y in the linear model and transformation linear model, and proposed one robust rank correlation screening (RRCS) method for ultra-high dimensional data. The sample marginal rank correlation is defined as

$$\omega_j = \frac{1}{n(n-1)} \sum_{i \ne l}^{n} I(X_{ij} > X_{lj}) I(Y_i > Y_l) - \frac{1}{4}.$$

The advantage of RRCS is that it is not only robust and invariant under the monotonic transformation but also save computational cost. To overcome the same problem mentioned by Fan and Lv [2008], iterative robust rank correlation screening (IRRCS) is also suggested by Li et al. [2012a]. For a class of generalized linear models, Fan and Song [2010] took the maximum marginal likelihood estimator (MMLE) or the marginal log-likelihood function's increment as a measurement of marginal correlation between X_j and Y. Assume that the conditional distribution of Y has the density function (1.2), $l(Y, \theta) = -\theta Y + b(\theta) - c(Y)$ and $\mathbb{P}_n h(X, Y) = n^{-1} \sum_{i=1}^{n} h(X_i, Y_i)$ is the empirical measure. Here, $\theta = \beta_0 + \sum_{j=1}^{n} X_j \beta_j$. The sample MMLE and increment

of negative log-likelihood function respectively are

$$\widehat{\boldsymbol{\beta}}^M = (\widehat{\beta}_{j0}^M, \widehat{\beta}_j^M) = \arg\min_{\beta_0,\beta_j} \mathbb{P}_n l(\beta_0 + \beta_j, Y),$$

and

$$L_{j,n} = \mathbb{P}_n\{l(\widehat{\beta}^M, Y) - l(\widehat{\beta}_{j0}^M + \widehat{\beta}_j^M, Y)\}.$$

Hence, for the given pre-specified threshold value γ_n and v_n, the index set of selected variables are

$$\widehat{M}_{\gamma_n} = \{1 \le j \le p : |\beta_j^M| > \gamma_n\},$$

and

$$\widehat{M}_{v_n} = \{1 \le j \le p : L_{j,n} > v_n\}.$$

Besides Pearson correlation, generalized correlation and rank correlation , another correlation-distance correlation (DC) can be used by Li et al. [2012b] to screen the variables, whose method is called DC-SIS. The distance covariance (dCov) was firstly introduced by Székely et al. [2007] to measure the association between two random vectors $\boldsymbol{U} \in \mathbb{R}^{d_1}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2}$. By using the marginal characteristic functions $\phi_{\boldsymbol{U}}(\boldsymbol{t})$ and $\phi_{\boldsymbol{V}}(\boldsymbol{s})$ and the joint characteristic function $\phi_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{t},\boldsymbol{s})$, its expression is

$$\mathrm{dCov}^2(\boldsymbol{U},\boldsymbol{V}) = \int_{\mathbb{R}^{d_1+d_2}} |\phi_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{t},\boldsymbol{s}) - \phi_{\boldsymbol{U}}(\boldsymbol{t})\phi_{\boldsymbol{V}}(\boldsymbol{s})|^2 \omega(\boldsymbol{t},\boldsymbol{s}) d\boldsymbol{t} d\boldsymbol{s},$$

where, $\omega(\boldsymbol{t},\boldsymbol{s}) = (c_{d_1} c_{d_2} |\boldsymbol{t}|_{d_1}^{1+d_1} |\boldsymbol{s}|_{d_2}^{1+d_2})^{-1}$ with $c_d = \pi^{(1+d)/2}/\Gamma((1+d)/2)$ and $|f|^2$ stands for $f \cdot \bar{f}$ with complex conjugate $\bar{f}$ of f. Hence , the distance correlation (DC) between $\boldsymbol{U}$ and $\boldsymbol{V}$ is expressed as

$$\mathrm{dcorr}(\boldsymbol{U},\boldsymbol{V}) = \frac{\mathrm{dCov}^2(\boldsymbol{U},\boldsymbol{V})}{\sqrt{\mathrm{dCov}^2(\boldsymbol{U},\boldsymbol{U})\mathrm{dCov}^2(\boldsymbol{V},\boldsymbol{V})}}.$$

Fan et al. [2011] continued to discuss the variable screening issue and proposed a nonparametric independence screening (NIS) for additive models, i.e,

$$Y = \sum_{j=1}^{n} m(X_j) + \varepsilon,$$

where $m_j(X_j)$ is an unknown smooth function with mean 0. NIS is a special case of SIS. The idea is to utilize $E(f_j^2(X_j))$ to represent the importance of the predictors, where $f_j(X_j)$ is the projection of Y onto X_j and is one nonparametric estimator of $m_j(X_j)$ by using a normalized B-spline basis $\mathbf{B}_j = \{B_{j_1}(x), \ldots, B_{jd_n}(x)\}$, that is, $f_j(x) = \sum_{k=1}^{n} \beta_{jk} B_{jk}(x)$. And the coefficients of function $f_j(X_j)$ can be estimated by marginal regression problem

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d_n}}{\arg\min}\, \mathbb{P}_n (Y - \boldsymbol{\beta}_j^T \mathbf{B}_j)^2.$$

And then, define that the sample version of $E(f_j^2(X_j))$ is $\|\hat{f}_j\|_n^2 = n^{-1} \sum_{i=1}^{n} \hat{f}_j(X_{ij})^2$, where $\hat{f}_j(X_{ij}) = \sum_{k=1}^{n} \widehat{\beta}_{jk} B_{jk}(x)$. Fan et al. [2014b] extended NIS to the varying coefficient models (VCM),

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

where $E(\varepsilon | \boldsymbol{X}, u) = 0$, $\beta_0(u)$ is the unknown smooth intercept function, and $\beta_j(u)$ is the unknown smooth univariate function with variable u. Another sure screening procedure-conditional correlation screening (CC-SIS) was proposed by Liu et al. [2014] for ultra-high dimensional varying coefficient model (VCM) and the conditional correlation between X_j and Y given u is defined as

$$\rho(X_j, Y | u) = \frac{\mathrm{cov}(X_j, Y | u)}{\sqrt{\mathrm{var}(X_j | u) \mathrm{var}(Y | u)}}.$$

In order to recover the hidden variables, that have impact on the response but cannot be selected by simple SIS, Barut et al. [2016] provided the conditional sure independent screening (CSIS) and applied the conditional MMLE to measure the significance of the predictor X_j. Suppose that $\boldsymbol{X}_{\mathcal{C}} = (X_1, \ldots, X_p)^T$ be the first known important variables, X_j is one of the remaining variables, the conditional MMLE is defined as

$$\widehat{\boldsymbol{\beta}}_{\mathcal{C}j} = \underset{\boldsymbol{\beta}_{\mathcal{C}}, \beta_j}{\arg\min}\, \mathbb{P}_n l(\boldsymbol{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}} + X_j \beta_j, Y).$$

For a given predefined threshold γ_n, the index set of the selected variables is $\mathcal{M}_{\gamma_n} =$

$\{j : |\beta_j| > \gamma_n\}$. Wang and Leng [2015] proposed a new screening technique called high dimensional ordinary least squares projection (HOLP) for variables selection in the linear model with large dimensionality p and small sample size n ($p > n$), to overcome the problem that some important variables have weak marginal correlations with the response. The idea of HOLP is to compute the estimator $\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}Y$ in the linear model (1.3), rank the each element of $\widehat{\boldsymbol{\beta}}$ and select the largest d of all $|\beta_j|$s or coefficients that exceed the pre-specified threshold. This procedure is very simple but effective for variable selection in the ultra-high dimensional linear model. Only one disadvantage is that the HOLP may not be extended to more general models such as the generalized linear model.

1.1.2 Interaction Effect Selection

As mentioned in the last subsection, statistical accuracy and model interpretability are awfully essential in the data analysis. With the development of the science and technology, data scientists face to the data with tremendous dimensionality. The simple models may not be able to meet the requirements of data research. Therefore, in order to reinforce the models' accuracy and interpretability, the importance is attached to a little more complex models, such as the linear model with explanatory variables including main effects and "interaction effects" or "cross-term" simultaneously, just like

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{1 \le j < k \le p} \beta_{jk} X_j X_k + \varepsilon \tag{1.5}$$

or the well-known quadratic model:

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

In the high or ultra-high dimensional feature space, the dimensionality p is tremendous and grows faster than sample size n with the speed of exponential order. If two-way interaction terms are taken into account, the length of the predictor vector will be $p + p(p-1)/2$. The number of covariates is more immense. For the sake of reducing the computational cost and predicting accurately the observations' behavior in the future, interaction selection or epistasis detection becomes a more and more

hot issue and lots of literature about this topic has sprung up in the recent few years.

In the last century, most of the statisticians believed that one interaction term is allowed into the model (1.5) if at least one of the corresponding main effects are also in the model (1.5). Such condition is known as several names. Nelder [1977] gave the name "marginality assumption ", Peixoto [1987, 1990] called it as "well-formulated models", the name "heredity assumption" was proposed by Chipman [1996]. Recently, "hierarchy assumption" was provided by Bien et al. [2013]. Here, the name "heredity" is utilized in this thesis. Two types of this restriction are included, which are called "strong heredity" and "weak heredity". Strong heredity means that an interaction between two predictors is significant only if both the corresponding main effects are included in the model, while at least one of two main effects is marginally significant if their interaction is present in the model under the weak heredity assumption. For instance, in the linear model (1.5), they imply respectively

$$\begin{aligned} &\text{Strong heredity}: \ \beta_{jk} \neq 0 \Longrightarrow \beta_j \neq 0 \text{ and } \beta_k \neq 0; \\ &\text{Weak heredity}: \ \beta_{jk} \neq 0 \Longrightarrow \beta_j^2 + \beta_k^2 \neq 0. \end{aligned}$$

However, the pure interaction exists. The heredity assumption is natural in some applications, but it will cause the pure interactions missing. In Genetics, Cordell [2002] published a review paper about detecting gene-gene interactions that underlie human diseases and indicated that pure interactions in the absence of main effects will be missed under the many existing methods. It stimulates us to look for one method that is available to pursue the interaction terms without the heredity assumption.

Interaction effects have played a critical role in many research areas. Genetics is taken as an example. In Genetics, the goal is to search for genetic factors that have impacts on common complex traits or diseases for modern geneticists. The scientific importance of gene-gene interactions has been demonstrated by many investigators and the existence of interactions between loci has been confirmed by them. Cox et al. [1999] illustrated that the interaction of genes on chromosomes 2 (NIDDM1) and 15 (near CYP19) can increase susceptibility to type 2 diabetes in Mexican Americans. Dong et al. [2003] found that extreme human obesity can be influenced by the epistatic interactions between loci on Chromosomes 20 and 10. In the light of the significance of

epistasis interaction in genetic studies, many approaches or algorithms were developed and applied to identify epistatic interaction during the past few years. Ritchie et al. [2001] proposed the multifactor dimensionality reduction (MDR) to detect two-way or high-order interaction among two or more loci in relatively small samples, and the pure epistatic interactions can also be identified by this method.

As the success and popularity of a powerful tool−genome-wide association study (GWAS) grows, many efficient methods are invented in GWAS. The genome-wide association studies (GWAS) focuses on the association among single nucleotide polymorphisms (SNPs) and provides the useful clues to help researchers understand the basis of the complex diseases. All SNPs are considered as the predictors in GWAS model and the alleles from locus A are denoted as "A" and "a", therefore one SNP has three genotypes: "AA", "Aa" and "aa". Besides a series of covariates, the remaining predictors are categorical variables with three levels. From a test point of view, three algorithms FastANOVA, FastChi and Convex Optimization-based Epistasis detection algorithm (COE) were proposed by Zhang et al. [2008], Zhang et al. [2009] and Zhang et al. [2010b], respectively. They used an upper bound to exclude a large number of SNP pairs but each SNP has been divided into a binary variable. Purcell et al. [2007] developed a software PLINK that can detect epistatic interactions. Zhang and Liu [2007] introduced a method called bayesian epistasis association mapping (BEAM) based on Markov chain Monte Carlo (MCMC). SNPHarvester (Yang et al. [2009]) and SNPRuler (Wan et al. [2010b]) are the filtering-based methods based on the χ^2 test and usually their first step is to reduce the number of SNPs by using some filtering methods. Two locus epistatic interactions were also considered by Zhang et al. [2010a]. They proposed an exhaustive algorithm−Tree-based epistasis association mapping (TEAM). Yang et al. [2010] applied the adaptive group Lasso with sparsity constraint to GWA studies with large-scale data and select main effects and epistasis simultaneously. Screen and Clean (SC) was proposed by Wu et al. [2010]. They first screen the main effect by logistic regression with LASSO penalty and then detect the interactions by controlling Type I error in the selected main effects. BOOST (Wan et al. [2010a]) contributes to the detection of epistatic interaction by likelihood ratio test. It transforms the original data to Boolean type and then uses the

logic operations to construct the contingency table. It is very efficient. And also, pure epistatic can be detected by BOOST. Li et al. [2014] proposed two-stage sure independence screening (TS-SIS) procedure for detecting gene-gene interactions in GWAS and claimed that TS-SIS is computationally efficient and has an outstanding finite performance. However, these methods only focus on genome-wide association studies, and cannot be applied to some other fields or more general models. And most methods are built on the heredity assumption. Especially, some methods cannot be applied to the quantitative traits unless the predictor and response value are properly discretized.

In statistics, the importance is also attached to the interaction effects and the growing literature is developed to identify them in the different models. Most of the existing methods, however, depend on the strong or weak heredity assumption. Choi et al. [2010] extended the LASSO method and identified the significant interaction terms in the linear model (1.5) and generalized linear models (1.1) under the strong heredity assumption. They proved that their method possessed the oracle property as mentioned in Fan and Li [2001] and Fan and Peng [2004], that is, it performs well if the true model is known in advance. The algorithm hierNet was developed by Bien et al. [2013] to select the interactions, which added a set of convex constraints to LASSO in the linear model (1.5) and constructed the sparsity interaction model with the strong or weak heredity assumption. For the linear model (1.5), Hao and Zhang [2014] also proposed two algorithms iFORT and iFORM and identified the interaction effects in a greedy fashion under the heredity assumption. Hao et al. [2016] went on to study the interaction selection problem for high dimensional quadratic regression via regularization and proposed a new regularization method, called Regularization Algorithm under Marginality Principle (RAMP), which means that RAMP still works under the heredity assumption. To pursue the interaction without any heredity assumption, Fan et al. [2016] suggested a flexible procedure, called the interaction pursuit (IP), in ultra-high dimensional linear interaction models (1.5). The idea of the method IP is to select the active interaction variables. A predictor X_j is called an active interaction variable if there exists another variable X_k $(k \neq j)$ such that $\beta_{jk} \neq 0$. And they utilize the Pearson correlation between X_j^2 and Y^2 $(\mathrm{Corr}(X_j^2, Y^2))$

to identify the active interaction variables. Kong et al. [2016] extended IP to the ultra-high dimensional linear interaction model with multiple responses:

$$\boldsymbol{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_M \boldsymbol{X} + \boldsymbol{\beta}_I \boldsymbol{Z} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^T$ is a $d \times 1$ vector of responses, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is a $p \times 1$ vector of main effects, $\boldsymbol{Z} = (X_1X_2, X_1X_3, \ldots, X_{p-1}X_p)^q$ is a $q \times 1$ vector of interactions with $q = p(p-1)/2$, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0d})^T$ is a d-dimensional vector of intercepts, $\boldsymbol{\beta}_M \in \mathbb{R}^{d \times p}$ and $\boldsymbol{\beta}_I \in \mathbb{R}^{d \times q}$ are coefficient matrices of main effects and cross-terms, respectively. They replaced the Pearson correlation ($\text{Corr}(X_j^2, Y^2)$) with distance correlation (DC) ($\text{dcorr}(X_j^2, \boldsymbol{Y}^2)$) provided by Székely et al. [2007] and proposed the new method IPDC to select the important active interaction variables in the screening step. However, the power of $\text{Corr}(X_j^2, Y^2)$ in IP and the power of $\text{dcorr}(X_j^2, \boldsymbol{Y}^2)$ in IPDC depend on the distributions of X_j. If the support set of X_j is $\{-1, 1\}$, these two methods are not beneficial for us to identify the cross terms. A new algorithm xyz was introduced by Thanei et al. [2016]. Suppose that $\boldsymbol{X} \in \{-1, 1\}^{n \times p}$ be a matrix of binary predictors and $\boldsymbol{Y} \in \{-1, 1\}^n$ be the response vector with binary values in this paper. Define that $\boldsymbol{Z}$ is a $n \times p$ matrix with $Z_{ij} = Y_i X_{ij}$. This paper uses a random projection vector $\boldsymbol{R} \in \mathbb{R}^n$ such that $\mathbf{x} = \boldsymbol{X}^T \boldsymbol{R} = (x_1, \ldots, x_p)^T$ and $\mathbf{z} = \boldsymbol{Z}^T \boldsymbol{R} = (z_1, \ldots, z_p)^T$. That is, it projects the matrixes $\boldsymbol{X}$ and $\boldsymbol{Z}$ to two $p \times 1$ vectors $\mathbf{x}$ and $\mathbf{z}$, respectively, and then finds a set of all pairs (j, k) such that $x_j = z_k$. In the final step, in this remaining set, this method selects the pairs that satisfy $|\boldsymbol{X}_j^T \boldsymbol{Z}_k|/n > \gamma$, where γ is a specified threshold. Thanei et al. [2016] claimed that by repeating L times this process with different random projections R, this algorithm can find all the potential candidate sets for interaction terms with high probability. And xyz algorithm is also extended to the continuous case by transforming $\boldsymbol{X}$ to a binary matrix.

1.2 Hypothesis Testing of Markov Processes

The Markov property was firstly proposed by Markov [1954], which represents one kind of memoryless property of a stochastic process. A stochastic process is actually

a collection of random variables $\boldsymbol{X} = \{X_t,\ t \in \mathbf{T}\}$, where $\mathbf{T}$ is an ordered index set. For each $t \in \mathbf{T}$, X_t is a random variable from one measurable space−sample space $(\Omega, \mathcal{F})$ to another one measurable space−state space $(E, \mathcal{G})$. The state space E will usually be $\mathbb{R}^d$ and $\mathcal{G}$ is the corresponding σ-algebra of Borel sets. A Markov process is a stochastic process that has no memory of the information in the past state, which indicates that given the present state, the future states and the past states are statistically conditional independent in this process.

Without loss of generality, assume that $\mathbf{T} = [0, \infty)$ and a stochastic process $\boldsymbol{X} = \{X_t,\ t \in \mathbf{T}\}$ is defined at a probability space $(\Omega, \mathcal{F}, P)$. Denote that $\mathcal{F}_i^j$ is the sub $\sigma-$field of the $\sigma-$field $\mathcal{F}$, generated by $\{X_t : i \leq t \leq j\}$, i.e., $\mathcal{F}_i^j = \sigma(X_t : i \leq t \leq j)$. Thus, $\mathcal{F}_0^t = \sigma(X_s : s \leq t)$ represents the history of $\boldsymbol{X}$ while $\mathcal{F}_t^\infty = \sigma(X_s : s \geq t)$ contains the future evolution of $\boldsymbol{X}$. Suppose that there exists a filtration $(\mathcal{F}_t)_{t\in\mathbf{T}}$ such that $\mathcal{F}_0^t \subset \mathcal{F}_t$, for all $t \in \mathbf{T}$. And then, the stochastic process $(X_t)_{t\in\mathbf{T}}$ is called a Markov process if the process meets

$$P(A \in \mathcal{F}_t^\infty | \mathcal{F}_t) = P(A \in \mathcal{F}_t^\infty | X_t), \quad \text{for all } t \in \mathbf{T}. \tag{1.6}$$

If the index set $\mathbf{T} = \mathbb{N}$, the definition (1.6) can be reformulated as

$$P(X_n = x_n | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}). \tag{1.7}$$

In this case, the process is a discrete time Markov process. From the point of conditional independence, the formula (1.6) is definitely equivalent to

$$P(A \cap B | X_t) = P(A | X_t) P(B | X_t), \quad A \in \mathcal{F}_t^\infty \text{ and } B \in \mathcal{F}_t. \tag{1.8}$$

Furthermore, denote that the transition function $F(A, t|x, s)$ is the conditional probability of $X_t \in A$ given $X_s = x$ for $t \geq s$, $x \in E$ and $A \in \mathcal{B}(E)$, that is, $F(A, t|x, s) = P(X_t \in A | X_s = x)$. Under the Markov property, this function satisfies the famous Chapman-Kolmogorov equation, which was proposed independently by

two mathematicians Sydney Chapman and Andrey Kolmogorov, that is,

$$F(A,t|x,s) = \int_E F(A,t|y,u)F(dy,u|x,s) \tag{1.9}$$

for all $x \in E$ and $s,t,u \in \mathbf{T}$ with $s \leq u \leq t$. If the Markov process $\boldsymbol{X}$ is time-homogenous, i.e., $F(A,t|x,s) = F(A,t-s|x,0)$, then the Chapman-Kolmogorov equation (1.9) is simplified to

$$F(A,t+s|x,0) = \int_E F(A,t|z,0)F(dz,s|x,0). \tag{1.10}$$

1.2.1 Markov Property in Various Models

The Markov property is a fundamental characterization and widely utilized in different kinds of models or processes of various areas such as time series, economics and finance. One of the processes is the Markovian decision processes (MDPs), which was early derived by Bellman [1957]. It relies on the Markov assumption, and provides a general framework for modeling sequential decision making. Lucas Jr and Prescott [1971] applied the MDP to the investment under uncertainty. And later, MDP was used by Lucas [1978] to analyze the stochastic behaviour of the asset pricing from the theoretical viewpoint. Weintraub et al. [2008] analyzed the Markov perfect industry dynamics for industrial organization and explored profit opportunities in a market.

The Markovian assumption is also very prevalent in another one kind of models-continuous time modeling. All diffusion processes are Markovian processes with the continuous sample path, which are regarded as the solutions of some stochastic differential equations (SDEs). For instance, the Onrstein-Uhlenbeck process or Vasicek process is the unique solution to the following stochastic differential equation

$$dX_t = r(\theta - X_t)dt + \sigma dW_t, \quad t \geq 0, \quad X_0 = x_0,$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion, $\sigma > 0$ is the diffusion or volatility coefficient and $r(\theta - X_t)$ is called the drift coefficient of the Vasicek process. And the

jump diffusion processes are also the Markov processes, they can be expressed as

$$dX_t = \mu(X_{t-})dt + \sigma(X_{t-})dW_t + J_{t-}dN_t,$$

where W_t is a standard Wiener process, N_t is a Poisson process with the rate λ and jump size 1, and the random variable J_{t-} is the jump size.

In Time Series, the Markov property is also an indispensable element. For example, the first order autoregressive model (AR(1)) satisfies the Markov property, that is,

$$X_t = \phi X_{t-1} + \varepsilon_t,$$

where ϕ is the unknown parameter, $\{\varepsilon_t\}_{t\geq 0}$ are independent and identically distributed (i.i.d.) and follow the standard normal distribution. Another one is the autoregressive conditional heteroscedastic (ARCH(p)) model, which was introduced by Engle [1982] to describe time-varying variability of inflation rates and can be formulated as

$$h_t = \phi_0 + \phi_1 X_{t-1}^2 + \cdots + \phi_p X_{t-p}^2 \quad \text{and} \quad X_t = h_t^{1/2}\eta_t,$$

with $\phi_0 > 0$, $\phi_i \geq 0$ for $i = 1, \ldots, p-1$ and $\phi_p > 0$, where $\{\eta_t\}_{t\geq 0}$ are i.i.d. random variables with mean 0 and variance 1. Apparently, ARCH(1) is a Markovian process.

1.2.2 Testing Markov property

Markov property plays an awfully important role in data analysis, but few existing papers indicate the tests for the Markov property in the past few years. Aitï-Sahalia [1996] firstly doubted whether the Markovian assumption is reasonable and proposed a test for whether the interest rates really follow the continuous-time Markov Diffusions. Later, Aït-Sahalia et al. [2010] went on to illustrate several statistics to test the Markov hypothesis for $\beta-$mixing stationary processes sampled at the regular time points. All of these two tests are based on the Chapman-Kolmogorov equation of

X_{t+1}, X_t andX_{t-1}, namely,

$$f(X_{t+1}|X_{t-1}) = \int f(X_{t+1}|X_t = x)f(X_t|X_{t-1})dx, \quad \text{for all } t \geq 1,$$

where $f(\cdot|\cdot)$ is the transition density functions estimated by the smoothed nonparametric kernel method. Although the Chapman-Kolmogorov equation (1.9) is one important property of Markov processes, it is only a necessary condition but not sufficient condition of the Markov property. Feller [1959] and Rosenblatt and Slepian [1962] indicated that the first order transition probabilities of some stochastic processes satisfy the Chapman-Kolmogorov equation (1.9) but they are not Markovian processes. As a result, non-Markovian processes may be missed by these tests.

De Matos and Fernandes [2007] found that financial transactions data are not evenly spaced in time and the test of Aitï-Sahalia [1996] is not available for this type of data. Thus, they developed another one nonparametric test for the Markov property with high frequency data and checked whether discrete-valued irregularly spaced financial transactions data followed the Markovian assumption. Because the Markov property (1.6) is equivalent to the conditional independence (1.7), the idea of this test is built on testing the conditional independence between X_{t+1} and X_{t-j} given X_t, that is,

$$f(X_{t+1}|X_t) = f(X_{t+1}|X_t, X_{t-j}) \quad \text{for all } t, j \geq 1.$$

In fact, for their simulated data sets, they took the case $j = 1$ and detected the following formula

$$f(X_{t+1}|X_t, X_{t-1}) = f(X_{t+1}|X_t)f(X_t, X_{t-1}) \quad \text{for all } t \geq 1. \tag{1.11}$$

Obviously, the Markov property can infer the formula (1.11), but the converse is not true. Actually, the equation (1.11) is also the necessary condition of the Markov property (1.6). The main disadvantage of these tests is that only one lag order is concerned and a majority of information may be lost in the process of testing.

Chen and Hong [2010] proposed one nonparametric regression-based goodness-of-fit test for multifactor continuous-time Markov models using the conditional characteristic function (CCF). In order to improve the existing tests, Chen and Hong [2012] provided another one nonparametric test for Markov property in time series by checking

$$\phi(u|X_t) = \phi(u|\mathcal{F}_t) \ a.s. \text{ for all } u \in \mathbb{R}^d \text{ and all } t \geq 1,$$

where $\phi(u|\cdot)$ is the conditional characterization function, namely,

$$\phi(u|\cdot) = \int e^{iu^T x} f(u|\cdot)dx$$

with $i = \sqrt{-1}$. Define that a complex-valued process $Z_{t+1}(u) = e^{iu^T X_{t+1}} - \phi(u|X_t)$. As one characterization of Markov property for one stochastic process is that a generalized residual process associated with the conditional characteristic (CCF) function is a martingale difference sequence (MDS), the Markov property is equivalent to

$$E(Z_{t+1}(u)|\mathcal{F}_t) = 0, \quad \text{for all } u \in \mathbb{R}^d \text{ and all } t \geq 1. \tag{1.12}$$

They viewed the process $\{Z_t(u)\}$ as a residual of the regression

$$e^{iu^T X_{t+1}} = E(e^{iu^T X_{t+1}}|X_t) + Z_{t+1}(u) = \phi(u|X_t) + Z_{t+1}(u)$$

and then defined the generalized covariance function

$$\Gamma_j(u, v) = \text{cov}[Z_t(u), e^{iu^T X_{t-|j|}}], \qquad u, v \in \mathbb{R}^d.$$

The formula (1.12) implies $\Gamma_j(u, v) = 0$, for all $u, v \in \mathbb{R}^d$ and all $j \neq 0$. They used this condition to construct the test statistics. They claimed that their test was available to discrete and continuous time processes with discretely observed data, and both univariate and multivariate time series processes, but there exists the drawback of the test statistics, that is, there is a gap between $E(Z_{t+1}(u)|\mathcal{F}_t) = 0$ and the final hypothesis $\Gamma_j(u, v) = 0$, which means that the final hypothesis is only one necessary condition for the formula (1.12).

1.3 Outline of the Thesis

In Chapter 2, we consider the interaction screening problem for ultra-high dimensional generalized linear models, and two-way interaction terms are considered in the models. The most essential thing, which is different from most of the existing approaches, is that our methods or algorithms do not depend on the strong or weak heredity assumption. And our proposed method can do the exhausting searching efficiently for all interactions terms in our models. We show that this method can identify the significant interactions with high probability. Simulation studies are carried out to detect the finite sample performance of this method by comparing it with other existing methods such as RAMP, xyz and IP. And also, our method is applied to three popular data sets: Prostate Cancer Data, Leukemia Data and Supermarket Data for illustration.

The testing about Markov property is another one important issue in the complex models. A few of papers focus on this important topic. In Chapter 3, we consider this hypothesis testing in the stationary times series models. The observations are β-mixing, and we take advantage of the equivalence between the Markov property and the conditional independence property and construct a new test statistic by the conditional characteristic functions. The asymptotic properties of the proposed test are studied under the null hypothesis and local alternative hypothesis. And also, numerical studies are implemented to assess the finite sample performance of our provided test procedures.

In Chapter 4, we discuss these results about these two topics-variable selection and hypothesis testing of Markovian assumption, which are present in this thesis. We provide some extensions about our methods and propose some interesting issues for future works.

Chapter 2

Sure Screening for Interaction Effect in Ultra-high Dimensional Generalized Linear Model

2.1 Introduction

As mentioned in Section 1.1.2, there is a huge literature devoted to studying the statistical properties and computational algorithms of screening interaction terms or epistatic interaction selection, such as BOOST (Wan et al. [2010a]), hierNet (Bien et al. [2013]), RAMP (Hao et al. [2016]) and xyz (Thanei et al. [2016]). In the existing literature, however, most of the methods or algorithms are under the heredity assumption.

Assume that given the predictor vector $\boldsymbol{x}$, the conditional distribution of the random variable Y belongs to an exponential family, whose probability density function has the canonical form

$$f_{Y|\boldsymbol{x}}(y|\boldsymbol{x}) = \exp\{y\theta(\boldsymbol{x}) - b(\theta(\boldsymbol{x})) + c(y)\} \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot)$ are some known functions and $\theta(\boldsymbol{x})$ is a canonical natural parameter. Here we ignore the dispersion parameter ϕ in (2.1), since we only concentrate on the estimation of mean regression function.

We consider the following generalized linear model with two-way interaction:

$$E(Y|X=\boldsymbol{x}) = b'(\theta(\boldsymbol{x})) = g^{-1}(\beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_iX_j) \tag{2.2}$$

for some link function $g(\cdot)$. And we focus on the canonical link function, hence $b' = g^{-1}$ and $\theta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_iX_j$. And also, we assume that each variable has been standardized with mean 0 and variance 1.

In the ultrahigh-dimensional regression model, we usually assume that the sparsity exists for variables or features, which means that only a few of variables or features are significantly correlated with response Y. As a result, our assumption is that only a small number of variables and their interactions contribute to the response Y, i.e., the true parameter vector $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star, \beta_{12}^\star, \beta_{13}^\star, \ldots, \beta_{(p-1)p}^\star)^T$ is sparse. Let

$$\mathcal{M}_\star = \{1 \le k \le p : \beta_k^\star \neq 0\}$$

and

$$\mathcal{N}_\star = \{(i,j) : \beta_{ij}^\star \neq 0,\ 1 \le i < j \le p\},$$

and denote that $t_n = |\mathcal{M}_\star|$ and $s_n = |\mathcal{N}_\star|$, then the non-sparsity size t_n and s_n must be very small. Our focus is to find an estimated set of $\mathcal{N}_\star$ in this chapter.

In this chapter, we propose a statistical framework-simple sure screening procedure (SSI) based on the increment of log-likelihood function to detect significant interactions for the class of models (2.2). Furthermore, we indicate that our proposed method possesses the sure screening properties and can control the false discovery rate at a reasonable size, hence, the regularized variable selection methods can be easily applied to get more accurate feature selection in the following model selection procedure. Especially, the heredity assumption is not needed in our method.

Computational efficiency is an essential concern for feature screening algorithms. As a result, we first provide one much efficient and general algorithm to exhaustively search the important interaction effects in terms of Rcpp. Rcpp is an R add-on package which facilitates extending R with C++ functions. Therefore, it will make our algorithm much more efficient and realize the exhaustive selection. Secondly, in

order to further strengthen the efficiency of screening algorithms, we propose a new efficient algorithm-the discretized sure screening for interaction, in short, DSSI, the idea of which is based on the data discretization and Boolean representation of the data. Data discretization has a long history in data analysis because discrete values play important roles in data mining and knowledge discovery, which is frequently used in the preprocessing of data (Liu et al. [2002]). Although the loss of information is inevitable, the most essential element is that the success of the discretization can significantly reinforce the interpretation of the results, awfully improve the efficiency of data analysis, and broaden the application of many learning algorithms. A lot of algorithms and methods can only handle the data with discrete type. For example, decision tree needs to discretize continuous variables during the process of tree building. Another technique of DSSI is the data's Boolean representation, which transforms the original data to Boolean type. It will reduce the storage and faster the computation by using the logic operation. The Boolean representation was applied by Wan et al. [2010a] to the "BOOST" method, which is one fast and efficient tool for an exhaustive search for epistatic interaction in genetics. It can screen all of the interactions and filter out nonsignificant interaction effects in GWA studies. And also, we extend this idea to the method SIS proposed by Fan and Lv [2008] and authenticate the extended version-discretized sure independent screening (DSIS). From the point of theory, we show that two pairs: SIS and DSIS, SSI and DSSI have the consistent screening results respectively, and numerical analysis demonstrates that they have same better performances in feature screening in the ultra-dimensional generalized linear model.

The rest of this chapter is organized as followed. In Section 2.2, we briefly introduce our methodology and main results, including sure screening procedure for interactions, sure screening properties and the uniform convergence of the marginal maximum likelihood estimator. The general algorithm SSI and new algorithm DSSI are presented in Section 2.3. In Section 2.4, we consider the relationship between SIS and DSIS, SSI and DSSI from the viewpoint of theory, and prove that they have consistent results in variable screening. Section 2.5 and 2.6 examine the finite sample performance of the proposed method SSI and DSSI on the simulated data compared

with other three methods: RRAMP, xyz and IP. Three real data sets in Section 2.7 are used to demonstrate the usage of our approaches. Our findings and conclusions are summarized in Section 2.8. The detailed proofs are relegated to Section 2.9.

2.2 Methodology and Main results

2.2.1 SSI for Two-Way Interaction in Generalized Linear Models

Assume that the set of variables $\{X_1, X_2, X_3, \ldots, X_p\}$ may be related to the response Y and the set $\{X_{12}, X_{13}, \ldots, X_{(p-1)p}\}$ includes all two-way interactions of the former set, in which $X_{ij} = X_i * X_j$, $1 \leq i < j \leq p$. Let $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{I}}^T)^T$, where $\boldsymbol{X}_{\mathcal{C}} = (X_0, X_1, X_2, X_3, \ldots, X_p)^T$ with $X_0 = 1$ and $\boldsymbol{X}_{\mathcal{I}} = (X_{12}, X_{13}, \ldots, X_{(p-1)p})^T$. Furthermore, assume that X_{ij} is centralized, i.e., $E(X_{ij}) = 0$, $1 \leq i < j \leq p$. And we wish to select the most important interactions from the set $\boldsymbol{X}_{\mathcal{I}}$ to better explain the response Y. The corresponding sets of coefficient are

$$\boldsymbol{\beta}_{\mathcal{C}} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p, \quad \text{and} \quad \boldsymbol{\beta}_{\mathcal{I}} = (\beta_{12}, \beta_{13}, \ldots, \beta_{(p-1)p})^T \in \mathbb{R}^q,$$

where $q = \binom{p}{2} = \frac{p(p-1)}{2}$. And then, our model (2.2) can be rewritten as

$$E(Y|X = \boldsymbol{x}) = g^{-1}(\boldsymbol{X}^T\boldsymbol{\beta}). \tag{2.3}$$

Fan et al. [2009] indicated that we can select the important variables by sorting the marginal likelihood functions. Later, Fan and Song [2010] pointed out that their technique can be considered as the marginal likelihood ratio screening, which builds on the difference between two marginal log-likelihood functions. Here, we still use this method-the likelihood ratio screening and its procedures are listed as follows.

Denote that a random sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ is from the model (2.2) with the canonical link. Let $\boldsymbol{X}_{ij} = (1, X_i, X_j, X_{ij})^T$ and $\boldsymbol{X}_{i,j} = (1, X_i, X_j)^T$. And their coefficients are expressed as $\boldsymbol{\beta}_{ij} = (\beta_{ij0}, \beta_i, \beta_j, \beta_{ij})^T$ and $\boldsymbol{\beta}_{i,j} = (\beta_{i,j0}, \beta_{i,}, \beta_{j,})^T$, respectively. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{ij}^M$ is expressed as the minimizer of the marginal

regression

$$\hat{\boldsymbol{\beta}}_{ij}^{M} = \arg\min_{\boldsymbol{\beta}_{ij}} \mathbb{P}_n\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)\}$$

where $l(\theta, Y) = b(\theta) - \theta Y - c(Y)$ and $\mathbb{P}_n f(\boldsymbol{X}, Y) = n^{-1}\sum_{i=1}^{n} f(\boldsymbol{X}_i, Y_i)$ is the empirical measure. Correspondingly, the population version of the minimizer of the componentwise regression is denoted as

$$\boldsymbol{\beta}_{ij}^{M} = \arg\min_{\boldsymbol{\beta}_{ij}} E\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)\}.$$

Similarly, if the interactions are not included in our model, we define

$$\hat{\boldsymbol{\beta}}_{i,j}^{M} = \arg\min_{\boldsymbol{\beta}_{i,j}} \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}, Y)\},$$

and its population version

$$\boldsymbol{\beta}_{i,j}^{M} = \arg\min_{\boldsymbol{\beta}_{i,j}} E\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}, Y)\}.$$

Finally, denote that

$$L_{ij,n} = \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\hat{\boldsymbol{\beta}}_{i,j}^{M}, Y) - l(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}_{ij}^{M}, Y)\}$$

and $\boldsymbol{L}_n = (L_{12,n}, \ldots, L_{(p-1)p,n})^T \in \mathbb{R}^q$. Here, $L_{ij,n}$ measures the strength of the interaction X_{ij} in our model. The larger $L_{ij,n}$, the more the interaction X_{ij} contributes to the response Y. Correspondingly, let

$$L_{ij}^{\star} = E\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^{M}, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^{M}, Y)\}$$

and $\boldsymbol{L}^{\star} = (L_{12}^{\star}, \ldots, L_{(p-1)p}^{\star})^T \in \mathbb{R}^q$. We sort the vector $\boldsymbol{L}_n$ in a descent order and select a set of variables

$$\widehat{\mathcal{N}}_{\gamma_n} = \{(i,j) : L_{ij,n} \geq \gamma_n,\ 1 \leq i < j \leq p\},$$

where γ_n is a predefined threshold value. In the following section, we will illustrate

that our method satisfies the sure screening property under some certain conditions.

Actually, the coefficient β_{ij}^M is able to measure the importance of the interaction terms. It is impossible that the joint regression parameter $\beta_{ij}^\star$ is same as the marginal regression coefficient β_{ij}^M, but we expect that, in most cases, $|\beta_{ij}^M|$ is larger than some threshold if $|\beta_{ij}^\star|$ exceeds another certain threshold. Here, we replace $|\beta_{ij}^M|$ by the marginal log-likelihood function's increment $L_{ij}^\star$.

For the marginal regression coefficients $\boldsymbol{\beta}_{ij}^M$, by the fact that $E(Y|\boldsymbol{X}) = b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)$, they satisfy the score equation

$$E\left\{b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M)\boldsymbol{X}_{ij}\right\} = E(Y\boldsymbol{X}_{ij}) = E\left\{b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)\boldsymbol{X}_{ij}\right\}. \tag{2.4}$$

If the marginal model does not include the interaction X_{ij}, we gain the score equation just like

$$E\left\{b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)\boldsymbol{X}_{i,j}\right\} = E(Y\boldsymbol{X}_{i,j}) = E\left\{b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)\boldsymbol{X}_{i,j}\right\}; \tag{2.5}$$

that is,

$$E\left(Y - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)\right)\boldsymbol{X}_{i,j} = 0. \tag{2.6}$$

When the marginal coefficient of the interaction term is equal to zero, i.e., $\beta_{ij}^M = 0$, by (2.4), the first 3 components of $\boldsymbol{\beta}_{ij}^M$, should be equal to $\boldsymbol{\beta}_{i,j}^M$ by the uniqueness of the solution of the score equation (2.6). Therefore, the score equation (2.4) on the interaction X_{ij} entails that

$$E\left\{b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)X_{ij}\right\} = E(YX_{ij}), \quad \text{or} \quad E\left\{(Y - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M\})X_{ij}\right\} = 0. \tag{2.7}$$

Follow that the definition of the conditional linear expectation, provided by Barut et al. [2016], is the best linearly fitted regression within the class of linear functions, we denote that

$$E_L(Y|\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) = b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) \quad \text{and} \quad E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) = b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M);$$

and then equation (2.7) becomes

$$E\left(Y - E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)\right)X_{ij} = 0 \tag{2.8}$$

Furthermore, define that the notation $E_L(X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)$ is the best linear regression fit of X_{ij} by using $\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M$. Then, equation (2.6) can be expressed as

$$E(Y - E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))\boldsymbol{X}_{i,j} = 0 \tag{2.9}$$

Combing equation (2.8) and (2.9), we define that

$$\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) \equiv E(X_{ij} - E_L(X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))(Y - E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)) = 0.$$

Consequently, we obtain that if $\beta_{ij}^M = 0$, $\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) = 0$.

2.2.2 Sure Screening Properties

Denote that $\boldsymbol{\beta}_{ij} = (\beta_{ij0}, \beta_i, \beta_j, \beta_{ij})^T$ be the four-dimensional parameter, and let $\boldsymbol{X}_{ij} = (1, X_i, X_j, X_{ij})^T$. Since the log-likelihood function is of the concavity in the generalized linear model with the canonical link function, the function $El(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)$ can arrive at its unique minimum $El(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)$ over $\boldsymbol{\beta}_{ij} \in \mathcal{B}$, in which $\boldsymbol{\beta}_{ij}^M = (\beta_{ij0}^M, \beta_i^M, \beta_j^M, \beta_{ij}^M)^T$ is an interior point of the set $\mathcal{B}$ and $\mathcal{B} = \{|\beta_{ij,0}^M| \leq B, |\beta_i^M| \leq B, |\beta_j^M| \leq B, |\beta_{ij}^M| \leq B\}$ is an area with the width B where the marginal likelihood is maximized. The following conditions are needed:

(A) The marginal Fisher information: $\boldsymbol{I}_{ij}(\boldsymbol{\beta}_{ij}) = E\{b''(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij})\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T\}$ is finite and positive definite at $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{ij}^M$, for $1 \leq i < j \leq p_n$. Moreover, $\|\boldsymbol{I}_{ij}(\boldsymbol{\beta}_{ij})\|_{\mathcal{B}} = \sup_{\boldsymbol{\beta}_{ij} \in \mathcal{B}, \|\boldsymbol{x}\|=1} \|\boldsymbol{I}_{ij}(\boldsymbol{\beta}_{ij})^{1/2}\boldsymbol{x}\|$ is bounded from above.

(B) (i) Let $\boldsymbol{X}_{i,j} = (1, X_i, X_j)^T$. For $(i, j) \in \mathcal{N}_\star$, there exists a constant $c_1 > 0$ such that $|\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)| \geq c_1 n^{-\kappa}$ for some $0 < \kappa < 1/4$.

(ii) Denote m_{ij} be the random variable defined by

$$m_{ij} = \frac{b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)}{\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M - \boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M},$$

and $E(m_{ij}X_{ij}^2) = E(m_{ij}X_i^2X_j^2) \leq c_2$ uniformly for some constant c_2, in which $1 \leq i < j \leq p$.

(C) For all $\boldsymbol{\beta}_{ij} \in \mathcal{B}$, $E(l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)) \geq V\|\boldsymbol{\beta}_{ij} - \boldsymbol{\beta}_{ij}^M\|^2$, for some constant $V > 0$, bounded from below uniformly for all $1 \leq i < j \leq p$.

(D) There exists some constants m_0, m_1, s_0, $s_1 > 0$ and $\alpha > 0$, such that for a sufficiently large $t > 0$,

$$P(|X_i| > t) \leq m_1 \exp\{-m_0 t^\alpha\} \qquad \text{for } 1 \leq i \leq p,$$

and that

$$E \exp(b(\boldsymbol{X}^T\boldsymbol{\beta}^\star + s_0) - b(\boldsymbol{X}^T\boldsymbol{\beta}^\star)) + E \exp(b(\boldsymbol{X}^T\boldsymbol{\beta}^\star - s_0) - b(\boldsymbol{X}^T\boldsymbol{\beta}^\star)) \leq s_1.$$

(E) For the function $b(\theta)$, the second derivative $b''(\theta)$ is one continuous function and $b''(\theta) > 0$. There exists $\varepsilon_1 > 0$ such that for all $1 \leq i < j \leq p$,

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_{ij}^M\| \leq \varepsilon_1} |E\{b(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}) I(|X_{ij}| > K_n)\}| \leq o(n^{-1}),$$

where $I(\cdot)$ is the indicator function and K_n is an arbitrarily large constant such that for a given $\boldsymbol{\beta}$ in $\mathcal{B}$, the function $l(\boldsymbol{x}^T\boldsymbol{\beta}, y)$ satisfies the Lipschitz property with positive constant k_n for all $(\boldsymbol{x}, y)$ in the set $\Omega_n = \{(\boldsymbol{x}, y) : \|\boldsymbol{x}\|_\infty \leq K_n, |y| \leq K_n^\star\}$ with $K_n^\star = m_0 K_n^\alpha / s_0$, in which $\|\cdot\|_\infty$ be the supremum norm.
(F) The variance $\mathrm{Var}(\boldsymbol{X}^T\boldsymbol{\beta}^\star) = \boldsymbol{\beta}^{\star T}\boldsymbol{\Sigma}\boldsymbol{\beta}^\star$, where $\boldsymbol{\Sigma} = \mathrm{diag}(0, \boldsymbol{\Sigma}_1)$ with $\boldsymbol{\Sigma}_1 = \mathrm{Var}(\boldsymbol{X})$, and $b''(\cdot)$ are bounded.
(G) The minimum eigenvalue of the matrix $E[m_{ij}\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T]$ is uniformly larger than a positive constant for any i, j, where m_{ij} is defined in Condition B(ii).
(H) Denote that $\boldsymbol{\beta}_{ij-}^M = (\beta_{i,j0}^M, 0, \ldots, 0, \beta_{i,}^M, 0, \ldots, 0, \beta_{j,}^M, 0, \ldots, 0)^T$, $\Delta\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{\mathcal{C}}^\star - \boldsymbol{\beta}_{ij-}^M$. Let $R_{ij} = E[X_{ij}\boldsymbol{X}_{\mathcal{C}}^T\Delta\boldsymbol{\beta}_{ij}]$ and $\boldsymbol{R} = (R_{12}, R_{13}, \ldots, R_{(p-1)p})^T$, hence, it holds that $\|\boldsymbol{R}\|_2^2 = o(\lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{I}}))$, where $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{I}})$ be the largest eigenvalue of the matrix $\boldsymbol{\Sigma}_{\mathcal{I}}$.

All of Conditions are similar to them that proposed by Fan and Song [2010] and Barut et al. [2016], and are satisfied by most of the generalized linear models such as linear regression and logistic regression. By the strict convexity property of $b(\theta)$, m_{ij} is almost surely larger than 0. If $b(\theta) = \theta^2/2$, then $m_{ij} = 1$ and Condition B(ii) is automatically satisfied by the uniform bounded property of $E(X_{ij}^2)$ since X_i and X_j is normalized. The first part of Condition (D) builds an exponential bound on the

tails of X_j. Actually, since the event $\{\omega : |X_{ij}(\omega)| > t\}$ is a subset of the union of $\{\omega : |X_i(\omega)| > \sqrt{t}\}$ and $\{\omega : |X_j(\omega)| > \sqrt{t}\}$, when $P(|X_i| > \sqrt{t}) \leq m_1' \exp\{-m_0 t^{\alpha/2}\}$ and $P(|X_j| > \sqrt{t}) \leq m_1' \exp\{-m_0 t^{\alpha/2}\}$, we have that

$$P(|X_{ij}| > t) \leq 2m_1' \exp\{-m_0 t^{\alpha/2}\} \quad \text{for } 1 \leq i < j \leq p.$$

And next, we can take $m_1 = 2m_1'$ and by $\exp\{-m_0 t^{\alpha}\} < \exp\{-m_0 t^{\alpha/2}\}$, the exponential bound on the tails is simultaneously available for main effect and interaction terms. Hence, the first part of Condition (D) also implies that an exponential bound is built on the tails of X_{ij}. And the second part of Condition (D) points out that the response variable Y possesses the exponentially light tail, as shown in Lemma 1 of Fan and Song [2010]. In our proof, we need this Lemma.

Since our purpose is to preserve the important interactions in our model, one critical question would be: at what level the interactions of variables should be preserved. For sure screening purposes, if one interaction X_{ij} is jointly important ($\beta_{ij}^{\star} \neq 0$), will it still be marginally important ($\beta_{ij}^{M} \neq 0$) ? On the other hand, for the model selection consistency purpose, when one interaction is jointly unimportant ($\beta_{ij}^{\star} = 0$), will it still be marginally unimportant ($\beta_{ij}^{M} = 0$) ? In this section, we will provide the answers for them.

Theorem 2.2.1 *For $1 \leq i < j \leq p$, the marginal likelihood increment $L_{ij}^{\star} = 0$ if and only if $\beta_{ij}^{M} = 0$.*

Theorem 2.2.2 *For $1 \leq i < j \leq p$, the marginal regression parameters $\beta_{ij}^{M} = 0$ if and only if $Cov_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M) = 0$.*

Corollary 2.2.1 *For $1 \leq i < j \leq p$, the marginal likelihood increment $L_{ij}^{\star} = 0$ if and only if $Cov_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M) = 0$.*

The above theorems and Corollary reveal that the increments of the log-likelihood function is a measurement of the relationship between the interaction and the mean response function, and also the marginal regression parameter is one kind of the measurement. They are equivalent and here we use the former.

To distinguish the active interactions $\{X_{ij} : (i,j) \in \mathcal{N}_\star\}$ and inactive interactions $\{X_{ij} : (i,j) \notin \mathcal{N}_\star\}$, we need to set up one appropriate threshold value γ_n, so that the minimum marginal signal strength is stronger than the stochastic noise and the sure screening property will be guaranteed. This will be shown in Theorem 2.2.3 and Theorem 2.2.4. Their proofs are listed in the last section of this chapter.

Theorem 2.2.3 *If Condition (B) holds, then there exist a positive constant c_3 such that*

$$\min_{(i,j)\in\mathcal{N}_\star} |\beta_{ij}^M| \geq c_3 n^{-\kappa}.$$

Theorem 2.2.4 *Under the conditions (B) and (C), we have*

$$\min_{(i,j)\in\mathcal{N}_\star} L_{ij}^\star \geq c_4 n^{-2\kappa}$$

for some positive constant c_4.

Next, we are going to establish the sure screening property. The crucial point is to build the uniform convergence of likelihood ratio screening. And then, we obtain the uniform convergence rate and sure screening property for it. The former will be beneficial for us to control the size of the selected set. The results are given in the following theorem.

Theorem 2.2.5 *Assume that Conditions (A), (B), (C), (D) and (E) hold. Let $k_n = b'(3K_nB + B) + m_0K_n^\alpha/s_0$, with K_n given in Condition (E).*

(i) If $n^{1-2\kappa}/(k_nK_n)^2 \to \infty$, then for any $c_5 > 0$, there exists a constant $c_6 > 0$ such that

$$\begin{aligned} & P\left(\max_{1\leq i<j\leq p} |\hat{\beta}_{ij}^M - \beta_{ij}^M| \geq c_5 n^{-\kappa}\right) \\ \leq\ & q\left(\exp(-c_6 n^{1-2\kappa}/(k_nK_n)^2) + nm_2\exp(-m_0K_n^{\alpha/2})\right), \end{aligned}$$

where $q = \frac{p(p-1)}{2}$ and $m_2 = 3m_1 + s_1$.

(ii) If $n^{1-2\kappa}/(k_nK_n)^2 \to \infty$, then for any $c_7 > 0$, there exists two constants $c_8 > 0$

and $c_9 > 0$ *such that*

$$\begin{aligned} & P\left(\max_{1\le i<j\le p} |L_{ij,n} - L^\star_{ij}| \ge c_7 n^{-2\kappa}\right) \\ \le\ & q\left(2\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp(-c_9 n^{1-4\kappa}) + 4nm_2\exp(-m_0 K_n^{\alpha/2})\right), \end{aligned}$$

(iii) In addition, by taking $\gamma_n = c_{10} n^{-2\kappa}$ *with* $c_{10} \le c_4/2$ *, we obtain*

$$\begin{aligned} & P(\mathcal{N}_\star \subset \widehat{\mathcal{N}}_{\gamma_n}) \\ \ge\ & 1 - s_n\left(2\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp(-c_9 n^{1-4\kappa}) + 4nm_2\exp(-m_0 K_n^{\alpha/2})\right), \end{aligned}$$

where $s_n = |\mathcal{N}_\star|$*, the size of active interactions.*

Note that the sure screening property given in Theorem 2.2.5(iii) only relates to the size s_n of active interactions. The dimensionality p or q does not matter for the purpose of sure screening. For generalized linear model (2.3), such as logistic regression, $b(\theta) = \ln(1+\exp(\theta))$, and $b'(\theta) = \frac{1}{1+\exp(-\theta)}$ is bounded. By Theorem 2.2.5(ii), the optimal order of K_n is $n^{(1-4\kappa)/(\alpha+2)}$, and

$$P\left(\max_{1\le i<j\le p} |L_{ij,n} - L^\star_{ij}| \ge c_7 n^{-2\kappa}\right) = O\left\{p^2\exp(-c_9 n^{(1-4\kappa)\alpha/(\alpha+2)})\right\}.$$

Thus, the tail probability will be exponentially small. That is, we can deal with the NP-dimensionality

$$\ln p = o\left(n^{(1-4\kappa)\alpha/(\alpha+2)}\right)$$

with $\alpha = \infty$ of special case of the bounded covariates and $\alpha = 2$ of normal covariates. Similar results for unconditional screening and conditional screening are shown in Fan and Song [2010] and Barut et al. [2016], respectively.

After illustrating the sure screening property of likelihood ratio screening for interactions, we will point out that the false selection rate can be controlled absolutely in the remaining part of this section. In other words, the size of the set $\widehat{\mathcal{N}}_{\gamma_n}$ can be controlled and the number of interactions would be actually reduced. Here, we provide a bound on the size of selected set of interactions in the following theorem.

Theorem 2.2.6 *Under Conditions (A)-(H), we have*

$$\begin{aligned} & P\left(|\widehat{\mathcal{N}}_{\gamma_n}| \leq O(n^{2\kappa}\lambda_{max}(\boldsymbol{\Sigma}_{\mathcal{I}}))\right) \\ \geq \; & 1 - q\left(2\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp(-c_9 n^{1-4\kappa}) + 4nm_2\exp(-m_0 K_n^{\alpha/2})\right). \end{aligned}$$

where $q = \frac{p(p-1)}{2}$ *and* $m_2 = 3m_1 + s_1$.

From the proof of Theorem 2.2.6, without Condition (H), Theorem 2.2.6 still holds with $\boldsymbol{\Sigma}_{\mathcal{I}}$ replaced by $\boldsymbol{\Sigma}_{\mathcal{I}} + \boldsymbol{R}\boldsymbol{R}^T$. Note that the right-hand side probability has been explained in the last paragraph. If $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{I}}) = O(n^{\tau})$, the size of the selected set has order $O(n^{2\kappa+\tau})$, the same order as in the approach of Fan and Lv [2008]. Our result is an extension of the work of Fan and Lv (2008). Similar results has been shown in Fan and Song [2010], Fan et al. [2011], Li et al. [2012a], and Barut et al. [2016].

2.3 Algorithm

In this section, we first introduce the general algorithm SSI and the more efficient method DSSI will be shown in the second part.

2.3.1 Description of the General Algorithm-SSI

Here, we only consider the interaction screening and the general algorithm works as follows:

Step 1. For any $1 \leq i < j \leq p$, compute

$$\hat{\boldsymbol{\beta}}_{ij}^M = \arg\min_{\boldsymbol{\beta}_{ij}} \mathbb{P}_n\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)\}, \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{i,j}^M = \arg\min_{\boldsymbol{\beta}_{i,j}} \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}, Y)\},$$

and then calculate the sample statistics

$$L_{ij,n} = \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\hat{\boldsymbol{\beta}}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}_{ij}^M, Y)\}.$$

Step 2. Choose the thresholding value $\gamma_n = c_{10}n^{-2\kappa}$ and select the following interactions:

$$\widehat{\mathcal{N}}_{\gamma_n} = \{(i,j) : L_{ij,n} \geq \gamma_n,\ 1 \leq i < j \leq p\}.$$

Usually, we take the d largest $L_{ij,n}$, and Fan and Lv [2008] proposed that $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ or $n-1$, where the braces indicate the floor function. In our numerical examples, to avoid the issues of choosing the thresholding parameter, we take the median minimum model size (MMMS) as one measure of the effectiveness of interaction screening methods.

Since the number of pairwise interaction increases quadratically with the dimension p of the data set, the efficiency of the algorithm must be considered as the essential point to assess any one of the algorithms. The traditional algorithms may be time-exhausting, so we consider **Rcpp** to complete our algorithm in the software R. It makes our algorithm more efficient. In order to further improve the efficiency, we will first transform the continuous features to discrete values, and then use Boolean representation to express the discrete values. According to these two procedures, another more efficient algorithm DSSI is shown in the following section.

2.3.2 Boolean Representation

Assume that the continuous data set $\boldsymbol{X}$ is one $n \times p$ matrix with n observations and p predictors, Y be the response. After discretizing data set $\boldsymbol{X}$ and response Y, each predictor $\widetilde{X}_i$ has l levels and $\widetilde{Y}$ has m categories. Here, we take $l = 3$ and $m = 2$ as an example and assume that $\widetilde{Y}$ has two values (0 and 1). Instead of using one column for each predictor $\widetilde{X}_i$, the new representation uses 3 columns since 3 values are included in each $\widetilde{X}_i$. Each column consists of two bit strings, one for samples with $\widetilde{Y} = 0$ and the other for them with $\widetilde{Y} = 1$, and each bit can represent one sample in the string. The values (0 and 1) illustrate whether the sample belongs to some category of each predictor X_i. For instance, we have one discretized data set $\widetilde{\boldsymbol{X}}$ with 2 predictors and 16 samples, where the first 8 rows represent samples with $\widetilde{Y} = 0$ and the other samples with $\widetilde{Y} = 1$:

$$\widetilde{\boldsymbol{X}}^T = \begin{array}{c} \\ \widetilde{Y} \\ \widetilde{X}_1 \\ \widetilde{X}_2 \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 3 & 1 & 2 & 3 & 2 & \vdots & 2 & 2 & 1 & 1 & 3 & 2 & 2 & 1 \\ 3 & 2 & 1 & 1 & 3 & 2 & 2 & 1 & \vdots & 2 & 3 & 2 & 3 & 1 & 2 & 3 & 2 \end{bmatrix}$$

and its Boolean representation is

$$\widetilde{\boldsymbol{X}}_{bit}^{T} = \begin{array}{c} \\ \widetilde{X}_1 = 1 \\ \widetilde{X}_1 = 2 \\ \widetilde{X}_1 = 3 \\ \widetilde{X}_2 = 1 \\ \widetilde{X}_2 = 2 \\ \widetilde{X}_2 = 3 \end{array} \begin{bmatrix} \widetilde{Y} = 0 & \widetilde{Y} = 1 \\ 10001000 & 00110001 \\ 00100101 & 11000110 \\ 01010010 & 00001000 \\ 00110001 & 00001000 \\ 01000110 & 10100101 \\ 10001000 & 01010010 \end{bmatrix}.$$

From the Boolean representation $\widetilde{\boldsymbol{X}}_{bit}$, we can easily find that the first sample belongs to the first category of X_1 and the third category of X_2. And also, we can quickly obtain the number of observations that belong to any two categories by taking the logic operation. For example, if we want to calculate the number of samples with $\widetilde{X}_1 = 2$ and $\widetilde{X}_2 = 2$ in the category $\widetilde{Y} = 0$, we just conduct the logical **AND** operation:

$$00100101 \quad \textbf{AND} \quad 01000110 = 00000100,$$

and then, we count the number of 1s in the final string "00000100", that is 1. This result is consistent to that in $\widetilde{\boldsymbol{X}}$. As a result, it is more efficient by using $\widetilde{\boldsymbol{X}}_{bit}$ to construct the contingency table for any two discretized predictors. Since the fast logic operation with $\widetilde{\boldsymbol{X}}_{bit}$ is utilized, we can accelerate our calculation in our algorithm. Obviously, $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{X}}_{bit}$ are equivalent and they store the same information. Because one byte can store 8 bits, $\widetilde{\boldsymbol{X}}_{bit}$ occupies smaller storage space in the computer comparing with $\widetilde{\boldsymbol{X}}$ although the number of $\widetilde{\boldsymbol{X}}_{bit}$'s columns is l times as large as that of $\widetilde{\boldsymbol{X}}$'s columns. As a result, the Boolean representation can reduce dramtically the storage space. This is another one of its advantages.

2.3.3 New Algorithm "DSSI"

In this subsection, we are going to illustrate the new algorithm "DSSI" (Discretized SSI). Before introducing this method, we need some preliminary knowledge. The first step of DSSI is to discretize one continuous attribute by creating one categorical

variable with a specified number of levels. Binning is the simplest method, which includes equal-width and equal-frequency. To ensure the power of screening, we choose the equal-frequency method, which means that the quantiles are used to split the domain of variables to several intervals. The number of intervals is called "arity" in the discretization context (See Liu et al. [2002]). Assume that the arity is denoted by a, and then $a-1$ is the maximum number of cut-points of the continuous features. Actually, there is a trade-off between the arity and accuracy of data analysis. Higher arity may make the results difficultly understood but be of high predictive accuracy while lower arity may have negative influence on the predictive accuracy but positively improve the efficiency. One natural question is: which one of the values is the best choice for the arity a. Usually, we recommend $a = 2$ or 3. Obviously, here we can choose different a_i for different continuous features X_i and the algorithm still works.

After discretization, all predictors and the response will be categorical variables. The second step is to estimate the increment of log-likelihood function by using these new predictors. Under the circumstances, we usually take the logistic model (for binary response) or baseline-category logit models (for the response with several categories) to fit the data set. Actually, the logistic regression models or baseline-category logit models have their corresponding log-linear regression models for contingency table when the predictor and the response are categorical (See Agresti [2002]). Based on this equivalence, the significance of interaction effects can be measured by the increment of log-likelihood function in the corresponding log-linear regression models.

Assume that we consider the following two logistic models-the logistic regression model with main effect and the full logistic regression model:

$$\text{logit}(P(Y = 1|X, Z)) = \beta_0 + \beta_i^X + \beta_j^Z \tag{2.10}$$

and

$$\text{logit}(P(Y = 1|X, Z)) = \beta_0 + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}. \tag{2.11}$$

Denote that $\widehat{l}_M$ and $\widehat{l}_F$ be the sample version of the negative maximum log-likelihood functions of the logistic regression model (2.10) with main effect and the full logistic regression model (2.11), respectively. The increment of the log-likelihood function is

defined as $\widehat{l}_M - \widehat{l}_F$. The corresponding log-linear regression models can be expressed as

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Z + \lambda_k^Y + \lambda_{ij}^{XZ} + \lambda_{ik}^{XY} + \lambda_{jk}^{ZY} \tag{2.12}$$

and

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Z + \lambda_k^Y + \lambda_{ij}^{XZ} + \lambda_{ik}^{XY} + \lambda_{jk}^{ZY} + \lambda_{ijk}^{XZY}. \tag{2.13}$$

Let $\widehat{l}_H$ and $\widehat{l}_S$ be the sample version of the negative maximum log-likelihood functions of the homogeneous association regression model (2.12) and the saturated model (2.13), respectively. $\widehat{l}_H - \widehat{l}_S$ is the corresponding increment of log-likelihood function. Thus, we can take advantage of $\widehat{l}_H - \widehat{l}_S$ to screen the interaction terms instead of using $\widehat{l}_M - \widehat{l}_F$.

Now we want to obtain the difference $\widehat{l}_H - \widehat{l}_S$. Suppose that we have one three-way $(I \times J \times K)$ table with cell counts $\{n_{ijk}\}$ of random variables X, Z and Y. And the kernel of the log-likelihood function for this contingency table is

$$L(\boldsymbol{\mu}) = \sum_{ijk} \log(\mu_{ijk}) - \sum_{ijk} \mu_{ijk}.$$

Here, we need some notations. Denote that $\pi_{i++} = \sum_{jk} \pi_{ijk}$ is the marginal probability of $X = i$ and $n_{i++} = \sum_{jk} n_{ijk}$ is the number of samples with $X = i$, $\pi_{ij+} = \sum_k \pi_{ijk}$ is the marginal probability of $X = i$ and $Z = j$ and $n_{ij+} = \sum_k n_{ijk}$ is the corresponding count. Similarly, $\pi_{+j+} = \sum_{ik} \pi_{ijk}$, $\pi_{++k} = \sum_{ij} \pi_{ijk}$, $\pi_{i+k} = \sum_j \pi_{ijk}$, $n_{i+k} = \sum_j n_{ijk}$, $\pi_{+jk} = \sum_i \pi_{ijk}$, $n_{+j+} = \sum_{ik} n_{ijk}$, $n_{++k} = \sum_{ij} n_{ijk}$, $n_{+jk} = \sum_i n_{ijk}$.

For the saturated model (2.13), we know that $\widehat{\mu}_{ijk} = n_{ijk}$ and directly get the estimation $\widehat{l}_S = \sum_{ijk} n_{ijk} - \sum_{ijk} \log(n_{ijk}) - \sum_{ijk} n_{ijk}$. And for the homogeneous association regression model (2.12), the iterative proportional fitting (IPF) algorithm is recommended by calculating the estimation of μ_{ijk}, which was introduced by Deming and Stephan [1940]. Three steps are included in the first cycle of the IPF algorithm:

$$\mu_{ijk}^{(1)} = \mu_{ijk}^{(0)} \frac{n_{ij+}}{\mu_{ij+}^{(0)}}, \quad \mu_{ijk}^{(2)} = \mu_{ijk}^{(1)} \frac{n_{i+k}}{\mu_{i+k}^{(1)}}, \quad \mu_{ijk}^{(3)} = \mu_{ijk}^{(2)} \frac{n_{+jk}}{\mu_{+jk}^{(2)}},$$

where $\mu_{ij+} = \sum_k \mu_{ijk}$, , $\mu_{i+k} = \sum_j \mu_{ijk}$, $\mu_{+jk} = \sum_i \mu_{ijk}$. This cycle does not stop until the process converges and the convergence property has been proved by Fienberg

[1970] and Haberman [1974]. We count the number n_{ijk} by using the Boolean representation, thus the contingency table for X and Z given Y can be quickly constructed in a faster manner. Finally, the estimation $\widehat{l}_H$ will be obtained.

In the second step, for our ultra-high dimensional generalized linear model (2.2), instead of calculating the increment $\widetilde{L}_{ij,n} = \widehat{l}_{M_{ij}} - \widehat{l}_{F_{ij}}$ for any pair of $\widetilde{X}_i$ and $\widetilde{X}_j$, we compute the new increment of the log-likelihood function $\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$. In the third step, by taking the thresholding value γ_n or choosing the d largest $\widetilde{L}'_{ij,n}$, where $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ or $n-1$. Here, the IPF method is needed to compute all $\widehat{l}_{H_{ij}}$ for any pair of $\widetilde{X}_i$ and $\widetilde{X}_j$. Finally, we obtain the selected set $\widehat{\mathcal{N}}_{\gamma_n}$. Our algorithm DSSI is summarized as follows:

Step 1. For any pair of the continuous variables X_i and X_j, $1 \le i < j \le p$, transform them to the corresponding discretized variables $\widetilde{X}_i$ with level l_i and $\widetilde{X}_j$ with level l_j, and change the response Y to a categorical variable $\widetilde{Y}$ if necessary.

Step 2. By using the IPF algorithm to estimate $\widetilde{l}_{H_{ij}}$ for all pairs of X_i and X_j, we compute $\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$.

Step 3. Choose the threshold γ_n and select the following interactions:

$$\widehat{\mathcal{N}}_{\gamma_n} = \{(i,j) : \widetilde{L}'_{ij,n} \ge \gamma_n,\ 1 \le i < j \le p\}.$$

Usually, we take the d largest $L_{ij,n}$, where $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ or $n-1$.

Sometimes, the dimension p is very large and may be on the order of tens of millions. The IPF method may be time-consuming for computing all $\widehat{l}_{H_{ij}}$. Here, we propose to use one tool to prune some interaction terms in the second step. For the homogeneous association regression model (2.12), Kirkwood Superposition Approximation (KSA), which was firstly proposed by Kirkwood [1935], is utilized to provide an estimator for μ_{ijk}. That is,

$$\widehat{\mu}_{ijk}^{KSA} = \frac{n}{\eta} \frac{\widehat{\pi}_{ij+}\widehat{\pi}_{i+k}\widehat{\pi}_{+jk}}{\widehat{\pi}_{i++}\widehat{\pi}_{+j+}\widehat{\pi}_{++k}},$$

where $\eta = \sum_{ijk} \frac{\widehat{\pi}_{ij+}\widehat{\pi}_{i+k}\widehat{\pi}_{+jk}}{\widehat{\pi}_{i++}\widehat{\pi}_{+j+}\widehat{\pi}_{++k}}$ is a normalization term, $n = \sum_{ijk} n_{ijk}$. And then, we get the approximation estimation $\widehat{l}_{KSA}$. Wan et al. [2010a] showed that $\widehat{l}_{KSA} - \widehat{l}_S$ is

an upper bound of $\widehat{l}_H - \widehat{l}_S$, i.e.,

$$0 \leq \widehat{l}_H - \widehat{l}_S \leq \widehat{l}_{KSA} - \widehat{l}_S.$$

Based on this boundary and by setting up one threshold γ_{KSA}, in the second step, we can filter out many insignificant interaction terms quickly and then reduce the size of a pool of all interaction effects. The value γ_{KSA} can be defined by the conservative Bonferroni correction or specified by user. Obviously, if $\gamma_{KSA} = 0$, no one interaction term is deleted in this step. In the final step, for the remaining interaction terms, we compute their $\widetilde{L}'_{ij,n}$ by the IPF algorithm and take the thresholding value γ_n or select the d largest $\widetilde{L}'_{ij,n}$, where $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ or $n-1$. As a result, the selected set $\widehat{\mathcal{N}}_{\gamma_n}$ is obtained. Here, γ_n can be taken as the Bonferroni correction $100 * (1 - 0.05 * p(p-1)/2)\%$ percentile decided by the χ^2 test with degree freedom $(l_i - 1)(l_j - 1)$ for any one interaction between $\widetilde{X}_i$ and $\widetilde{X}_j$. In summary, our algorithm DSSI with KSA is listed as follows:

Step 1. For any pair of continuous variables X_i and X_j, $1 \leq i < j \leq p$, transform them to corresponding discretized variables $\widetilde{X}_i$ with level l_i and $\widetilde{X}_j$ with level l_j, and change the response Y to a categorical variable $\widetilde{Y}$ if necessary.

Step 2. By using the KSA to approximate $\widetilde{l}_{H_{ij}}$ by the IPF algorithm for all pairs of X_i and X_j, we compute $\widehat{l}_{KSA_{ij}} - \widehat{l}_{S_{ij}}$ and set up the threshold γ_{KSA} to remove a part of interaction terms.

Step 3. For the remaining interaction effects, we compute $\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$ and further identify the important interaction effects by χ^2-test with degree freedom $(l_i - 1)(l_j - 1)$, or directly select the d largest $\widetilde{L}'_{ij,n}$.

So far, we have specified the procedures of our new algorithm "DSSI". Apparently, the new method "DSSI" will be much more efficient than the method "SSI". Since DSSI is based on the sure screening property with maximum likelihood function shown in Section 2.2, this procedure guarantees that once the important interactions enter the pool of important candidates, the probability of selecting the correct ones is very high, actually which can approach to 1 as the sample size tends to infinity. Furthermore, the theorems in the next section tell us that DSSI and SSI will have the consistent results in the sure screening of interaction effects.

2.4 Sure Screening Properties after Discretization

In this section, we first study some properties of the discretized sure screening method and build a bridge between original SIS and discretized SIS (DSIS) for the linear model. Next, we focus on the relationship between SSI and DSSI in our model.

2.4.1 Relationship between SIS and DSIS

We follow the assumption of Fan and Song [2010], and consider the variable or feature selection of the generalized linear model without interaction effects:

$$Y = b'(\boldsymbol{X}^T\boldsymbol{\beta}) + \varepsilon. \tag{2.14}$$

where $\boldsymbol{X} = (X_1,\ X_2,\ \ldots,\ X_p)^T$ is a $p \times 1$ random vector, $\boldsymbol{\beta} = \{\beta_1,\ \beta_2,\ \ldots,\ \beta_p\}$ is the parameter vector, Y is the response, $b'(\cdot)$ is the canonical link function, and assume that

$$\mathcal{M}_\star = \{1 \le k \le p :\ \beta_k \neq 0\}$$

is the set of indexes of nonzero parameter. Define the marginal log-likelihood increment

$$L_k^\star = E\{l(\beta_0^M, Y) - l(\boldsymbol{X}_k^T\boldsymbol{\beta}_k^M, Y)\}, \qquad k = 1, 2, \ldots, p$$

where $\beta_0^M = \arg\min_{\beta_0} El(\beta_0, Y)$, $\boldsymbol{X}_k^T = \{1, X_k\}$, $\boldsymbol{\beta}_k^M = \{\beta_{k,0}, \beta_k^M\}^T$ and

$$\boldsymbol{\beta}_k^M = \arg\min_{\boldsymbol{\beta}_k} El(\boldsymbol{X}_k^T\boldsymbol{\beta}_k, Y).$$

Furthermore, $E(Y) = E(X_k) = 0$ and $E(Y^2) = E(X_k^2) = 1$, $k = 1, 2, \ldots, p$. Let $\rho_k = \mathrm{Corr}(Y, X_k)$ and (Y_1, X_{1k}), (Y_2, X_{2k}) be the independent copies of (Y, X_k).

Assume that S^{X_k} and S^Y are the support sets of variables X_k and Y, respectively. Denote that $\{P_i^{X_k}\}_{i=1}^l$ and $\{P_j^Y\}_{j=1}^m$ are partitions of their supports, which means that

$$\bigcup_{i=1}^{l} P_i^{X_k} = S^{X_k} \quad \text{and} \quad P_{i_1}^{X_k} \bigcap P_{i_2}^{X_k} = \emptyset \ \text{ for } i_1 \neq i_2;$$

and

$$\bigcup_{j=1}^{m} P_j^Y = S^Y \quad \text{and} \quad P_{j_1}^Y \bigcap P_{j_2}^Y = \emptyset \ \text{ for } j_1 \neq j_2;$$

where l and m are two positive constants. Here, the $l-$quantiles and $m-$quantiles are considered as the break points for the partitions of variables X_k and Y. Define

$$\widetilde{X}_k = \begin{cases} 0, & X_k \in P_1^{X_k} \\ 1, & X_k \in P_2^{X_k} \\ \vdots & \vdots \\ l-1, & X_k \in P_l^{X_k} \end{cases} \quad \text{and} \quad \widetilde{Y} = \begin{cases} 0, & Y \in P_1^Y \\ 1, & Y \in P_2^Y \\ \vdots & \vdots \\ m-1, & Y \in P_m^Y \end{cases},$$

and then variables X_k and Y are discretized to two categorical variables $\widetilde{X}_k$ and $\widetilde{Y}$, respectively. Furthermore, denote that $\widetilde{X}_{k_i} = I(X_k \in P_i^{X_k})$, $1 \leq i \leq l$ and $\widetilde{Y}_j = I(Y \in P_j^Y)$, $1 \leq j \leq m$, where $I(\cdot)$ is the indicator function. After discretization, we have the new increment of log-likelihood function

$$\widetilde{L}_k^\star = E\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_k^T \widetilde{\boldsymbol{\beta}}_k^M, \widetilde{Y})\}, \qquad k = 1, 2, \ldots, p.$$

Remark 2.4.1 *In this paper, we consider continuous response Y and 2-quantile (median) for the response Y, that is, $m = 2$ and*

$$\widetilde{Y} = \begin{cases} 0, & Y \leq M_d(Y) \\ 1, & Y > M_d(Y) \end{cases},$$

where $M_d(Y)$ is the median of the response Y. Actually, we will first give the proof of the following theorem for the case $m = 2$ and $l = 2$, and then extend the case $l = 2$ to $l \geq 3$.

Before illustrating our theorem, we need some marginally symmetric conditions, which were provided by Li et al. [2012a].

(M1) Denote $\Delta\varepsilon_k = Y_1 - Y_2 - \rho_k(X_{1k} - X_{2k})$ and $\Delta X_k = X_{1k} - X_{2k}$, then the conditional distribution of $\Delta\varepsilon_k$ given ΔX_k is a symmetric finite mixture distribution, i.e., $f_{\Delta\varepsilon_k|\Delta X_k}(t) = \pi_{0k} f_0(t, \sigma_0^2|\Delta X_k) + (1 - \pi_{0k}) f_1(t, \sigma_1^2|\Delta X_k)$, where $f_0(t, \sigma_0^2|\Delta X_k)$ is symmetric unimodal probability distribution and $f_1(t, \sigma_1^2|\Delta X_k)$ is a symmetric

probability distribution function and σ_0^2, σ_1^2 are conditional variances related to ΔX_k, $k \in \mathcal{M}_\star$. Furthermore, there exists a given positive constant $\pi^\star \in (0, 1]$ such that $\pi_{0k} \geq \pi^\star$ for any $k \in \mathcal{M}_\star$.

(M2) $c_{\mathcal{M}_\star} = \min_{k \in \mathcal{M}_\star} E|X_k|$ is a positive constant and is free of p.

(M3) The predictors $\boldsymbol{X}_i = (X_{i1},\ \ldots,\ X_{ip})^T$ and the error term ε_i are independent, $i = 1,\ 2,\ \ldots,\ n$.

Theorem 2.4.1 *Under the marginally symmetric condition (M1)-(M3) and the condition of Theorem 3 in Fan and Song (2010), i.e., for $k \in \mathcal{M}_\star$,*

$$|Cov(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star),\ X_k)| \geq C_1 n^{-\kappa}$$

where C_1 is a positive constant and $\kappa < 1/2$. After using 2-quantile and $l-$quantiles to discretize the response Y and the predictor X_k, we have

(1) at least one $\widetilde{X}_{k_i}$ such that

$$|Cov(\widetilde{Y},\ \widetilde{X}_{k_i})| \geq C_2 n^{-\kappa}$$

for some positive constant C_2.

(2) Furthermore,

$$\min_{k \in \mathcal{M}_\star} \widetilde{L}_k^\star \geq C_3 n^{-2\kappa}$$

for some positive constant C_3 and $\widetilde{L}_k^\star$ is the corresponding increments of the log-likelihood after discretization.

Theorem 2.4.1 ensures that if the original variables are associated with the response or other variables, they are also related to each other after discretization. Therefore, discretization provides us another efficient way to screen the variable in the generalized linear models. We can replace SIS by Discretized SIS (DSIS). It stimulates us to apply discretization to the interaction pursuit. In next subsection, we get the similar result between SSI and discretized SSI (DSSI).

2.4.2 Relationship between SSI and DSSI

In this section, we consider the generalized linear model (2.2) with interaction terms and study the relationship SSI and DSSI. We follow the above assumption of our model and furthermore, suppose that $E(Y) = 0$ and $E(Y^2) = 1$.

Assume that S^{X_i}, S^{X_j} and S^Y are the support sets of variables X_i, X_j and Y, respectively. Denote that $\{P_s^{X_i}\}_{s=1}^{l_1}$, $\{P_t^{X_j}\}_{t=1}^{l_2}$and $\{P_k^Y\}_{k=1}^m$ are partitions of their supports, which means that

$$\bigcup_{s=1}^{l} P_s^{X_i} = S^{X_i} \quad \text{and} \quad P_{s_1}^{X_i} \bigcap P_{s_2}^{X_i} = \emptyset \ \text{ for } s_1 \neq s_2;$$

$$\bigcup_{t=1}^{l} P_t^{X_j} = S^{X_j} \quad \text{and} \quad P_{t_1}^{X_j} \bigcap P_{t_2}^{X_j} = \emptyset \ \text{ for } t_1 \neq t_2;$$

and

$$\bigcup_{k=1}^{m} P_k^{Y} = S^{Y} \quad \text{and} \quad P_{k_1}^{Y} \bigcap P_{k_2}^{Y} = \emptyset \ \text{ for } k_1 \neq k_2;$$

where l_1, l_2 and m are positive constants. Here, we still consider the l_1-quantiles, l_2-quantiles and $m-$quantiles as the break points for the partitions of variables X_i, X_j and Y, respectively. Define

$$\widetilde{X}_i = \begin{cases} 0, & X_i \in P_1^{X_i} \\ 1, & X_i \in P_2^{X_i} \\ \vdots & \vdots \\ l_1 - 1, & X_i \in P_{l_1}^{X_i} \end{cases} \quad \text{and} \quad \widetilde{X}_j = \begin{cases} 0, & X_j \in P_1^{X_j} \\ 1, & X_j \in P_2^{X_j} \\ \vdots & \vdots \\ l_2 - 1, & X_j \in P_{l_2}^{X_j} \end{cases}.$$

Furthermore, denote that

$$\widetilde{X}^{ij} = \begin{cases} 0, & X_i \in P_1^{X_i} \ \text{ and } \ X_j \in P_1^{X_j} \\ 1, & X_i \in P_1^{X_i} \ \text{ and } \ X_j \in P_2^{X_j} \\ \vdots & \vdots \\ l_1 * l_2 - 1, & X_i \in P_{l_1}^{X_i} \ \text{ and } \ X_j \in P_{l_2}^{X_j} \end{cases}$$

And also, we define that the discretized response $\widetilde{Y}$,

$$\widetilde{Y} = \begin{cases} 0, & Y \in P_1^Y \\ 1, & Y \in P_2^Y \\ \vdots & \vdots \\ m-1, & Y \in P_m^Y \end{cases}.$$

Hence, we have the new categorical predictor $\widetilde{X}_i$, $\widetilde{X}_j$ and response $\widetilde{Y}$, respectively. And also, we get the new interaction variable $\widetilde{X}^{ij}$. Furthermore, denote that

$$\widetilde{X}_{st}^{ij} = I\left(\left\{X_i \in P_s^{X_i}\right\} \bigcap \left\{X_j \in P_t^{X_j}\right\}\right), \quad 1 \le s \le l_1, \;\; 1 \le t \le l_2$$

and $\widetilde{Y}_j = I(Y \in P_j^Y)$, $1 \le j \le m$, where $I(\cdot)$ is the indicator function. After discretization, the new increment of log-likelihood function in population version is defined as

$$\widetilde{L}_{ij}^{\star} = E\{l(\widetilde{\boldsymbol{X}}_{i,j}^T \widetilde{\boldsymbol{\beta}}_{i,j}^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_{ij}^T \widetilde{\boldsymbol{\beta}}_{ij}^M, \widetilde{Y})\}, \quad 1 \le i < j \le p.$$

Remark 2.4.2 *Here, we still consider continuous response Y and use 2-quantile (median) to split the response Y, that is, $m = 2$ and*

$$\widetilde{Y} = \begin{cases} 0, & Y \le M_d(Y) \\ 1, & Y > M_d(Y) \end{cases},$$

where $M_d(Y)$ is the median of the response Y. To prove the following theorem, we will first give the proof for the case $m = 2$ and $l_1 = l_2 = 2$, and then extend the case $l_1 = l_2 = 2$ to the case $l_1 \ge 3$ and $l_2 \ge 3$.

Similar to the last part, we still need some marginally symmetric conditions. Let $\zeta_{ij} = Y - b'(\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M)$, and denote that $(Y_1, X_{1i}, X_{1j}, X_{1ij}, \zeta_{1ij})$, $(Y_2, X_{2i}, X_{2j}, X_{2ij}, \zeta_{2ij})$ be the independent copies of $(Y, X_i, X_j, X_{ij}, \zeta_{ij})$. We further centralize ζ_{ij} and denote that $\rho_{ij} = \mathrm{Cov}(\zeta_{ij}, X_{ij})/\sqrt{\mathrm{Var}(\zeta_{ij})\mathrm{Var}(X_{ij})}$.

(M1′) Denote $\Delta\varepsilon_{ij} = \zeta_{1ij} - \zeta_{2ij} - \rho_{ij}(X_{1ij} - X_{2ij})$ and $\Delta X_{ij} = X_{1ij} - X_{2ij}$, then the conditional distribution of $\Delta\varepsilon_{ij}$ given ΔX_{ij} is a symmetric finite mixture distribution,

i.e., $f_{\Delta\varepsilon_{ij}|\Delta X_{ij}}(t) = \pi_{0ij} f_0(t, \sigma_0^2|\Delta X_{ij}) + (1-\pi_{0ij}) f_1(t, \sigma_1^2|\Delta X_{ij})$, where $f_0(t, \sigma_0^2|\Delta X_{ij})$ is symmetric unimodal probability distribution and $f_1(t, \sigma_1^2|\Delta X_{ij})$ is a symmetric probability distribution function and σ_0^2, σ_1^2 are conditional variances related to ΔX_{ij}, $i, j \in \mathcal{N}_\star$. Furthermore, there exists a given positive constant $\pi^\star \in (0, 1]$ such that $\pi_{0ij} \geq \pi^\star$ for any $i, j \in \mathcal{N}_\star$.

(M2′) $c_{\mathcal{N}_\star} = \min_{i,j\in\mathcal{N}_\star} E|X_{ij}|$ is a positive constant and is free of p.

(M3′) The predictors $\boldsymbol{X} = (X_1, \ \ldots, \ X_p)^T$ and the error term ε are independent.

Remark 2.4.3 *In fact, the marginally symmetric condition (M1)' is also easily satisfied. Denote that $\varepsilon_{ij} = \zeta_{ij} - \rho_{ij} X_{ij}$. A special case is that under the linear model, the conditional distribution of ε_{ij} given X_{ij} does not dependent on X_{ij} and it has K modes, where K is finite. It implies that the conditional distribution $\varepsilon_{ij}|X_{ij}$ is same as the distribution of ε_{ij}. Suppose that ε_{1ij}, ε_{2ij} follow a distribution $f_\varepsilon(t)$ with K modes, that is, $f_\varepsilon(t) = \sum_{k=1}^K \pi_k f_k(t)$, where $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$. Moreover, assume that $f_{lm}^\star(t)$, $1 \leq l, m \leq K$, are the distributions of the difference $Z_l - Z_m$, where Z_l and Z_m are independent and follow the distributions $f_l(t)$ and $f_m(t)$, respectively. Therefore, the distribution of $\Delta\varepsilon_{ij} = \varepsilon_{1ij} - \varepsilon_{2ij}$ can be expressed as*

$$\begin{aligned} f_{\Delta\varepsilon}(t) &= \sum_l \sum_m \pi_l \pi_m f_{lm}^\star(t) = \sum_l \pi_l^2 f_{ll}^\star(t) + \sum_{l\neq m} \pi_l \pi_m f_{lm}^\star(t) \\ &= \big(\sum_l \pi_l^2\big) \sum_l \frac{\pi_l^2}{\sum_l \pi_l^2} f_{ll}^\star(t) + \big(1 - \sum_l \pi_l^2\big) \sum_{l\neq m} \frac{\pi_l \pi_m}{1 - \sum_l \pi_l^2} f_{lm}^\star(t) \\ &\triangleq \pi_0^\star f_0^\star(t) + (1 - \pi_0^\star) f_1^\star(t). \end{aligned}$$

Obviously, $f_{ll}^\star(t)$ are symmetric unimodal distributions because of the unimodal distributions $f_l(t)$, and then $f_0^\star(t)$ is symmetric and unimodal. And $f_1^\star(t)$ is a symmetric and multimodal density function. Moreover, $\pi_0^\star = \sum_l \pi_l^2 \geq (\sum_l \pi_l^2)^2/K = 1/K$.

Theorem 2.4.2 *Under the marginally symmetric conditions (M1′)−(M3′) and the condition: for $i, j \in \mathcal{N}_\star$,*

$$|Cov_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M)| \geq c_1 n^{-\kappa}$$

where c_1 is a positive constant and $\kappa < 1/4$. After using 2-quantile, l_1-quantiles and

l_2-quantiles to discretize the response Y and the predictors X_i, X_j, we obtain

(1) at least one $\widetilde{X}_{st}^{ij}$ such that

$$|Cov_L(\widetilde{Y},\ \widetilde{X}_{st}^{ij}|\widetilde{\boldsymbol{X}}_{i,j}^T\widetilde{\boldsymbol{\beta}}_{i,j}^M)| \geq c_{10}n^{-\kappa}$$

for some positive constant c_{10}.

(2) Furthermore,

$$\min_{i,j\in\mathcal{N}_\star} \widetilde{L}_{ij}^\star \geq c_{11}n^{-2\kappa}$$

for some positive constant c_{11} and $\widetilde{L}_{ij}^\star$ is the corresponding increments of the log-likelihood after discretization.

Theorem 2.4.2 claims that the important interaction terms are still significant after discretization under some certain conditons. Consequently, we can use more efficient algorithm Discretized SSI to detect efficiently the important interaction effects.

2.5 Numerical Studies I

In this section, we will demonstrate the performance of our method-SSI about the interaction screening on the simulated data and verify the theoretical results. Here, the methods-*xyz* (Thanei et al. [2016]), RAMP (Hao et al. [2016]) and IP (Fan et al. [2016]) are taken into account to compare their performance on the estimation and prediction. Furthermore, the finite sample performance of discretized SSI (DSSI) is also examined.

2.5.1 The setup of Simulation studies

In the first part of simulation study, the data $(\boldsymbol{X}_1^T, Y_1),\ (\boldsymbol{X}_2^T, Y_2), \ldots, (\boldsymbol{X}_n^T, Y_n)$ are independent copies of the population $(\boldsymbol{X}^T, Y)$, where the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ is a normal distribution or binomial distribution. And also we design three types of interaction models with different heredity assumptions:

(1) strong heredity assumption: the interaction effect is significant and the corresponding main effects are also included in the model;

(2) weak heredity assumption: the interaction effect is significant and at least one of

the corresponding main effects are included in the model;
(3) anti-heredity assumption: the interaction effect is significant and the corresponding main effects are not included in the model.

Following the setting of Fan and Song [2010], the predictors are generated from

$$X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1 + a_j^2}},$$

where ε and $\{\varepsilon_j\}_{j=1}^{\lfloor p/3 \rfloor}$ are independent and identically distributed standard normal random variables; $\{\varepsilon_j\}_{j=\lfloor p/3 \rfloor+1}^{\lfloor 2p/3 \rfloor}$ are independent and identically distributed, and follow a double exponential distribution with the location parameter 0 and scale parameter 1; and $\{\varepsilon_j\}_{j=\lfloor 2p/3 \rfloor+1}^{p}$ are independent and identically distributed and follow a mixture normal distribution with two components $N(-1, 1)$, $N(1, 0.5)$ and equal mixture proportion. The covariates are standardized. The constants $\{a_j\}_{j=1}^{r}$ are identical and taken as some values such that the correlation $\rho = \text{corr}(X_i, X_j) = 0, 0.2, 0.4, 0.6, 0.8$ among the first r predictors, and $a_j = 0$ for $j > r$.

We fix the number of predictors with $p = 400$, 2000 and 5000, and consider the size of nonzero interaction coefficients as $s = 3$, and present the numerical result with $r = 15$. All methods are evaluated by aggregating the median minimal model size (MMMS) of the selected model and its associated RSD, which is the interquartile range (IQR) divided by 1.34, by considering 100 simulated results. Time (in seconds) represents the average time of computing the test statistics of all the interaction terms in 100 simulated data sets. Denote that Pr be the percentage of the numbers of results that include the true model by using the method $RAMP$ and xyz. Since the method $RAMP$ is considered under the strong and weak heredity assumption in the package "RAMP", and anti-heredity is not covered in that package, we use the function RAMP with weak assumption (RAMP-w) to screen the interaction terms in the model without heredity assumption. The tuning parameter is selected by AIC. For xyz algorithms, we use the projection times $L = 10$ and the number of interaction terms selected is 500, which is the largest number of the interaction terms that we can get by this algorithm in the package "xyz". The linear model and logistic model are chosen to take the simulation studies.

2.5.2 Example 1-Linear Model

The generated data are n independent and identically distributed copies of the population $(\boldsymbol{X}^T, Y)$, in which the response Y follows a linear model with $Y = \boldsymbol{X}^T\boldsymbol{\beta}^\star + \varepsilon$. We follow the above predictors' setting and consider the following three interaction models linking the predictors X_j's to the response Y.

(M1): $Y = \sum_{i=1}^{5} \beta_i X_i + \beta_{13} X_1 X_3 + \beta_{24} X_2 X_4 + \beta_{35} X_3 X_5 + \varepsilon$,

(M2): $Y = \sum_{i=1}^{5} \beta_i X_i + \beta_{610} X_6 X_{10} + \beta_{814} X_8 X_{14} + \beta_{1215} X_{12} X_{15} + \varepsilon$,

(M3): $Y = \sum_{i=1}^{5} \beta_i X_i + \beta_{16} X_1 X_6 + \beta_{310} X_3 X_{10} + \beta_{515} X_5 X_{15} + \varepsilon$;

where the coefficients of main effect and interaction effect are 1 and 2 respectively. For models (M1)-(M3), all of them include the five main terms X_1, X_2, X_3, X_4, X_5. Three interaction terms X_1X_3, X_2X_4 and X_3X_5 are included in the model (M1), which satisfies the strong heredity assumption. In the model (M2), only pure interaction effects exist (anti-heredity), i.e., $X_6X_{10},\ X_8X_{14},\ X_{12}X_{15}$. The weak heredity assumption is included in the models (M3). For interaction terms $X_1X_6,\ X_3X_{10},\ X_5X_{15}$, the corresponding main effect X_1, X_3 and X_5 are in the model (M3). Here, the sample size is taken as 100 and 200. The results of these three models by using three methods can be found in the Table 2.1-2.3.

Based on these results in the above tables, among all of these designed scenarios, our method SSI performs well in linear model with our setting. The xyz algorithm is the least time consuming but not be of the best performance. By comparing SSI with RAMP, the time of SSI is less than that of RAMP when p is not large, while RAMP is more efficient than SSI when p is sufficiently large. However, Pr is not always 1 for RAMP, which implies that the results from RAMP cannot cover all important interaction terms sometimes. Especially, for model (M2), Pr of RAMP is the worst one in three methods (M1)-(M3). For MMMS, RAMP and SSI are able to reduce the number of interaction terms to less than n. The performance of xyz is the worst one. And also, RAMP performs a little better than SSI does about MMMS, but the results of RAMP are not better in Model (M2) that does not have the heredity assumption. Furthermore, sometimes RAMP cannot select any one of interaction terms, for example, $p = 400$, $\rho = 0$ in model (M2). Moreover, these three methods

Table 2.1: The MMMS and the associated RSD (in the parenthesis) for the linear model with $p = 400$

SS1							
ρ	n	MMMS	Time	MMMS	Time	MMMS	Time
		(M1)		(M2)		(M3)	
0.0	100	12(50)	1.73	3(0)	1.87	3(0)	2.20
0.2	100	4(6)	1.73	3(0)	1.92	3(0)	1.89
0.4	100	5(5)	1.63	4(3)	1.96	4(2)	1.87
0.6	100	8(5)	1.60	10(9)	1.95	8(5)	1.88
0.8	100	12.5(10)	1.61	17(11)	1.88	13(7)	1.87
0.0	200	3(0)	4.17	3(0)	3.20	3(0)	2.46
0.2	200	3(0)	4.29	3(0)	3.67	3(0)	2.44
0.4	200	3(1)	3.27	5(3)	3.82	5(4)	2.41
0.6	200	6(5)	3.15	12(9)	3.56	8(5)	2.41
0.8	200	9(5)	3.15	17(13)	3.70	13(7)	2.41

xyz										
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M1)			(M2)			(M3)	
0.0	100	0.04	198.5(122)	0.0128	0.03	149(126)	0.0142	0.04	286(130)	0.0210
0.2	100	0.08	133(65)	0.0121	0.06	242.5(147)	0.0131	0.01	73(0)	0.0164
0.4	100	0.25	98(158)	0.0126	0.15	84(56)	0.0131	0.23	145(120)	0.0131
0.6	100	0.54	161.5(102)	0.0132	0.47	181(90)	0.0145	0.35	193(77)	0.0145
0.8	100	0.79	116(59)	0.0139	0.74	129.5(58)	0.0143	0.73	123(57)	0.0154
0	200	0.07	16(6)	0.0121	0.04	14(12)	0.0096	0.08	23(36)	0.0095
0.2	200	0.11	20(47)	0.0119	0.04	26(43)	0.0099	0.06	98(101)	0.0099
0.4	200	0.44	123(128)	0.0118	0.19	97(145)	0.0101	0.25	128(137)	0.0101
0.6	200	0.63	125(77)	0.0132	0.47	130(63)	0.0104	0.56	148.5(63)	0.0107
0.8	200	0.93	103(38)	0.0130	0.91	100(44)	0.0112	0.89	103(49)	0.0118
RAMP										
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M1)			(M2)			(M3)	
0	100	0.53	3(0)	6.086	0	0(0)	11.395	0.47	6(7)	10.866
0.2	100	0.39	3(0)	5.407	0.10	14(7)	16.018	0.84	4(1)	12.643
0.4	100	0.24	3(0)	5.886	0.21	7(3)	16.347	0.94	4(1)	12.007
0.6	100	0.03	3(0)	6.309	0.24	10(7)	25.665	0.94	5(1)	13.949
0.8	100	0	0(0)	6.490	0.32	9(9)	23.991	0.96	5(2)	20.724
0	200	0.92	3(0)	19.459	0	0(0)	41.933	0.97	3(0)	22.439
0.2	200	0.80	3(0)	14.554	0.12	3.5(8)	47.947	1	3(0)	27.983
0.4	200	0.68	3(0)	12.545	0.20	6(10)	26.479	1	3(1)	25.656
0.6	200	0.45	3(1)	14.363	0.21	6(6)	23.101	1	4(1)	26.679
0.8	200	0.08	4(0)	14.467	0.24	6.5(4)	21.594	1	4(1)	21.922

Table 2.2: The MMMS and the associated RSD (in the parenthesis) for the linear model with $p = 2000$

SS1							
ρ	n	MMMS	Time	MMMS	Time	MMMS	Time
		(M1)		(M2)		M3)	
0	100	93.5(786)	61.76	433(1651)	43.65	123.5(528)	46.76
0.2	100	21(124)	61.53	152.5(642)	52.16	96.5(463)	48.96
0.4	100	7(9)	63.43	13(24)	66.22	10(12)	61.02
0.6	100	9(6)	62.11	21.5(17)	62.41	15.5(10)	60.91
0.8	100	12.5(10)	59.76	34.5(19)	61.54	23(14)	57.12
0	200	3(0)	69.28	3(0)	64.80	3(0)	67.12
0.2	200	3(0)	57.28	3(0)	59.55	3(0)	59.63
0.4	200	4(1)	51.28	4(5)	59.92	4(2)	57.98
0.6	200	6(4)	60.33	10(8)	79.24	9(7)	79.45
0.8	200	7.5(5)	59.42	17(14)	60.86	13(7)	61.55

xyz										
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M1)			(M2)			(M3)	
0	100	0.00	0(0)	0.0624	0.00	0(0)	0.0888	0.00	0(0)	0.0762
0.2	100	0.00	0(0)	0.0639	0.00	0(0)	0.0959	0.01	309(0)	0.0743
0.4	100	0.02	154.5(34)	0.0654	0.02	113.5(32)	0.0983	0.03	294(38)	0.0762
0.6	100	0.15	231(162)	0.0685	0.07	369(57)	0.1015	0.12	335.5(81)	0.078
0.8	100	0.38	193.5(110)	0.0675	0.31	200(123)	0.1032	0.35	183(130)	0.0793
0.0	200	0.00	0(0)	0.0532	0.00	0(0)	0.0511	0.01	151(0)	0.0500
0.2	200	0.03	136(62)	0.0524	0.00	0(0)	0.0689	0.02	117.5(78)	0.0504
0.4	200	0.10	103.5(108)	0.0619	0.03	21(134)	0.0701	0.03	117(163)	0.0511
0.6	200	0.21	231(118)	0.0568	0.16	224.5(196)	0.0618	0.16	386(163)	0.0526
0.8	200	0.65	209(89)	0.0549	0.60	210.5(142)	0.0544	0.62	223.5(102)	0.0546
RAMP										
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M1)			(M2)			(M3)	
0	100	0.14	3(0)	6.62	0	0(0)	36.62	0.14	11.5(5)	37.63
0.2	100	0.27	3(0)	6.61	0.02	11(5)	38.86	0.66	4(3)	42.64
0.4	100	0.14	3(0)	6.53	0.11	6(5)	44.23	0.84	4(2)	59.34
0.6	100	0.03	3(0)	6.05	0.21	12(11)	60.36	0.93	5(1)	62.62
0.8	100	0	0(0)	5.57	0.20	13.5(6)	44.91	0.97	6(2)	42.74
0	200	0.81	3(0)	10.56	0	0(0)	126.08	0.88	3(1)	118.35
0.2	200	0.82	3(0)	15.91	0.05	3(1)	194.92	0.99	3(0)	225.87
0.4	200	0.52	3(0)	17.52	0.16	3.5(1)	142.36	1	3(1)	196.26
0.6	200	0.27	3(1)	16.60	0.14	13.5 (10)	141.69	1	4(1)	136.17
0.8	200	0	0(0)	14.65	0.26	8(17)	106.72	1	4(1)	106.11

Table 2.3: The MMMS and the associated RSD (in the parenthesis) for the linear model with $p = 5000$

				SSI			
ρ	n	MMMS	Time	MMMS	Time	MMMS	Time
		(M1)		(M2)		(M3)	
0	100	628.5(4903)	869.8	1801.5 (13480)	431.7	986.5(4562)	248.1
0.2	100	87.5(708)	824.0	815(8283)	251.7	330.5(1569)	248.2
0.4	100	6(8)	388.3	29.5(141)	254.4	11(27)	248.6
0.6	100	9(7)	385.1	24(21)	255.5	14(15)	383.7
0.8	100	10(11)	386.1	27.5(18)	248.7	23(15)	444.3
0	200	3(1)	394.4	3(0)	424.1	3(0)	424.1
0.2	200	3(0)	405.4	3(1)	518.7	3(1)	518.7
0.4	200	4(1)	806.1	5(4)	1191.5	4(3)	1191.4
0.6	200	6(4)	803.9	9.5(11)	1152.5	10(8)	1146.8
0.8	200	7(5)	744.5	16(11)	819.2	14(10)	836.0

					xyz					
rho	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M1)			(M2)			(M3)	
0	100	0.00	0(0)	0.2653	0.00	0(0)	0.2271	0.00	0(0)	0.1704
0.2	100	0.00	0(0)	0.2115	0.00	0(0)	0.268	0.00	0(0)	0.1741
0.4	100	0.00	0(0)	0.2847	0.00	0(0)	0.2087	0.00	0(0)	0.1789
0.6	100	0.04	175.5(156)	0.2898	0.01	251(0)	0.2084	0.01	419(0)	0.1813
0.8	100	0.07	254(141)	0.2968	0.03	389(82)	0.213	0.05	225(87)	0.1825
0	200	0.00	0(0)	0.1615	0.00	0(0)	0.1343	0.00	0(0)	0.138
0.2	200	0.00	0(0)	0.1578	0.00	0(0)	0.1331	0.01	447(0)	0.1462
0.4	200	0.05	13(4)	0.1571	0.01	227(0)	0.1354	0.01	27(0)	0.1536
0.6	200	0.09	269(235)	0.1602	0.08	17(13)	0.1388	0.02	27.5(9)	0.1569
0.8	200	0.22	252.5(169)	0.1516	0.23	289(129)	0.1417	0.17	276(123)	0.1605
					RAMP					
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M1)			(M2)			(M3)	
0	100	0.07	3(0)	4.81	0	0(0)	66.47	0.03	25(16)	63.83
0.2	100	0.22	3(0)	5.44	0.01	13(0)	73.07	0.51	6(4)	79.14
0.4	100	0.08	3(0)	5.47	0.12	5(6)	92.91	0.83	5(3)	90.24
0.6	100	0	0(0)	5.59	0.17	7(8)	99.77	0.98	5(2)	94.38
0.8	100	0	0(0)	5.47	0.19	8(4)	96.52	0.87	6(2)	95.26
0	200	0.79	3(0)	14.42	0	0(0)	222.57	0.72	3(1)	315.45
0.2	200	0.71	3(0)	14.28	0.02	4.5(1)	301.78	0.95	3(0)	245.21
0.4	200	0.47	3(0)	13.45	0.18	4(8)	321.22	1	3(1)	261.86
0.6	200	0.14	3(0)	12.62	0.19	6(9)	284.26	1	4(1)	264.76
0.8	200	0.01	4(0)	12.38	0.29	7(15)	351.89	1	5(1)	308.32

will perform better when the sample size n becomes larger.

2.5.3 Example 2-Logistic Model

In this section, the generated data are n independent and identically distributed copies of the population $(\boldsymbol{X}^T, Y)$, where the conditional distribution of the variable Y given $\boldsymbol{X} = \boldsymbol{x}$ is a binomial distribution with

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{X}^T\boldsymbol{\beta}^\star.$$

We use the same setup of covariates and the same values $\boldsymbol{\beta}^\star$ as that in the linear model, and consider the following three interaction models.

(M4): $\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \sum_{i=1}^{5}\beta_i X_i + \beta_{13}X_1X_3 + \beta_{24}X_2X_4 + \beta_{35}X_3X_5,$

(M5): $\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \sum_{i=1}^{5}\beta_i X_i + \beta_{610}X_6X_{10} + \beta_{814}X_8X_{14} + \beta_{1215}X_{12}X_{15},$

(M6): $\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \sum_{i=1}^{5}\beta_i X_i + \beta_{16}X_1X_6 + \beta_{310}X_3X_{10} + \beta_{515}X_5X_{15};$

Here, the coefficients of main effects and interaction effects are still chosen as 1 and 2 respectively. All three models have five main effects X_1, X_2, X_3, X_4 and X_5. The model (M4) has the strong heredity assumption with interaction terms X_1X_3, X_2X_4 and X_3X_5 ; the anti heredity assumption is satisfied by the model (M5) with cross terms X_6X_{10}, X_8X_{14}, $X_{12}X_{15}$; and the models (M6) is under the weak heredity assumption, which has interaction terms X_1X_6, X_3X_{10}, X_5X_{15}. The sample size is defined as 200 and 300. The results of three methods applied into (M4)-(M6) are presented in the tables 2.4-2.6.

Tables 2.4-2.6 list all results about three methods in the logistic models. By these results, we can draw some conclusions: (1) From the perspective of efficiency, the xyz algorithm is still best and the time consumed by SSI is a little more than that of RAMP; (2) By considering Pr, RAMP outperforms xyz but its Pr is still a little low; (3) For MMMS, we choose the projection numbers $L = 10$, 50 and 100, the results of xyz do not involve any one of the interaction terms. In the model (M5), RAMP provides the same results. When the dimension p is small, SSI performs better than other two methods. When $p = 2000$ or 5000, the performance of our method is not good when the sample size is only 200 compared with RAMP. By increasing the

Table 2.4: The MMMS and the associated RSD (in the parenthesis) for logistic model with $p = 400$

SS1							
ρ	n	MMMS	Time	MMMS	Time	MMMS	Time
		(M4)		(M5)		(M6)	
0	200	9(46)	11.88	7(12)	9.31	4(6)	9.48
0.2	200	3(1)	9.96	8(24)	10.04	5(8)	10.55
0.4	200	3(1)	10.57	5(5)	10.33	3(1)	10.95
0.6	200	4(3)	10.75	10(10)	10.44	6(6)	11.14
0.8	200	11(9)	10.82	32.5(20)	10.42	21(15)	11.09
0	300	3(1)	16.10	3(0)	16.016	3(0)	15.99
0.2	300	3(0)	15.53	3(0)	17.95	3(0)	17.29
0.4	300	3(0)	16.10	3(1)	17.97	3(0)	18.03
0.6	300	3(1)	18.03	5.5(4)	18.21	4(2)	17.41
0.8	300	7(7)	18.23	22(16)	18.48	14(12)	17.97

xyz										
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M4)			(M5)			(M6)	
0	200	0	0(0)	0.0169	0	0(0)	0.0167	0	0(0)	0.0170
0.2	200	0	0(0)	0.0165	0	0(0)	0.0177	0	0(0)	0.0165
0.4	200	0.02	324.5(6)	0.0169	0.01	335(0)	0.0163	0.01	338(0)	0.0168
0.6	200	0.05	365(66)	0.0176	0.04	366.5(17)	0.0172	0.04	368.5(46)	0.0174
0.8	200	0.15	375(52)	0.0183	0.23	347(48)	0.0168	0.2	345(22)	0.0174
0	300	0	0(0)	0.0216	0	0(0)	0.0288	0	0(0)	0.0215
0.2	300	0	0(0)	0.0211	0	0(0)	0.0217	0	0(0)	0.0211
0.4	300	0	0(0)	0.0231	0.02	371(17)	0.0204	0	0(0)	0.0205
0.6	300	0.07	339(38)	0.0204	0.02	389(4)	0.0229	0.05	341(14)	0.0205
0.8	300	0.19	330(30)	0.0207	0.17	370(46)	0.0207	0.2	363(44)	0.0206
RAMP										
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M4)			(M5)			(M6)	
0	200	0.33	3(1)	24.59	0	0(0)	65.09	0.16	8.5(5)	62.39
0.2	200	0.07	3(0)	25.58	0	0(0)	59.82	0.35	6(4)	56.83
0.4	200	0.00	6(0)	20.09	0	0(0)	53.15	0.13	5(7)	55.15
0.6	200	0.00	0(0)	31.07	0	0(0)	46.31	0.04	7(7)	45.18
0.8	200	0.00	0(0)	18.76	0	0(0)	40.09	0.00	0(0)	41.46
0	300	0.82	3(0)	33.94	0	0(0)	131.91	0.56	8(7)	113.81
0.2	300	0.16	3(1)	35.16	0	0(0)	112.46	0.57	4(3)	95.82
0.4	300	0	0(0)	31.91	0	0(0)	78.00	0.41	6(6)	79.84
0.6	300	0	0(0)	30.27	0	0(0)	71.98	0.11	9(8)	76.69
0.8	300	0	0(0)	26.95	0	0(0)	107.48	0.02	17.5(9)	86.15

Table 2.5: The MMMS and the associated RSD (in the parenthesis) for logistic model with $p = 2000$

					SS1				
ρ	n	MMMS	Time	MMMS	Time	MMMS	Time		
		(M4)		(M5)		(M6)			
0	200	153(677)	238.80	107(383)	247.70	72.5(378)	244.78		
0.2	200	9(34)	242.00	73(289)	265.83	42.5(325)	261.22		
0.4	200	3(2)	262.72	8.5(25)	272.73	4.5(9)	269.12		
0.6	200	4(3)	267.42	10(10)	271.16	6.5(6)	272.65		
0.8	200	12(10)	273.37	28.5(24)	265.56	22(19)	271.2		
0	300	5.5(24)	479.18	4(5)	821.08	3(1)	470.92		
0.2	300	3(0)	465.85	3(5)	951.26	3(1)	638.17		
0.4	300	3(0)	499.65	3(1)	509.85	3(0)	933.61		
0.6	300	3(1)	529.42	6(4)	509.79	4(4)	976.17		
0.8	300	8.5(7)	984.82	22(20)	492.48	14(13)	996.39		

						xyz				
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M4)			(M5)			(M6)	
0	200	0	0(0)	0.0700	0	0(0)	0.0721	0	0(0)	0.0623
0.2	200	0	0(0)	0.0687	0	0(0)	0.0838	0	0(0)	0.0621
0.4	200	0	0(0)	0.0781	0	0(0)	0.0798	0	0(0)	0.0626
0.6	200	0	0(0)	0.0680	0	0(0)	0.0699	0	0(0)	0.0626
0.8	200	0	0(0)	0.0629	0	0(0)	0.0645	0	0(0)	0.0626
0	300	0	0(0)	0.0784	0	0(0)	0.0761	0	0(0)	0.1161
0.2	300	0	0(0)	0.0811	0	0(0)	0.0762	0	0(0)	0.1190
0.4	300	0	0(0)	0.0858	0	0(0)	0.0769	0	0(0)	0.1039
0.6	300	0	0(0)	0.0879	0	0(0)	0.0770	0	0(0)	0.0900
0.8	300	0	0(0)	0.0942	0	0(0)	0.0780	0	0(0)	0.0776
						RAMP				
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M4)			(M5)			(M6)	
0	200	0.06	3(0)	22.38	0	0(0)	169.32	0.05	19(4)	206.53
0.2	200	0.02	3(0)	20.22	0	0(0)	200.53	0.21	6(4)	193.03
0.4	200	0.00	0(0)	16.04	0	0(0)	175.64	0.03	4(4)	221.53
0.6	200	0.00	0(0)	11.28	0	0(0)	239.09	0.02	20.5(4)	182.39
0.8	200	0.00	0(0)	11.67	0	0(0)	187.78	0.00	0(0)	152.55
0	300	0.43	3(0)	58.5	0	0(0)	314.23	0.22	13(10)	303.25
0.2	300	0.13	3(0)	59.61	0	0(0)	228.41	0.34	5(3)	220.29
0.4	300	0	0(0)	37.53	0	0(0)	213.25	0.26	4.5(7)	208.62
0.6	300	0	0(0)	23.32	0	0(0)	199.5	0.03	8(1)	199.44
0.8	300	0	0(0)	25.11	0	0(0)	187.99	0	0(0)	195.71

Table 2.6: The MMMS and the associated RSD (in the parenthesis) for logistic model with $p = 5000$

				SS1				
ρ	n	MMMS	Time	MMMS	Time	MMMS	Time	
		(M4)		(M5)		(M6)		
0	200	2711.5(12891)	1551.9	964.5(2845)	1493.8	325 (2440)	1528.0	
0.2	200	24.5(158)	1686.2	639.5(3758)	1759.9	252.5(1746)	1720.3	
0.4	200	4(5)	1798.8	35.5(343)	1834.9	11.5(53)	1805.1	
0.6	200	5(3)	1669.4	11(12)	1697.3	7(8)	1667.1	
0.8	200	13(9)	1715.9	28.5(18)	1717.3	23(14)	1708.1	
0	300	19(93)	3733.0	4(11)	2317.9	3(5)	2343.2	
0.2	300	3(1)	2461.2	5(19)	2400.0	4(7)	2407.4	
0.4	300	3(0)	2259.1	3(1)	2398.9	3(1)	2413.0	
0.6	300	3(1)	2299.1	5(5)	2685.1	4(2)	2698.2	
0.8	300	8(6)	2477.5	19.5(14)	2999.2	13(11)	3008.4	

					xyz					
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M4)			(M5)			(M6)	
0	200	0	0(0)	0.1863	0	0(0)	0.1683	0	0(0)	0.1543
0.2	200	0	0(0)	0.2007	0	0(0)	0.1707	0	0(0)	0.1533
0.4	200	0	0(0)	0.1885	0	0(0)	0.1638	0	0(0)	0.1545
0.6	200	0	0(0)	0.1781	0	0(0)	0.2014	0	0(0)	0.1561
0.8	200	0	0(0)	0.1597	0	0(0)	0.1753	0	0(0)	0.1560
0	300	0	0(0)	0.1917	0	0(0)	0.2007	0	0(0)	0.2225
0.2	300	0	0(0)	0.1885	0	0(0)	0.2027	0	0(0)	0.2160
0.4	300	0	0(0)	0.1891	0	0(0)	0.2113	0	0(0)	0.2138
0.6	300	0	0(0)	0.1919	0	0(0)	0.2126	0	0(0)	0.2066
0.8	300	0	0(0)	0.1912	0	0(0)	0.2143	0	0(0)	0.1998
					RAMP					
ρ	n	Pr	MMMS	Time	Pr	MMMS	Time	Pr	MMMS	Time
			(M4)			(M5)			(M6)	
0	200	0.02	3(0)	58.87	0	0(0)	1003.4	0.00	0(0)	766.43
0.2	200	0.00	0(0)	68.98	0	0(0)	1276.2	0.04	11.5(9)	1185.6
0.4	200	0.01	3(0)	78.11	0	0(0)	1093.9	0.02	10.5(1)	1053.2
0.6	200	0.00	0(0)	58.68	0	0(0)	1016.1	0.01	10(0)	1058.1
0.8	200	0.00	0(0)	48.21	0	0(0)	833.9	0.00	0(0)	764.7
0	300	0.31	3(0)	119.71	0	0(0)	1902.00	0.16	9(7)	1822.5
0.2	300	0.02	3(0)	221.85	0	0(0)	2545.9	0.29	6(6)	2527.1
0.4	300	0	0(0)	184.68	0	0(0)	2321.5	0.19	7(6)	2292.1
0.6	300	0	0(0)	145.84	0	0(0)	1490.3	0.03	6(8)	1413.1
0.8	300	0	0(0)	111.50	0	0(0)	1465.1	0	0(0)	1474.7

sample size to 300, our method's performance awfully exceeds other two methods.

In summary, among all of these designed scenarios, our method SSI performs well although its time is a little more than other methods'.

2.5.4 Prediction Performance-SSI

In this part, we consider the performance of prediction in the linear model by comparing three methods: SSI, *xyz* and RAMP. We apply the out of sample R^2 to evaluate the prediction performance of these three methods. The out of sample R^2 is defined by

$$R^2 = 100\% \times \left\{ 1 - \frac{\sum (Y_i^* - \boldsymbol{X}_i^{*T} \hat{\boldsymbol{\beta}})^2}{\sum (Y_i^* - \bar{Y}^*)^2} \right\},$$

where $(\boldsymbol{X}_i^*, Y_i^*)$ is the testing data and $\hat{\boldsymbol{\beta}}$ is the estimate of the coefficient based on the training data. All methods are evaluated by summarizing the median of the out of sample R^2 of the selected model and its associated RSD, which is taken as the interquartile range (IQR) divided by 1.34. All of these values are based on 100 simulated results.

The setup of the predictors still follows the setting of Fan and Song [2010], just like the last section. All of the samples are generated by the linear models (M1)-(M3). Here, the dimension and the sample size of data sets are:

$$\{(p,\ n) : (400, 200), (400, 300), (2000, 300), (2000, 400), (5000, 300), (5000, 400)\}.$$

And ρ is taken as five values: 0, 0.2, 0.4, 0.6, 0.8. For each sample data set, we choose the $n_1 = 75\% * n$ of the data set as the training data and the remaining samples as the testing data. For the method SSI, we first select $n_1 - 1$ main effects, $n_1 - 1$ quadratic components and $n_1 - 1$ interaction terms, and then apply the LASSO penalty to these terms in the training set. As a result, by using the testing data, we calculate the out of sample R^2. For *xyz* algorithm, we directly apply the function "*xyz_regression*" to the training set and then get the out of sample R^2 based on the testing data. For the third method RAMP, we also use the function $RAMP$ with AIC criterion to select the interaction effects in the training set and compute R^2 by the testing data. By default, LASSO penalty is used in the RAMP. All results are shown in Table 2.7-2.9.

Table 2.7: The Median of out-of-sample R^2 and the associated RSD (in the parenthesis) for linear model with $p = 400$

			SSI			
	(M1)		(M2)		(M3)	
ρ	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
0	0.81010(0.0917)	0.89446(0.0209)	0.83067(0.0597)	0.89941(0.0215)	0.83940(0.0581)	0.89556(0.0203)
0.2	0.91994(0.0277)	0.94314(0.0197)	0.91664(0.0273)	0.93653(0.0136)	0.91472(0.0331)	0.93876(0.0131)
0.4	0.95067(0.0233)	0.96311(0.0121)	0.95345(0.0169)	0.96104(0.0099)	0.95283(0.0207)	0.95830(0.0114)
0.6	0.97275(0.0107)	0.97536(0.0073)	0.96947(0.0129)	0.97501(0.0090)	0.97200(0.0101)	0.97359(0.0078)
0.8	0.98152(0.0099)	0.98375(0.0061)	0.98081(0.0106)	0.98348(0.0057)	0.98086(0.0111)	0.98322(0.0063)
			xyz			
	(M1)		(M2)		(M3)	
ρ	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
0	0.54926(0.1284)	0.79347(0.0575)	0.55194(0.1122)	0.81049(0.0504)	0.55065(0.1286)	0.79504(0.0594)
0.2	0.78792(0.0903)	0.86519(0.0511)	0.79247(0.0863)	0.86407(0.0385)	0.77353(0.0768)	0.85750(0.0443)
0.4	0.77608(0.1298)	0.79045(0.0949)	0.78930(0.1070)	0.78971(0.0939)	0.76304(0.1035)	0.77736(0.0920)
0.6	0.69978(0.2149)	0.67514(0.2030)	0.69902(0.1824)	0.69524(0.1557)	0.69716(0.2174)	0.70987(0.1856)
0.8	0.56368(0.2474)	0.60657(0.2027)	0.60709(0.2400)	0.53215(0.1832)	0.57206(0.2080)	0.55701(0.2473)
			RAMP			
	(M1)		(M2)		(M3)	
ρ	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
0	0.68868(0.5037)	0.78503(0.0758)	0.99981(0.0005)	0.99975(0.0005)	0.99994(0.0001)	0.99994(0.0001)
0.2	0.79104(0.4578)	0.87020(0.0444)	0.99976(0.0005)	0.99974(0.0004)	0.99982(0.0003)	0.99981(0.0003)
0.4	0.79195(0.5881)	0.90441(0.0678)	0.99896(0.0026)	0.99876(0.0019)	0.99857(0.0033)	0.99841(0.0018)
0.6	0.34543(0.6467)	0.73968(0.3226)	0.99354(0.0092)	0.99439(0.0077)	0.99204(0.0072)	0.99346(0.0094)
0.8	0.50299(0.4228)	0.60585(0.2526)	0.97875(0.0201)	0.97613(0.0241)	0.97801(0.0231)	0.97482(0.0254)

Table 2.8: The Median of out-of-sample R^2 and the associated RSD (in the parenthesis) for linear model with $p = 2000$

			SSI			
	(M1)		(M2)		(M3)	
ρ	$n = 300$	$n = 400$	$n = 300$	$n = 400$	$n = 300$	$n = 400$
0	0.85418(0.0620)	0.89557(0.0259)	0.85351(0.0408)	0.89846(0.0159)	0.84968(0.0477)	0.89401(0.0227)
0.2	0.92892(0.0254)	0.94463(0.0182)	0.92695(0.0176)	0.94096(0.0153)	0.92617(0.0215)	0.94189(0.0151)
0.4	0.96085(0.0143)	0.96768(0.0076)	0.95684(0.0129)	0.96434(0.0087)	0.95788(0.0169)	0.96478(0.0099)
0.6	0.97231(0.0108)	0.97518(0.0062)	0.97272(0.0113)	0.97487(0.0064)	0.97208(0.0117)	0.97422(0.0082)
0.8	0.98157(0.0073)	0.98381(0.0050)	0.98181(0.0059)	0.98377(0.0052)	0.98122(0.0062)	0.98399(0.0048)
			xyz			
	(M1)		(M2)		(M3)	
ρ	$n = 300$	$n = 400$	$n = 300$	$n = 400$	$n = 300$	$n = 400$
0	0.58010(0.1142)	0.81290(0.0510)	0.56533(0.1357)	0.82146(0.0531)	0.59076(0.1177)	0.81772(0.0551)
0.2	0.83377(0.0643)	0.87378(0.0340)	0.82684(0.0524)	0.87017(0.0307)	0.82472(0.0463)	0.87502(0.0375)
0.4	0.78623(0.0918)	0.79693(0.0744)	0.78523(0.0829)	0.79198(0.0689)	0.78762(0.0885)	0.78743(0.0711)
0.6	0.68467(0.1739)	0.69228(0.1547)	0.68087(0.1551)	0.66030(0.1249)	0.65133(0.1464)	0.70160(0.1457)
0.8	0.58967(0.1798)	0.57715(0.2144)	0.59163(0.2129)	0.55211(0.1469)	0.53657(0.2327)	0.57465(0.1806)
			RAMP			
	(M1)		(M2)		(M3)	
ρ	$n = 300$	$n = 400$	$n = 300$	$n = 400$	$n = 300$	$n = 400$
0	0.81433(0.0749)	0.82122(0.0696)	0.99987(0.0002)	0.99984(0.0002)	0.99997(0.0000)	0.99997(0.0000)
0.2	0.87472(0.0594)	0.87001(0.0505)	0.99989(0.0002)	0.99991(0.0001)	0.99986(0.0002)	0.99989(0.0001)
0.4	0.89756(0.2944)	0.92444(0.0509)	0.99897(0.0016)	0.99899(0.0014)	0.99895(0.0021)	0.99892(0.0013)
0.6	0.59375(0.5608)	0.90287(0.3518)	0.99373(0.0060)	0.99362(0.0057)	0.99371(0.0072)	0.99248(0.0073)
0.8	0.50794(0.3565)	0.62332(0.3557)	0.98185(0.0170)	0.97892(0.0170)	0.98190(0.0153)	0.97614(0.0196)

Table 2.9: The Median of out-of-sample R^2 and the associated RSD (in the parenthesis) for linear model with $p = 5000$

			SSI			
	(M1)		(M2)		(M3)	
ρ	$n = 300$	$n = 400$	$n = 300$	$n = 400$	$n = 300$	$n = 400$
0	0.77501(0.1168)	0.87139(0.0502)	0.79931(0.0760)	0.87859(0.0324)	0.79359(0.0915)	0.87419(0.0336)
0.2	0.91867(0.0220)	0.93855(0.0143)	0.91300(0.0250)	0.93135(0.0128)	0.91402(0.0203)	0.93694(0.0157)
0.4	0.95934(0.0143)	0.96368(0.0101)	0.95513(0.0148)	0.96153(0.0097)	0.95616(0.0151)	0.96112(0.0092)
0.6	0.97362(0.0105)	0.97497(0.0095)	0.97299(0.0082)	0.97551(0.0068)	0.97298(0.0089)	0.97596(0.0091)
0.8	0.98199(0.0073)	0.98344(0.0053)	0.98036(0.0088)	0.98298(0.0057)	0.98181(0.0065)	0.98299(0.0063)
			xyz			
	(M1)		(M2)		(M3)	
ρ	$n = 300$	$n = 400$	$n = 300$	$n = 400$	$n = 300$	$n = 400$
0	0.42767(0.2121)	0.74193(0.0881)	0.47859(0.1429)	0.72840(0.1398)	0.50059(0.1584)	0.75243(0.1200)
0.2	0.78602(0.0876)	0.86104(0.0490)	0.76705(0.1102)	0.85495(0.0435)	0.75615(0.1921)	0.86346(0.0421)
0.4	0.78611(0.1104)	0.78351(0.0874)	0.77673(0.0898)	0.77793(0.0821)	0.75709(0.1292)	0.76752(0.0861)
0.6	0.64861(0.1748)	0.67471(0.1676)	0.63452(0.1586)	0.65403(0.1447)	0.64539(0.1684)	0.67878(0.1319)
0.8	0.52948(0.2067)	0.53357(0.1760)	0.55832(0.2328)	0.57193(0.1822)	0.55593(0.2159)	0.54611(0.2235)
			RAMP			
	(M1)		(M2)		(M3)	
ρ	$n = 300$	$n = 400$	$n = 300$	$n = 400$	$n = 300$	$n = 400$
0	0.80433(0.1780)	0.80874(0.0776)	0.99985(0.0003)	0.99994(0.0001)	0.99998(0.0000)	0.99997(0.0000)
0.2	0.84934(0.4036)	0.88609(0.0697)	0.99978(0.0004)	0.99988(0.0002)	0.99989(0.0002)	0.99990(0.0001)
0.4	0.80287(0.4993)	0.92411(0.2935)	0.99888(0.0014)	0.99915(0.0013)	0.99885(0.0019)	0.99896(0.0017)
0.6	0.51192(0.5455)	0.91247(0.4386)	0.99399(0.0071)	0.99346(0.0067)	0.99500(0.0066)	0.99375(0.0058)
0.8	0.45297(0.2610)	0.59992(0.2817)	0.97640(0.0239)	0.98088(0.0175)	0.97225(0.0232)	0.97797(0.0191)

From Table 2.7-2.9, we can come to a conclusion that (1) as the sample size n increases, all the performance of prediction will become better for all three methods; (2) among these three algorithms, the xyz algorithm still performs worst; comparing SSI with RAMP, SSI outperforms RAMP in the model (M1), and each of them has its own advantages in the Model (M2) and (M3). Actually, the out of sample R^2 are very close in these two models by using SSI and RAMP. Therefore, from the point of prediction, SSI still has better behaviour.

2.5.5 Comparison between SSI and IP

In this section, we will compare the performance of the two methods SSI and IP in the screening interaction terms. The method IP was proposed by Fan et al. [2016]. Following the setting of the last section, n i.i.d. observations are generated from each of the three models (M1)-(M3). And we consider the following different settings: $p = 400, 2000, 5000$, $n = 100, 200$ and $\rho = 0, 0.2, 0.4, 0.6, 0.8$ and repeat each experiment 100 times. We retain the top $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ interactions in each experiment and compare the recovery percentage of each important interaction and the percentage of retaining

all significant ones, where the braces indicate the floor function.

Table 2.10: The percentages of retaining each important interaction (inter1, inter2, inter3) and all important ones (All) by SSI and IP for linear model when $p = 400$

SSI													
		(M1)				(M2)				(M3)			
ρ	n	inter1	inter2	inter3	All	inter1	inter2	inter3	All	inter1	inter2	inter3	All
0	100	0.85	0.83	0.88	0.59	0.86	0.75	0.82	0.51	0.83	0.84	0.87	0.58
0.2	100	0.98	0.84	0.99	0.82	0.85	0.8	0.79	0.54	0.9	0.89	0.81	0.65
0.4	100	1	0.99	0.99	0.98	0.97	0.9	0.87	0.74	0.97	0.97	0.93	0.87
0.6	100	1	1	0.99	0.99	0.93	0.91	0.89	0.73	0.94	0.97	0.94	0.85
0.8	100	1	1	0.99	0.99	0.94	0.94	0.9	0.78	0.96	0.98	0.97	0.91
0	200	1	1	1	1	1	1	1	1	1	1	1	1
0.2	200	1	1	1	1	1	1	1	1	1	1	1	1
0.4	200	1	1	1	1	1	1	1	1	1	1	1	1
0.6	200	1	1	1	1	1	1	1	1	1	1	1	1
0.8	200	1	1	1	1	1	1	1	1	1	1	1	1
IP													
		(M1)				(M2)				(M3)			
ρ	n	inter1	inter2	inter3	ALL	inter1	inter2	inter3	ALL	inter1	inter2	inter3	ALL
0	100	0.55	0.61	0.64	0.17	0.3	0.17	0.18	0	0.43	0.3	0.25	0.02
0.2	100	0.9	0.95	0.87	0.79	0.52	0.57	0.47	0.12	0.63	0.68	0.65	0.25
0.4	100	0.99	0.99	0.99	0.97	0.93	0.92	0.89	0.77	0.93	0.92	0.91	0.79
0.6	100	1	1	1	1	1	0.98	0.97	0.95	0.99	0.98	1	0.97
0.8	100	1	1	1	1	1	1	1	1	1	1	1	1
0	200	0.99	0.89	0.99	0.87	0.56	0.56	0.6	0.16	0.66	0.69	0.78	0.34
0.2	200	1	0.99	1	0.99	0.91	0.93	0.9	0.78	0.92	0.92	0.92	0.8
0.4	200	1	1	1	1	1	1	1	1	1	1	1	1
0.6	200	1	1	1	1	1	1	1	1	1	1	1	1
0.8	200	1	1	1	1	1	1	1	1	1	1	1	1

From the results of Table 2.10-2.12, SSI and IP can almost recover all important interactions when ρ increases and the top $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ interactions are remained. Furthermore, in most cases of our settings, SSI can recover more important interactions than IP does. For example, when $p = 2000$, $n = 100$ and $\rho = 0$, the recovering percentage of IP is 0 while SSI has 19% of 100 simulated results to retain all three interactions in Model (M2). These results indicate that our method SSI can improve the screening accuracy for interaction effects.

2.5.6 Simulation Studies-DSSI

In this part, we continue to consider the performance of DSSI in the interaction screening by using the linear models (M1)-(M3) and logistic models (M4)-(M6). The setting of parameters (β, r, ρ) is same as the simulation studies of SSI in the last section. Firstly, we still evaluate DSSI by computing the median minimal model size

Table 2.11: The percentages of retaining each important interaction (inter1, inter2, inter3) and all important ones (All) by SSI and IP for linear model when $p = 2000$

SSI													
		(M1)				(M2)				(M3)			
ρ	n	inter1	inter2	inter3	All	inter1	inter2	inter3	All	inter1	inter2	inter3	All
0	100	0.64	0.6	0.72	0.23	0.64	0.57	0.63	0.19	0.61	0.62	0.66	0.21
0.2	100	0.85	0.67	0.92	0.49	0.63	0.6	0.63	0.24	0.75	0.73	0.7	0.33
0.4	100	0.96	0.9	0.98	0.85	0.9	0.82	0.9	0.69	0.92	0.92	0.92	0.78
0.6	100	0.99	0.97	0.99	0.95	0.92	0.9	0.94	0.79	0.97	0.95	0.92	0.85
0.8	100	1	1	1	1	0.96	0.9	0.96	0.83	0.99	0.97	0.98	0.94
0	200	0.99	1	1	0.99	0.98	1	0.96	0.94	0.99	0.99	1	0.98
0.2	200	1	0.99	1	0.99	0.99	0.97	0.99	0.96	1	1	1	1
0.4	200	1	1	1	1	1	0.99	1	0.99	1	1	1	1
0.6	200	1	1	1	1	1	0.98	1	0.98	1	0.99	1	0.99
0.8	200	1	1	1	1	0.99	1	1	0.99	1	0.99	1	0.99
IP													
		(M1)				(M2)				(M3)			
rho	n	inter1	inter2	inter3	ALL	inter1	inter2	inter3	ALL	inter1	inter2	inter3	ALL
0	100	0.31	0.3	0.39	0.04	0.04	0.08	0.05	0	0.12	0.09	0.15	0
0.2	100	0.78	0.69	0.75	0.51	0.25	0.15	0.18	0	0.33	0.31	0.34	0.02
0.4	100	0.98	0.91	0.94	0.87	0.69	0.65	0.72	0.39	0.78	0.78	0.77	0.56
0.6	100	0.99	0.98	0.97	0.95	0.95	0.96	0.95	0.87	0.98	0.95	0.96	0.9
0.8	100	1	0.99	1	0.99	1	1	1	1	1	1	1	1
0	200	0.81	0.76	0.88	0.53	0.15	0.22	0.19	0	0.41	0.41	0.34	0.03
0.2	200	0.99	0.98	0.99	0.96	0.59	0.56	0.49	0.18	0.81	0.77	0.72	0.49
0.4	200	1	1	1	1	0.97	0.96	0.98	0.91	1	1	1	1
0.6	200	1	1	1	1	1	1	1	1	1	1	1	1
0.8	200	1	1	1	1	1	1	1	1	1	1	1	1

Table 2.12: The percentages of retaining each important interaction (inter1, inter2, inter3) and all important ones (All) by SSI and IP for linear model when $p = 5000$

SSI													
		(M1)				(M2)				(M3)			
ρ	n	inter1	inter2	inter3	All	inter1	inter2	inter3	All	inter1	inter2	inter3	All
0	100	0.58	0.56	0.62	0.18	0.48	0.52	0.38	0.05	0.5	0.48	0.56	0.08
0.2	100	0.77	0.53	0.86	0.37	0.51	0.4	0.46	0.07	0.48	0.54	0.61	0.13
0.4	100	1	0.87	0.99	0.86	0.81	0.75	0.82	0.51	0.85	0.85	0.91	0.66
0.6	100	1	0.97	1	0.97	0.92	0.89	0.91	0.74	0.97	0.96	0.98	0.91
0.8	100	1	0.98	1	0.98	0.95	0.94	0.97	0.86	0.99	0.98	0.99	0.96
0	200	1	0.94	0.98	0.92	0.97	0.98	1	0.95	0.98	0.97	1	0.96
0.2	200	1	0.94	0.99	0.93	0.97	0.96	0.98	0.91	0.98	1	0.99	0.97
0.4	200	1	0.99	1	0.99	1	0.99	0.99	0.98	0.99	1	1	0.99
0.6	200	1	1	1	1	1	0.99	1	0.99	1	1	1	1
0.8	200	1	1	1	1	1	0.99	1	0.99	1	1	1	1
IP													
0	100	0.19	0.13	0.22	0.01	0.08	0.03	0.04	0	0.02	0.03	0.02	0
0.2	100	0.64	0.58	0.67	0.35	0.15	0.2	0.17	0.01	0.11	0.2	0.29	0.01
0.4	100	0.85	0.83	0.82	0.67	0.68	0.67	0.72	0.41	0.61	0.7	0.67	0.4
0.6	100	0.93	0.93	0.89	0.84	0.93	0.95	0.93	0.84	0.91	0.94	0.9	0.81
0.8	100	0.99	1	1	0.99	0.99	1	1	0.99	0.98	1	0.99	0.97
0	200	0.73	0.61	0.76	0.32	0.17	0.13	0.23	0	0.25	0.25	0.26	0.01
0.2	200	0.96	0.97	0.98	0.93	0.43	0.53	0.51	0.1	0.58	0.64	0.65	0.25
0.4	200	1	1	1	1	0.96	0.95	0.93	0.85	0.95	0.99	0.99	0.93
0.6	200	1	1	1	1	1	1	1	1	1	1	1	1
0.8	200	1	1	1	1	1	1	1	1	1	1	1	1

(MMMS) of the selected model and its associated RSD, which is the interquartile range (IQR) divided by 1.34, according to 100 simulated results. Time (in seconds) represents the average time of computing the test statistics of all the interaction terms in 100 simulated data sets. Secondly, we also examine the prediction performance of DSSI by computing the median of the out of sample R^2 of the model with remaining interactions and its RSD. In the process of estimating MMMS, the following different settings are considered: (p, n)=(400, 300), (2000, 400), (5000, 500). When we assess the prediction performance of DSSI, we choose the different setups with $p = 400, 2000, 5000$ and $n = 200, 300$. Here, in the first step of DSSI, we transform all predictors to the categorical variables with the same level $l = 3$, and the response Y is simply changed into a binary response. And the threshold γ_{KSA} is chosen as 0, which means that all interactions do not be deleted in Step 2 of DSSI with KSA.

Table 2.13 and Table 2.14 list the interaction screening results of DSSI for all settings in the linear models (M1)-(M3) and the logistic models (M4)-(M5), respectively. Compared with the results of SSI in Table 2.1-2.6, the time is dramatically reduced.

Table 2.13: The MMMS and the associated RSD (in the parenthesis) for linear model when $p = 400,\ 2000,\ 5000$

p	ρ	n	MMMS	Time	MMMS	Time	MMMS	Time
			(M1)		(M2)		(M3)	
	0	300	9.5(40)	0.5199	15.5(47)	0.5166	10.5(49)	0.5136
	0.2	300	4(10)	0.5169	26.5(125)	0.5196	13.5(43)	0.5181
400	0.4	300	4(2)	0.5167	31.5(59)	0.5186	6(14)	0.5209
	0.6	300	10(10)	0.5169	27(69)	0.5172	13.5(23)	0.5161
	0.8	300	33(42)	0.5213	78(182)	0.5170	57.5(77)	0.5149
	0	400	9.5(60)	9.8733	9(28)	9.8792	4(11)	9.8846
	0.2	400	3(2)	9.8780	16(117)	9.8774	10.5(39)	9.8868
2000	0.4	400	3(0)	9.8736	17(65)	9.9024	7(15)	9.8780
	0.6	400	5(5)	9.9009	25.5(110)	9.9449	11(30)	9.9284
	0.8	400	27.5(30)	9.9413	74(165)	9.9622	44.5(55)	9.9541
	0	500	3(3)	61.3658	4.5(24)	61.2465	3(3)	61.0923
	0.2	500	3(0)	61.3542	8(73)	61.0974	3(12)	60.9617
5000	0.4	500	3(0)	61.2537	6(40)	61.1416	3(5)	61.0074
	0.6	500	4(2)	61.4720	7(12)	61.1504	7(8)	60.9372
	0.8	500	22(18)	61.4567	51.5(74)	61.4286	38.5(30)	61.1283

Table 2.14: The MMMS and the associated RSD (in the parenthesis) for logistic model when $p = 400,\ 2000,\ 5000$

p	ρ	n	MMMS	Time	MMMS	Time	MMMS	Time
			(M4)		(M5)		(M6)	
	0	300	50.5(271)	0.5232	35.5(277)	0.5274	23(99)	0.5213
	0.2	300	9(38)	0.5159	112.5(270)	0.5163	55(262)	0.5168
400	0.4	300	5(13)	0.5144	84(409)	0.5250	40.5(193)	0.5175
	0.6	300	9(23)	0.5183	109(253)	0.5193	43(149)	0.5205
	0.8	300	176.5(419)	0.5244	632(1924)	0.5147	426.5(1520)	0.5232
	0	400	65(378)	9.9543	46(335)	10.0584	27(83)	10.0244
	0.2	400	14.5(76)	9.9846	77.5(1219)	9.9962	26(130)	10.0198
2000	0.4	400	4(11)	10.0796	93(666)	10.0289	35.5(192)	9.9967
	0.6	400	7(38)	10.0638	138.5(841)	10.0641	40.5(424)	9.9704
	0.8	400	202.5(1179)	9.98360	2468(9846)	10.0595	941.5(4418)	9.9626
	0	500	24.5(123)	61.5150	9.5(156)	61.0798	5(30)	61.0308
	0.2	500	3(5)	61.1904	25(87)	61.5622	15(71)	61.2446
5000	0.4	500	3(1)	61.3408	29.5(274)	61.1146	4(26)	61.2586
	0.6	500	3(3)	61.3469	59.5(355)	61.7705	10.5(81)	61.4825
	0.8	500	85(604)	61.3699	711.5(11615)	61.7749	233(2377)	61.4225

Because of the loss of information after discretization, the accuracy is influenced by the discretization. Despite the loss, DSSI also can reduce the large scale interaction terms to a moderate scale. As the correlation ρ increases, the MMMS usually increase for all DSSI, which is consistent with SSI. DSSI is more efficient in general and its performance is also excellent.

Table 2.15: The Median of out-of-sample R^2 and the associated RSD (in the parenthesis) for linear model when $p = 400$

	(M1)		(M2)		(M3)	
ρ	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
0	0.39418(0.2433)	0.44239(0.2507)	0.41878(0.2594)	0.38171(0.2543)	0.48669(0.2668)	0.50839(0.2564)
0.2	0.70465(0.2308)	0.72533(0.2513)	0.59010(0.1961)	0.53819(0.2162)	0.56643(0.227)	0.54490(0.2220)
0.4	0.87971(0.1118)	0.88638(0.092)	0.77860(0.1470)	0.78056(0.1221)	0.81810(0.1256)	0.83392(0.1151)
0.6	0.94658(0.0354)	0.95455(0.0254)	0.91935(0.0525)	0.92743(0.0393)	0.92810(0.0566)	0.93893(0.0376)
0.8	0.97249(0.0128)	0.97349(0.0093)	0.96778(0.0182)	0.97081(0.0111)	0.96700(0.017)	0.97255(0.0103)

Table 2.16: The Median of out-of-sample R^2 and the associated RSD (in the parenthesis) for linear model when $p = 2000$

	(M1)		(M2)		(M3)	
ρ	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
0	0.55740(0.2759)	0.88789(0.2313)	0.43949(0.2155)	0.69261(0.2388)	0.58231(0.2514)	0.90176(0.2300)
0.2	0.77366(0.2393)	0.94240(0.0936)	0.54718(0.1928)	0.81883(0.1540)	0.65819(0.2001)	0.83530(0.1517)
0.4	0.94327(0.0733)	0.96435(0.0142)	0.81898(0.1064)	0.89886(0.0802)	0.84920(0.1049)	0.95153(0.0557)
0.6	0.96784(0.0220)	0.97565(0.0093)	0.94019(0.0402)	0.96446(0.0216)	0.94766(0.0380)	0.97066(0.0123)
0.8	0.97713(0.0114)	0.98119(0.0049)	0.97209(0.0123)	0.97914(0.0089)	0.97360(0.0111)	0.97840(0.0088)

Table 2.17: The Median of out-of-sample R^2 and the associated RSD (in the parenthesis) for linear model when $p = 5000$

	(M1)		(M2)		(M3)	
ρ	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
0	0.36979(0.2517)	0.89957(0.1710)	0.29226(0.2301)	0.91059(0.1601)	0.36990(0.2450)	0.91182(0.0698)
0.2	0.65909(0.1660)	0.94585(0.0369)	0.43198(0.1987)	0.86889(0.1895)	0.50751(0.2528)	0.92668(0.1389)
0.4	0.86543(0.1471)	0.96397(0.0221)	0.78673(0.1266)	0.94488(0.0595)	0.76558(0.1424)	0.94830(0.0422)
0.6	0.94798(0.0304)	0.97578(0.0197)	0.91401(0.0423)	0.96372(0.0215)	0.92051(0.0363)	0.96827(0.0171)
0.8	0.97376(0.0116)	0.98337(0.0143)	0.97152(0.0120)	0.97934(0.0121)	0.97202(0.0111)	0.97871(0.0171)

Table 2.15-2.17 list all results of prediction in the linear model (M1)-(M3). We can obtain the following conclusion: (1) when n is small, the accuracy of prediction is not good for all settings, since it is affected by the discretization, and as n increases, the out of sample R^2 becomes more close to 1; (2) the out of sample R^2s are not bad by comparing them with that of other methods in Table 2.7-2.9. Actually, they outperform the algorithm xyz.

In summary, although the algorithm DSSI is influenced by the discretization and

the information is a little loss, DSSI can efficiently filter out the unimportant interactions and also its prediction is superior.

2.6 Numerical Studies II

In this section, we will continue to demonstrate the performance of the new algorithm DSSI with KSA in the statistical simulation. Firstly, we take the logistic model with the dimension $p = 50$ as a toy example in Section 2.6.1, and discuss which value of the arity l is appropriate for our data analysis. Later, we will provide several results in the different designed scenarios. We choose the threshold value $\gamma_{KSA} = 30$ through this section. All of the designed scenarios are evaluated by summarizing Pr, which is the percentage of retaining significant interaction effect based on 100 simulated results, MMMS and the associated RSD based on the simulated results which include the important interaction terms.

2.6.1 Example 3-Logistic Model

Now, following the setting of Example 1 in Fan et al. [2016], three interaction models are considered, where the conditional distribution of the response Y given $\boldsymbol{X} = \boldsymbol{x}$ is a binomial distribution.
(M7 strong heredity): $\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = 2X_1 + 2X_6 + 3X_1X_6$,
(M8 anti heredity): $\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = 2X_{10} + 2X_{16} + 3X_1X_6$,
(M9 weak heredity): $\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = 2X_1 + 2X_{10} + 3X_1X_6$;
where the predictor vector $\boldsymbol{x} = (X_1, \ldots, X_p)^T$ follows the multivariate normal distribution with zero mean vector and convariance matrix $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{1\leq i,j\leq p}$. A sample of n i.i.d. observations are generated from the above three models respectively. In this part, we first focus on the choice of the arity l by the setting of $p = 50$, $n = 200,\ 300,\ 400$ and $\rho = 0,\ 0.5$. Secondly, in order to evaluate the algorithm DSSI with KSA, we consider the dimension of the covariates as p =50, 100, 2000 and 5000, take the parameter ρ as 0 and 0.5, and vary the sample size from 200 to 400 for different scenarios.

Table 2.18 lists the comparison results for the values $l = 3,\ 4,\ 5$ and 6 in recovering

Table 2.18: The Pr, MMMS and the associated RSD (in the parenthesis) for logistic model with $p = 50$

$\rho = 0$		$l = 3$		$l = 4$		$l = 5$		$l = 6$	
	n	Pr	DSSI	Pr	DSSI	Pr	DSSI	Pr	DSSI
M7	200	0.74	1(0)	1	1(0)	0.99	1(1)	0.93	3(2)
	300	0.97	1(0)	1	1(0)	0.99	1(0)	0.99	1(0)
	400	0.99	1(0)	1	1(0)	1	1(0)	1	1(0)
M8	200	0.77	1(0)	0.97	1(0)	1	1(1)	0.99	3(1)
	300	1	1(0)	1	1(0)	1	1(0)	1	1(0)
	400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
M9	200	0.86	1(0)	0.98	1(0)	1	1(1)	0.95	2(1)
	300	0.99	1(0)	1	1(0)	1	1(0)	0.99	1(0)
	400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
$\rho = 0.5$		$l = 3$		$l = 4$		$l = 5$		$l = 6$	
M7	200	0.79	1(0)	1	1(0)	0.97	1(0.75)	0.92	4(4)
	300	1	1(0)	1	1(0)	1	1(0)	1	1(0)
	400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
M8	200	0.8	1(0)	0.98	1(0)	1	1(0.75)	0.93	3(2)
	300	0.98	1(0)	1	1(0)	1	1(0)	0.99	1(0)
	400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
M9	200	0.92	1(0)	0.97	1(0)	1	1(1)	0.98	3(3)
	300	1	1(0)	1	1(0)	1	1(0)	0.99	1(0)
	400	1	1(0)	1	1(0)	1	1(0)	1	1(0)

the significant interaction effect and retaining the important ones when $p = 50$. From visualization of the results in Table 2.18, we draw the conclusions of our experiment: (1) no matter which models, as the sample size n increases, the Pr would be close or equal to 100% and screening accuracy will be better ; (2) as the arity l increases, some pattern appears, i.e, the Pr and the screening accuracy will increase a little and then decrease a little with respect to the arity l. Consequently, DSSI performs well in all of the situations and the arity l will be taken based on the sample size n and we recommend that no more than 5 is enough for most cases.

Based on the results of Tables 2.18-2.19, we can find that the algorithm DSSI with KSA is very powerful to select the interaction terms although discretization can lose some information. Actually, we can choose different threshold values for different dimensions. For example, for $p = 50$, by Bonferroni correction, at the 0.05 significance level, the value $\tau = 25.45$ can be used. Hence, the power will be increased.

Table 2.19: The Pr, MMMS and the associated RSD (in the parenthesis) for logistic model when $p = 100$, 2000 and 5000

			$\rho = 0$, 1= 3		$\rho = 0,\ l = 4$		$\rho = 0.5,\ l = 3$		$\rho = 0.5,\ l = 4$	
	p	n	Pr	DSSI	Pr	DSSI	Pr	DSSI	Pr	DSSI
		200	0.75	1(0)	1	1(1)	0.83	1(0)	0.99	1(0)
M7	100	300	0.97	1(0)	1	1(0)	0.99	1(0)	1	1(0)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.82	1(0)	0.97	1(1)	0.87	1(0)	0.95	1(1)
M8	100	300	0.98	1(0)	1	1(0)	0.99	1(0)	1	1(0)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.92	1(0)	1	1(1)	0.87	1(0)	1	1(1)
M9	100	300	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.76	1(0)	0.99	39(117)	0.78	1(1)	1	37(89)
M7	2000	300	0.97	1(0)	1	1(1)	0.99	1(0)	1	1(1)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.8	1(1)	0.98	39(156)	0.77	1(1)	0.96	46.5(163)
M8	2000	300	0.98	1(0)	1	1(1)	0.99	1(0)	0.99	2(5)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.87	1(0)	1	47(161)	0.85	1(0)	0.99	51(162)
M9	2000	300	1	1(0)	1	1(4)	1	1(0)	1	1(3)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.76	1(3)	1	283(842)	0.78	1(1)	1	275.5(959)
M7	5000	300	1	1(0)	1	2(9)	0.97	1(0)	1	2(4)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.81	1(4)	0.95	301(1767)	0.76	1(2.2)	0.97	607(1497)
M8	5000	300	0.98	1(0)	1	3(29)	0.99	1(0)	1	3(16)
		400	1	1(0)	1	1(0)	0.99	1(0)	1	1(0)
		200	0.86	1(4)	0.92	420(1123)	0.92	1(2)	1	356(991)
M9	5000	300	1	1(0)	1	2(9)	1	1(0)	1	6(23)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)

2.6.2 Example 4-Linear Model

In this subsection, We still follow the setup of Example 1 in Fan et al. [2016] and emphasize the three interaction models, where the conditional distribution of the response Y given $\boldsymbol{X} = \boldsymbol{x}$ is a normal distribution.

(M10 strong heredity): $Y = 2X_1 + 2X_6 + 3X_1X_6 + \varepsilon_1$,

(M11 anti heredity): $Y = 2X_{10} + 2X_{16} + 3X_1X_6 + \varepsilon_2$,

(M12 weak heredity): $Y = 2X_1 + 2X_{10} + 3X_1X_6 + \varepsilon_3$;

where the predictor vector $\boldsymbol{x} = (X_1, \ldots, X_p)^T$ follows the multivariate normal distribution with zero mean vector and convariance matrix $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{1\leq i,j\leq p}$ and the noises $\varepsilon_1 \sim N(0, 2.5^2)$, $\varepsilon_2 \sim N(0, 2^2)$ and $\varepsilon_3 \sim N(0, 2^2)$. One set of i.i.d. observations with sample size n are generated from the above three models respectively.

We still first discretize the continuous predictor into the categorical variables with finite l levels. And for the response Y, we simply split it into two parts by using the median. Next, we use the Method "DSSI" to pick up all of the significant interaction effects. Here, we consider the following different settings: $n = 200, 300, 400$; $p = 100, 2000, 5000$; $\rho = 0, 0.5$ and $l = 3, 4$. All of the results are presented in the Table 2.20, based on the 100 repeated experiments. Surprisingly, the method "DSSI" still performs well although we simply transform the response Y to the binary variable.

By the results of Table 2.18-2.20, we would find that in two kinds of models, (1) when the dimension p is small, the Pr and MMMS are a little influenced by the arity l; (2) when dimension p increases and sample size n is not large, sometimes we obtain a bad result such as $p = 500$ and $n = 200$ in the model (M10) with $l = 4$, in this case, $l = 3$ is more appropriate; (3) as the sample size n becomes larger, the Pr is close to 1 and the MMMS is more accurate. All in all, DSSI with KSA has a good performance in the linear models and logistic models only if the arity l is not large.

2.6.3 Comparison between SSI and DSSI

In this subsection, we would like to compare the performance of the method "SSI" and "DSSI", based on the setting of models in the section 2.6.1 and 2.6.2. Here, we do not prune any one of interaction effects in the second step of method "DSSI", which means that the threshold value γ_{KSA} is 0. Here, we choose the arity $l = 3$.

Table 2.20: The Pr, MMMS and the associated RSD (in the parenthesis) for linear model when $p = 100$, 2000 and 5000

			$\rho = 0,\ l = 3$		$\rho = 0,\ l = 4$		$\rho = 0.5,\ l = 3$		$\rho = 0.5,\ l = 4$	
	p	n	Pr	DSSI	Pr	DSSI	Pr	DSSI	Pr	DSSI
		200	0.73	1(0)	1	1(1)	0.74	1(0)	0.99	1(1)
M10	100	300	0.96	1(0)	1	1(0)	0.99	1(0)	1	1(0)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.76	1(0)	0.95	1(1)	0.86	1(0)	0.97	1(1)
M11	100	300	0.97	1(0)	1	1(0)	0.95	1(0)	1	1(0)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.75	1(0)	0.99	1(1)	0.78	1(0)	0.96	1(1)
M12	100	300	0.99	1(0)	1	1(0)	0.98	1(0)	1	1(0)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.66	1(1.3)	0.99	170(294)	0.66	1(1.3)	0.96	132(456)
M10	2000	300	0.93	1(0)	1	3(8)	0.96	1(0)	1	2(10)
		400	0.99	1(0)	1	1(1)	1	1(0)	1	1(0)
		200	0.76	1(0.75)	0.93	89(231)	0.72	1(1)	0.92	105.5(207)
M11	2000	300	0.97	1(0)	1	1(3)	0.98	1(0)	0.99	1(4)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.77	1(0)	0.97	89(206)	0.85	1(1)	0.96	84(249)
M12	2000	300	0.96	1(0)	1	2(8)	1	1(0)	1	2.5(12)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.7	2(7.5)	0.99	1265(3268)	0.7	1.5(5)	0.97	783(2127)
M10	5000	300	0.97	1(0)	1	16(73)	0.97	1(0)	1	10.5(88)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(0)
		200	0.79	2(10.8)	0.95	660(1314)	0.74	2(4)	0.96	587.5(1444)
M11	5000	300	0.99	1(0)	1	5.5(39)	0.98	1(0)	1	8(50)
		400	0.99	1(0)	1	1(1)	1	1(0)	1	1(0)
		200	0.86	1(4)	0.97	765(1810)	0.76	1(4)	0.97	608(1805)
M12	5000	300	0.99	1(0)	1	8(49)	0.99	1(0)	1	11(49)
		400	1	1(0)	1	1(0)	1	1(0)	1	1(1)

The setting of dimension p and sample size n are (50, 100), (400, 200), (2000, 200) and (5000, 200) with correlation $\rho = 0,\ 0.5$.

Table 2.21: The comparison between SSI and DSSI for logistic models

p	Model	$\rho = 0$, $l = 3$			ρ=0.5, $l = 3$		
		n	SSI	DSSI	n	SSI	DSSI
	M7		1(0)	2(11)		1(0)	1(6)
50	M8	100	1(0)	2(7)	100	1(0)	1(4)
	M9		1(0)	1(3)		1(0)	1(2)
	M7		1(0)	1(0)		1(0)	1(0)
400	M8	200	1(0)	1(0)	200	1(0)	1(0)
	M9		1(0)	1(0)		1(0)	1(0)
	M7		1(0)	1(6)		1(0)	1(1)
2000	M8	200	1(0)	2(11)	200	1(0)	2(12)
	M9		1(0)	1(1)		1(0)	1(0)
	M7		1(0)	2(39)		1(0)	2(17)
5000	M8	200	1(0)	7(33)	200	1(0)	3(26)
	M9		1(0)	1(16)		1(0)	1.5(9)

Table 2.22: The comparison between SSI and DSSI for linear model

p	model	$\rho = 0$, level=3			$\rho = 0.5$, level=3		
		n	SSI	DSSI	n	SSI	DSSI
	M10		1(0)	4(13)		1(0)	3(8)
50	M11	100	1(0)	2(8)	100	1(0)	2(5)
	M12		1(0)	2(5)		1(0)	2(7)
	M10		1(0)	1(4)		1(0)	1(2)
400	M11	200	1(0)	1(0)	200	1(0)	1(0)
	M12		1(0)	1(0)		1(0)	1(0)
	M10		1(0)	1(16)		1(0)	3(23)
2000	M11	200	1(0)	2(5)	200	1(0)	1(7)
	M12		1(0)	1(2)		1(0)	1(5)
	M10		1(0)	37.5(374)		1(0)	9(221)
5000	M11	200	1(0)	7(128)	200	1(0)	3(31)
	M12		1(0)	2.5(39)		1(0)	2(21)

Here, all experiments are still evaluated by calculating the median minimal model size (MMMS) of the selected model and its corresponding RSD, which is the interquartile range (IQR) divided by 1.34, in the 100 simulated results. The results are shown in Tables 2.21 and 2.22. These two tables summarize the comparison results for our algorithm "SSI" and "DSSI" when the dimension p varies from 50 to 5000. We can find that although the false selection rate of "DSSI"is a little larger than that of "SSI", "DSSI" is still powerful for the interaction selection. These results help us verify our theoretical results of Theorem 2.4.2. Moreover, all of the sure screening

methods are very rough, therefore we can conclude that "DSSI" is available for the interaction selection if we would like to improve our method's efficiency.

2.7 Real Data Analysis

In this section, we will illustrate our method SSI and DSSI through the analysis of three real data sets: Prostate Cancer Data, Leukemia Data and Supermarket Data.

2.7.1 Prostate Cancer Data

Here, we use the prostate cancer data studied in Singh et al. [2002]. It has been studied in many papers to evaluate the classification methods or the screening methods for high dimensional data, such as Pochet et al. [2004], Fan and Fan [2008], Hall et al. [2009], Hall and Xue [2014] and Fan et al. [2016] . This data set involves the tumor group and normal group with $p = 12600$ genes. Their sample sizes are 77 and 59 respectively. A four-step procedure is used in the data's preprocessing, which was suggested by Hall and Xue [2014]. The first step is to make the intensities positive by truncating them; the second step is to filter out the gene with little variation in intensity; the third step is to replace intensities by base 10 logarithms; the last step is to standardize every data vector. After this procedure, we remain $p = 3239$ genes for the following data analysis.

SIS is utilized to independently select the top $n-1$ main effects, top $n-1$ quadratic terms. For interaction terms, we use SSI and DSSI to choose the top $n-1$ interaction effects, respectively. All predictors are transformed into categorical variables and their levels are 3 in DSSI. We use the threshold $\gamma_{KSA} = 30$ to prune some interaction terms and then $n-1$ interaction terms are selected in the remaining interactions. The final step of SSI and DSSI is to refine the results by using the LASSO penalty. AIC and BIC are used to choose the tuning parameter for the penalty function. Finally, we get these significant main terms and interaction effects, which are presented in the table 2.23.

From the results of Table 2.23, we can find that many same genes such as HPN, LMO3, HSPD1 exist in the two significant sets before and after discretization, al-

Table 2.23: Comparison of results between before and after discretization in prostate cancer data analysis

Main effects with SSI									
AIC	HPN	LMO3	HSPD1	PTGDS	S100A4	ATP2C1	PLA2G7	MAF	RGS10
	RAP1GAP	GUCY1A3	TP63						
BIC	HPN	LMO3	HSPD1	S100A4	ATP2C1	PLA2G7	SLC25A6	MAF	RGS10
	RAP1GAP	GUCY1A3							
Interaction terms with SSI									
AIC	Gene1	MYOF	SEMA6C	MYOF	TPTEP1				
	Gene2	ERCC1	RBPMS	PSMD10	CALM1				
BIC	Gene1	MYOF							
	Gene2	PSMD10							
Main effects with DSSI									
AIC	HPN	LMO3	HSPD1	S100A4	ATP2C1	NELL2	PLA2G7	MAF	ZMPSTE24
	GUCY1A3								
BIC	HPN	LMO3	CALM1	HSPD1	TARP				
Interaction terms with DSSI									
AIC	Gene1	SEPT9	PSD4	MCCC2	DMTN	FAM32A			
	Gene2	ERG	ABL1	TSPYL4	ABL1	ABL1			
BIC	Gene1	—							
	Gene2	—							

though we only did the screening procedure one time. Those genes may be essential in diagnosing prostate cancer. For instance, many papers, such as Dhanasekaran et al. [2001], Luo et al. [2001], Magee et al. [2001] and Luo et al. [2002], reveal that the Hepsin (HPN) is regarded as one of the highest differences in expression ratio between normal and prostate cancer. After using the technique "Discretization", our selection results are not much more affected. We still get some important main effects and interaction terms. Especially, the gene "ERG" can be found by DSSI as a part of interaction $SEPT9 \times ERG$. The gene "ERG" is an important one which has been studied by a wide range of research that investigate the influence on the prostate cancer, such as Demichelis et al. [2007], Klezovitch et al. [2008].

2.7.2 Leukemia Data

Leukemia data was firstly studied by Golub et al. [1999]. This data set is available at `http://web.stanford.edu/~hastie/CASI/data.html`. It includes 3751 genes and 72 samples, which comes from two classes: 47 in class ALL (acute lymphocytic leukemia) and 25 in class AML (acute myelogenous leukemia). We consider the class as the response Y ($Y = 1$ if ALL; $Y = 0$ if AML) and consider 3751 gene expression levels as the predictors. On the one hand, we still apply SIS and SSI only once to select the top 71 main effects, top 71 quadratic terms and top 71 interaction effects, respectively. On the other hand, we transform the continuous gene expressions to cat-

egorical predictors with 3 values and then use DSSI to choose the top 71 interactions. For the remaining main effects, quadratic terms and interactions, the LASSO penalty will be further applied to obtain more refined results, in which tuning parameter is selected by the criteria AIC and BIC. The significant main terms and interaction effects are presented in the table 2.24.

Table 2.24: The main effects and interaction effects of Leukemia data

Main effects with SSI										
AIC	956	2481	3441	979	1182	456	1652	1219	626	672
BIC	956	2481	3441	979	1182	456	1652	1219	626	672
Interaction effects with SSI										
	AIC	**BIC**								
Gene 1	1136	—								
Gene 2	2619	—								
Main effects with DSSI										
AIC	956	2481	3441	979	1182	456	1652	1219	626	672
BIC	956	2481	3441	979	1182	456	1652	1219	626	672
Interaction effects with DSSI										
	AIC	**BIC**								
Gene 1	—	—								
Gene 2	—	—								

Based on the results of Table 2.24, we can find that different methods have similar results. Before and After discretization, we obtain the almost same main effects, and get one significant interaction effect with SSI, but do not pick out the interaction terms with DSSI. Therefore, we believe that the significant interaction terms may not be present in the original data set or the sample size is not large enough, which results in lacking information after discretization.

2.7.3 Supermarket Data

The supermarket data was collected from a major supermarket located in northern China and has been analyzed by Wang [2009], Hao and Zhang [2014] and Hao et al. [2016], which includes 6398 predictors and 464 observations. The response is the number of customers on a particular day and each of predictors is the corresponding sale volume of the product. The supermarket manager wonder which products would be more associated with the number of customers, which means that he or she wants to select most informative products to predict the response.

Here, we randomly select 400 observations as the training data and the remaining 64 observations as the testing data and then use the out of sample R^2 to evaluate the prediction performance of our methods based on 100 random splits. The out of sample R^2 is defined by

$$R^2 = 100\% \times \left\{ 1 - \frac{\sum (Y_i^* - \boldsymbol{X}_i^{*T} \hat{\boldsymbol{\beta}})^2}{\sum (Y_i^* - \bar{Y}^*)^2} \right\},$$

where $(\boldsymbol{X}_i^*, Y_i^*)$ is the testing data and $\hat{\boldsymbol{\beta}}$ is the estimate of the coefficient based on the training data. For each split, we use SIS to select the top 399 main effects, top 399 terms about the square of main effects, and apply our methods to obtain the top 399 important interaction effects simultaneously. Next, we continue to use LASSO with cross-validation (CV) to get more refined results. The average performance is summarized in Table 2.25, which includes the average sizes of main effects and interaction effects, the average R^2 and their standard errors over 100 random splits. Besides the results of our methods, Table 2.25 displays the out-of-sample R^2 by other methods: FR-SCAD [Wang [2009]], iFORT & iFORM [Hao and Zhang [2014]] and RAMP [Hao et al. [2016]]. The corresponding results are extracted directly from their papers.

Table 2.25: Average results and the standard errors (in parentheses) on the supermarket data set

	main size	inter size	R^2(%)
SSI	107.7(0.73)	10.9(0.37)	92.73(0.14)
DSSI	98.81(0.64)	95.21(2.30)	92.29(0.14)
RAMP-AIC	229.18(1.68)	94.53 (1.06)	90.48(0.23)
RAMP-BIC	101.17(3.25)	34.36(1.65)	91.18(0.20)
RAMP-EBIC	29.27(1.01)	3.07(0.29)	89.67(0.31)
RAMP-GIC	30.71(0.92)	3.20(0.30)	90.08(0.28)
iFORT	———	———	88.91(0.17)
iFORM	———	———	88.66(0.18)
FR-SCAD	12(0.11)	———	88.00(0.17)

Obviously, the method "SSI" and "DSSI" have better performance in the prediction, and also "SSI" performs a little better than other methods do. Moreover, compared to the results of the method RAMP, the method "DSSI" with CV outperforms the RAMP with the tuning parameter selection methods (AIC, BIC, EBIC,

GIC) for the out-of-sample R^2 values with associated standard error. Although the method "SSI" with CV performs worse than RAMP with BIC, its prediction performance is better than RAMP with AIC, EBIC, GIC , iFORT, iFORM and FR-SCAD. Moreover, our standard errors are relatively smaller, which means that our methods are more robust.

2.8 Conclusion

This chapter studies the method of screening important interaction effects in the ultra-high dimensional generalized linear model. We propose a simple and new procedure SSI to detect the significant interaction effects in the high or ultra-high dimensional generalized linear model space. The most important thing is that our method is independent of the heredity assumption. And also we investigate the sure screening properties of the proposal method from theoretical insight. Furthermore, we prove that our new method can control the false discovery rate at a reasonable size. Moreover, we provide one efficient algorithms "DSSI" to realize our proposed sure screening method in practice. DSSI is based on the discretization and Boolean representation. The basic idea is to transform the continuous variables to discrete variables and approximate the increment of log-likelihood function by Kirkwood Superposition Approximation (KSA). Using the idea of discretization and KSA, "DSSI" becomes more efficient and powerful. From the point of theory, we build a bridge between SIS and discretized SIS, SSI and DSSI, show that all of them have the consistent results. In the numerical studies, we consider several models and some real data analyses to demonstrate our algorithm. The simulation results and real data analysis point out that the proposed procedure not only performs well in the selection of interaction terms, but also can improve the model's prediction accuracy.

2.9 Appendix-Technical Proof of the Theorems

Proof of Theorem 2.2.1: If $\beta_{ij}^M = 0$, by the model identifiability, $\beta_{i,j0}^M = \beta_{ij0}^M$, $\beta_{i,}^M = \beta_i^M$ and $\beta_{j,}^M = \beta_j^M$. Therefore, $L_{ij}^\star = 0$. On the other hand, if $L_{ij}^\star = 0$, based on Condition (C), we can know $\boldsymbol{\beta}_{i,j}^M = \boldsymbol{\beta}_{ij}^M$, which infers that $\beta_{i,j0}^M = \beta_{ij0}^M$, $\beta_{i,}^M = \beta_i^M$, $\beta_{j,}^M = \beta_j^M$ and $\beta_{ij}^M = 0$.

Proof of Theorem 2.2.2: Note that the condition $\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) = 0$ is equivalent to $E\{(Y - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))X_{ij}\} = 0$. Firstly, we prove the necessary part. The marginal regression coefficients $\boldsymbol{\beta}_{ij}^M$ satisfy the score equation

$$E\{b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M)\boldsymbol{X}_{ij}\} = E(Y\boldsymbol{X}_{ij}) = E(E(Y|\boldsymbol{X})\boldsymbol{X}_{ij}) = E(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)\boldsymbol{X}_{ij}), \tag{2.15}$$

i.e.,

$$E\{b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M)\boldsymbol{X}_{i,j}\} = E(Y\boldsymbol{X}_{i,j}) = E(E(Y|\boldsymbol{X})\boldsymbol{X}_{i,j}) = E(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)\boldsymbol{X}_{i,j}), \tag{2.16}$$

and the coefficients $\boldsymbol{\beta}_{i,j}^M$ satisfy the score equation

$$E\{b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)\boldsymbol{X}_{i,j}\} = E(Y\boldsymbol{X}_{i,j}) = E(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)\boldsymbol{X}_{i,j}). \tag{2.17}$$

If $\beta_{ij}^M = 0$, by the equation (2.16), the first 3 components of $\boldsymbol{\beta}_{ij}^M$, should be equal to $\boldsymbol{\beta}_{i,j}^M$ by the uniqueness of the solution of the score equation (2.17). Therefore, the score equation (2.15) on the component X_{ij} entails

$$E\{b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)X_{ij}\} = E(YX_{ij}), \tag{2.18}$$

it follows that $E\{(Y - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))X_{ij}\} = 0$, i.e., $Cov(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) = 0$.
On the other hand, similarly, if $E\{(Y - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))X_{ij}\} = 0$, we have the equation (2.18) holds. And then, together with the equation (2.17), we can know that $((\boldsymbol{\beta}_{i,j}^M)^T, 0)^T$ is a solution to the equation (2.15). By the property of solution's uniqueness, it indicates that $\boldsymbol{\beta}_{ij}^M = ((\boldsymbol{\beta}_{i,j}^M)^T, 0)^T$, which implies that $\beta_{ij}^M = 0$.

Proof of Corollary 2.2.1: By Theorem 2.2.1 and 2.2.2, we can easily obtain this

Corollary.

Proof of Theorem 2.2.3: Denote that the matrix $\mathbf{A} = E(m_{ij}\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T)$ and partition it as

$$\mathbf{A} = \begin{bmatrix} E(m_{ij}\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^T) & E(m_{ij}\boldsymbol{X}_{i,j}X_{ij}) \\ E(m_{ij}X_{ij}\boldsymbol{X}_{i,j}^T) & E(m_{ij}X_{ij}^2) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

Hence, the matrix $\mathbf{A}$ is a positive definite matrix. As a matter of fact, by the convexity of the function $b(\cdot)$, $m_{ij} > 0$ almost surely. Therefore, for any nonzero constant vector a, $a^T\mathbf{A}a = E(m_{ij}a^T\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^Ta) = E(m_{ij}a^T\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^Ta) = E(m_{ij}(a^T\boldsymbol{X}_{ij})^2) > 0$ and the inverse matrix A_{11}^{-1} exists.

Based on the equation (2.16) and (2.17), we obtain that

$$E\{b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M)\boldsymbol{X}_{i,j}\} = E\{b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)\boldsymbol{X}_{i,j}\},$$

i.e., $E\{[b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)]\boldsymbol{X}_{i,j}\} = 0$. Let $\Delta_{\boldsymbol{\beta}_{ij}} = (\beta_{ij0}^M, \beta_i^M, \beta_j^M)^T - \boldsymbol{\beta}_{i,j}^M$ and by the definition of m_{ij}, we can get

$$E\{m_{ij}(\boldsymbol{X}_{i,j}^T\Delta_{\boldsymbol{\beta}_{ij}} + X_{ij}\beta_{ij}^M)\boldsymbol{X}_{i,j}\} = 0,$$

that is, $\Delta_{\boldsymbol{\beta}_{ij}} = -A_{22}^{-1}A_{12}\beta_{ij}^M$. Moreover,

$$\begin{aligned} \mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) &= E(Y - E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))X_{ij} \\ &= E\{[b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)]X_{ij}\} \\ &= E\{m_{ij}[\boldsymbol{X}_{i,j}^T\Delta_{\boldsymbol{\beta}_{ij}} + X_{ij}\beta_{ij}^M]X_{ij}\} \\ &= A_{21}\Delta_{\boldsymbol{\beta}_{ij}} + A_{22}\beta_{ij}^M \\ &= (A_{22} - A_{21}A_{22}^{-1}A_{12})\beta_{ij}^M. \end{aligned}$$

By the positive definiteness of Matrix A, $A_{22} - A_{21}A_{22}^{-1}A_{12} > 0$, hence, by Condition (B),

$$|\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)| = |A_{22} - A_{21}A_{22}^{-1}A_{12}||\beta_{ij}^M| \le A_{22}|\beta_{ij}^M| \le c_2|\beta_{ij}^M|.$$

Finally, let $c_3 = \frac{c_1}{c_2}$, we can clarify that $|\beta_{ij}^M| \geq c_3 n^{-\kappa}$.The conclusion follows. $\square$

Proof of Theorem 2.2.4: If Condition (B) holds, by Theorem 2.2.3, we have $|\beta_{ij}^M| \geq c_3 n^{-\kappa}$ for $1 \leq i < j \leq p$. And then, by Condition (C), we have

$$\begin{aligned} L_{ij}^{\star} &= E\{l(\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{ij}^M)\} \\ &= E\{l(\boldsymbol{X}_{ij}^T((\boldsymbol{\beta}_{i,j}^M)^T, 0)^T, Y) - l(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{ij}^M)\} \\ &\geq V\|(\boldsymbol{\beta}_{i,j}^M)^T, 0)^T - \boldsymbol{\beta}_{ij}^M\|^2 \\ &\geq V|\beta_{ij}^M|^2 \\ &\geq V c_3^2 n^{-2\kappa}. \end{aligned}$$

Finally, let $c_4 = V c_3^2$, then we can get $\min_{(i,j)\in\mathcal{N}^\star} L_{ij}^{\star} \geq c_4 n^{-2\kappa}$. $\square$

In order to prove Theorem 2.2.5, we need some results, which are shown in Fan and Song (2010) and Barut *et al.* (2016), and they are listed as follows.

Denote $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} El(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ as the population parameters. Suppose that $\boldsymbol{\beta}_0$ is an interior point of one sufficiently large, convex and compact set $\mathbf{B} \subseteq \mathbf{R}^p$. The following several conditions are required in our proof.

(A1) The Fisher information

$$\boldsymbol{I}(\boldsymbol{\beta}) = E\left\{\left[\frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\right]\left[\frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\right]^T\right\}$$

is finite and $\boldsymbol{I}(\boldsymbol{\beta}_0)$ is a positive definite matrix. Furthermore,

$$\|\boldsymbol{I}(\boldsymbol{\beta})\|_{\mathbf{B}} = \sup_{\boldsymbol{\beta}\in\mathbf{B}, \|\boldsymbol{x}=1\|} \|\boldsymbol{I}(\boldsymbol{\beta})^{1/2}\boldsymbol{x}\|$$

exists.

(B1) The function $l(\boldsymbol{x}^T\boldsymbol{\beta}, Y)$ satisfy the Lipschitz property with positive constant k_n, which means that for any $\boldsymbol{\beta}$,$\boldsymbol{\beta}' \in \mathbf{B}$ and $(\boldsymbol{x}, y) \in \Omega_n = \{(\boldsymbol{x}, y) : \|\boldsymbol{x}\|_\infty \leq K_n, |y| \leq$

$K_n^\star\}$ with sufficiently large positive constant K_n and $K_n^\star$, we have

$$|l(\boldsymbol{x}^T\boldsymbol{\beta}, Y) - l(\boldsymbol{x}^T\boldsymbol{\beta}', Y)| \leq k_n|\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}'|.$$

Furthermore, there exists a sufficiently large constant C such that

$$\sup_{\boldsymbol{\beta}\in\mathbf{B},\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq b_n} |E\{[l(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)](1 - I_n(\boldsymbol{X}, Y))\}| \leq o(p/n),$$

in which $b_n = Ck_nV_n^{-1}(p/n)^{1/2}$, V_n is defined in Condition (C1) and $I_n(\boldsymbol{x}, y) = I((\boldsymbol{x}, y) \in \Omega_n)$.

(C1) The function $l(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ is convex with respect to $\boldsymbol{\beta}$ and

$$E[l(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)] \geq V_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2,$$

for some positive constant V_n and here $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq b_n$, b_n is defined in Condition (B1).

The proof of Theorem 2.2.5 needs an exponential bound for the tail probability of the quasi maximum likelihood estimator $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n l(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$.

Lemma 2.9.1 *(Fan and Song (2010)) Under conditions (A1)-(C1), for any $t > 0$,*

$$P\left(\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq 16K_n(1 + t)/V_n\right) \leq \exp(-2t^2/K_n^2) + nP(\Omega_n^c).$$

Lemma 2.9.2 *(Fan and Song (2010)) Under Condition (D), for any $t > 0$,*

$$P(|Y| \geq m_0t^\alpha/s_0) \leq s_1\exp(-m_0t^\alpha).$$

Lemma 2.9.3 *Under Conditions (A1)-(C1), for some positive constants c_6, c_7, c_9*

and κ, it holds that

$$\begin{aligned} & P\left(|\mathbb{P}_n\{l(\boldsymbol{X}^T\hat{\boldsymbol{\beta}},Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\}| \geq c_7 n^{-2\kappa}\right) \\ \leq \;& \exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + 2\exp\left(-c_9 n^{1-4\kappa}\right) + 2nP(\Omega_n^c). \end{aligned}$$

Proof.

$$\begin{aligned} & |\mathbb{P}_n\{l(\boldsymbol{X}^T\hat{\boldsymbol{\beta}},Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\}| \\ = \;& |\mathbb{P}_n\{l(\boldsymbol{X}^T\hat{\boldsymbol{\beta}},Y)\} - \mathbb{P}_n\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\} + \mathbb{P}_n\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\}| \\ \leq \;& |\mathbb{P}_n\{l(\boldsymbol{X}^T\hat{\boldsymbol{\beta}},Y)\} - \mathbb{P}_n\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\}| + |\mathbb{P}_n\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0,Y)\}| \\ \triangleq \;& S_1 + S_2 \end{aligned}$$

For the terms S_1, by Taylor's expansion and $\mathbb{P}_n l'(\boldsymbol{X}^T\hat{\boldsymbol{\beta}},Y) = 0$, we obtain

$$S_1 = \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T g(\boldsymbol{\xi}_n)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq \frac{1}{2} D_0 \lambda_{max}(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^T)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2,$$

where $D_0 = \sup_x b''(x)$, $\lambda_{max}(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^T)$ is the maximum eigenvalue of the sample variance matrix $\mathbb{P}_n b''(\boldsymbol{\xi}_n^T\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^T$, and $\boldsymbol{\xi}_n$ lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Based on Lemma 2.9.1 and taking $1 + t = c_5 V n^{1/2-\kappa}/(16 k_n)$,

$$P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \geq c_5^2 n^{-2\kappa}) = P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq c_5 n^{-\kappa}) \leq \exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + nP(\Omega_n^c).$$

Furthermore, by the Hoeffding inequality (Hoeffding(1963)), for a random variable X and any given $K > 0$, we have

$$P\left((\mathbb{P}_n - E)X^k \boldsymbol{I}(|X| \leq K) > \epsilon\right) \leq \exp\left(-2n\varepsilon^2/(4K^{2k})\right)$$

for any $k \geq 0$ and $\epsilon > 0$. Therefore, with the expectation on a set with negligible probability and a fixed constat p, it follows that

$$\lambda_{max}(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^T) \leq p\lambda_{max}(E\boldsymbol{X}\boldsymbol{X}^T) = O(1).$$

Consequently, $S_1 \leq D_1\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ for some $D_1 > 0$ and then taking $c_7 = D_1 c_5^2$,

$$\begin{aligned} & P(S_1 \geq c_7 n^{-2\kappa}) \leq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \geq c_5^2 n^{-2\kappa}) \\ \leq \; & \exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + nP(\Omega_n^c). \end{aligned}$$

For the term S_2, for any $\varepsilon > 0$,

$$\begin{aligned} & P\left(|\mathbb{P}_n\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}| > \varepsilon\right) \\ \leq \; & P\left(|(\mathbb{P}_n - E)\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}| > \varepsilon, \Omega_n\right) + nP(\Omega_n^c). \end{aligned}$$

Since $l(\boldsymbol{x}^T\boldsymbol{\beta}_0, y)$ satisfies the Lipschitz property, it can be bounded by some interval with length $C > 0$ on the set Ω_n for each $1 \leq i \leq n$, using Hoeffding inequality again, we can easily get

$$P\left(|(\mathbb{P}_n - E)\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}| > \varepsilon, \Omega_n\right) \leq 2\exp\left(-2n\varepsilon^2/C^2\right).$$

By taking $\varepsilon = c_7 n^{-2\kappa}$, we can obtain

$$\begin{aligned} & P\left(|\mathbb{P}_n\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}| > c_7 n^{-2\kappa}\right) \\ \leq \; & P\left(|(\mathbb{P}_n - E)\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}| > c_7 n^{-2\kappa}, \Omega_n\right) + nP(\Omega_n^c) \\ \leq \; & 2\exp\left(-2c_7^2 n^{1-4\kappa}/C^2\right) + nP(\Omega_n^c). \end{aligned}$$

Finally, taking $c_9 = 2c_7^2/C^2$, we get

$$\begin{aligned} & P\left(|\mathbb{P}_n\{l(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}, Y)\} - E\{l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}| \geq c_7 n^{-2\kappa}\right) \\ \leq \; & P(S_1 \geq c_7 n^{-2\kappa}) + P(S_2 \geq c_7 n^{-2\kappa}) \\ \leq \; & \exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + 2\exp\left(-c_9 n^{1-4\kappa}\right) + 2nP(\Omega_n^c). \end{aligned}$$

Proof of Theorem 2.2.5: (i) We want to use Lemma 2.9.1 to get the exponential bound for the tail property, hence we need to check the conditions (A1)-(C1). By the conditions (A)-(E), we can easily find that most of the conditions are satisfied expect

for the tail part of condition (B1). Now we check this part. In our case,

$$\begin{aligned}
& |E\{[l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)](1 - I_n(\boldsymbol{X}_{ij}, Y))\}| \\
= \; & |E\{[b(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}) - Y\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}) - b(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) + Y\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M](1 - I_n(\boldsymbol{X}_{ij}, Y))\}| \\
\leq \; & |E\{b(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij})I(|X_{ij} > K_n|)\}| + |E\{b(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M)I(|X_{ij} > K_n|)\}| \\
& + |E\{Y\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}(1 - I_n(\boldsymbol{X}_{ij}, Y))\}| + |E\{Y\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M(1 - I_n(\boldsymbol{X}_{ij}, Y))\}|
\end{aligned}$$

By condition (E), we can know that the first two terms are of order $o(1/n)$. For the last two terms, using the Cauchy-Schwarz inequality and Lemma 2.9.2 with $K_n^\star = m_0 K_n^\alpha / s_0$,

$$\begin{aligned}
& |E\{Y\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}(1 - I_n(\boldsymbol{X}_{ij}, Y))\}| \\
\leq \; & (E|Y\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}|^2)^{1/2} * (E|1 - I_n(\boldsymbol{X}_{ij}, Y)|^2)^{1/2} \\
\leq \; & C * (P(1 - I_n(\boldsymbol{X}_{ij}, Y)))^{1/2} \leq C * [P(|X_{ij}| > K_n) + P(|Y| > K_n^\star)]^{1/2} \\
\leq \; & C * [m_1 \exp(-m_0 K_n^{\alpha/2}) + s_1 \exp(-m_0 K_n^\alpha)]^{1/2} \\
\leq \; & C * [(m_1 + s_1) \exp(-m_0 K_n^{\alpha/2})]^{1/2}
\end{aligned}$$

When n tends to infinity, the last two terms can be very small. In summary, the tail part of condition (B1) can be satisfied. As a result, we obtain that

$$P\left(\sqrt{n}\|\hat{\boldsymbol{\beta}}_{ij}^M - \boldsymbol{\beta}_{ij}^M\| \geq 16K_n(1+t)/V\right) \leq \exp(-2t^2/K_n^2) + nP(\Omega_n^c).$$

And then, using condition (D) and Lemma 2.9.2 with $m_2 = 3m_1 + s_1$, we can get

$$\begin{aligned}
P(\Omega_n^c) \; & \leq \; P(\|\boldsymbol{X}_{ij}\|_\infty > K_n) + P(|Y| > K_n^\star) \\
& \leq \; 3m_1 \exp(-m_0 K_n^{\alpha/2}) + s_1 \exp(-m_0 K_n^\alpha) \\
& \leq \; m_2 \exp(-m_0 K_n^{\alpha/2}).
\end{aligned}$$

Next, by using the above inequalities and taking $1+t = c_5 V n^{1/2-\kappa}/(16k_n)$, it indicates

that

$$\begin{aligned} & P(|\hat{\beta}_{ij}^M - \beta_{ij}^M| \geq c_5 n^{-\kappa}) \\ \leq \; & P(\|\hat{\boldsymbol{\beta}}_{ij}^M - \boldsymbol{\beta}_{ij}^M\| \geq c_5 n^{-\kappa}) \\ \leq \; & \exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + n m_2 \exp(-m_0 K_n^{\alpha/2}) \end{aligned}$$

for some positive constant c_6. Consequently, by Bonferroni's inequality with $q = \frac{p(p-1)}{2}$, we have

$$\begin{aligned} & P\left(\max_{1\leq i<j\leq p} |\hat{\beta}_{ij}^M - \beta_{ij}^M| \geq c_5 n^{-\kappa}\right) \\ \leq \; & q\left(\exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + n m_2 \exp(-m_0 K_n^{\alpha/2})\right). \end{aligned}$$

(ii) By definition of $L_{ij,n}$ and $L_{ij}^\star$,

$$\begin{aligned} L_{ij,n} &= \mathbb{P}_n\{l(\hat{\beta}_{i,j0}^M + \hat{\beta}_{i,}^M X_i + \hat{\beta}_{j,}^M X_j, Y) \\ &\quad - l(\hat{\beta}_{ij0}^M + \hat{\beta}_i^M X_i + \hat{\beta}_j^M X_j + \hat{\beta}_{ij}^M X_{ij}, Y)\} \\ &= \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T \hat{\boldsymbol{\beta}}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T \hat{\boldsymbol{\beta}}_{ij}^M, Y)\} \end{aligned}$$

and

$$\begin{aligned} L_{ij}^\star &= E\{l(\beta_{i,j0}^M + \beta_{i,}^M X_i + \beta_{j,}^M X_j, Y) \\ &\quad - l(\beta_{ij0}^M + \beta_i^M X_i + \beta_j^M X_j + \beta_{ij}^M X_{ij}, Y)\} \\ &= E\{l(\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{ij}^M, Y)\} \end{aligned}$$

Hence,

$$\begin{aligned}
& |L_{ij,n} - L_{ij}^{\star}| \\
= \; & |\mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\hat{\boldsymbol{\beta}}_{i,j}^M, Y)\} - E\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M, Y)\} \\
& -\mathbb{P}_n\{l(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}_{ij}^M, Y)\} + E\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)\}| \\
\leq \; & |\mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\hat{\boldsymbol{\beta}}_{i,j}^M, Y)\} - E\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M, Y)\}| \\
& +|\mathbb{P}_n\{l(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}_{ij}^M, Y)\} - E\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)\}| \\
\triangleq \; & T_1 + T_2
\end{aligned}$$

By using Lemma 2.9.3, we easily get

$$\begin{aligned}
& P(|L_{ij,n} - L_{ij}^{\star}| \geq c_7 n^{-2\kappa}) \\
\leq \; & P(T_1 \geq c_7 n^{-2\kappa}) + P(T_2 \geq c_7 n^{-2\kappa}) \\
\leq \; & 2\exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp\left(-c_9 n^{1-4\kappa}\right) + 4nP(\Omega_n^c) \\
\leq \; & 2\exp(-c_6 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp\left(-c_9 n^{1-4\kappa}\right) + 4nm_2\exp(-m_0 K_n^{\alpha/2}).
\end{aligned}$$

Consequently, by Bonferroni's inequality with $q = \frac{p(p-1)}{2}$ and $c_8 = c_6$, we have

$$\begin{aligned}
& P\left(\max_{1\leq i<j\leq p}|L_{ij,n} - L_{ij}^{\star}| \geq c_7 n^{-2\kappa}\right) \\
\leq \; & q\left(2\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp\left(-c_9 n^{1-4\kappa}\right) + 4nm_2\exp(-m_0 K_n^{\alpha/2})\right).
\end{aligned}$$

(iii) Define that the event

$$A_n = \left\{\max_{(i,j)\in\mathcal{N}^{\star}}|L_{ij,n} - L_{ij}^{\star}| \leq c_4 n^{-2\kappa}/2\right\}.$$

By Theorem 2.2.4, we have $\min_{(i,j)\in\mathcal{N}_{\star}}|L_{ij}^{\star}| \geq c_4 n^{-2\kappa}$, and then for all $(i,j) \in \mathcal{N}_{\star}$,

$$L_{ij,n} = |L_{ij,n} - L_{ij}^{\star} + L_{ij}^{\star}| \geq L_{ij}^{\star} - |L_{ij,n} - L_{ij}^{\star}| \geq c_4 n^{-2\kappa}/2.$$

Taking $\gamma_n = c_{10} n^{-2\kappa}$ with $c_{10} \leq c_4/2$, we can know that $\mathcal{N}_{\star} \subset \widehat{\mathcal{N}}_{\gamma_n}$. Furthermore,

$P(A_n) \leq P(\mathcal{N}_\star \subset \widehat{\mathcal{N}}_{\gamma_n})$. And then, by Theorem 2.2.5(ii), we have

$$P(A_n^c) \leq s_n \left(2\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp\left(-c_9 n^{1-4\kappa}\right) + 4nm_2\exp(-m_0 K_n^{\alpha/2})\right).$$

Finally,

$$\begin{aligned} & P(\mathcal{N}_\star \subset \widehat{\mathcal{N}}_{\gamma_n}) \\ \geq\ & 1 - s_n \left(2\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + 4\exp\left(-c_9 n^{1-4\kappa}\right) + 4nm_2\exp(-m_0 K_n^{\alpha/2})\right). \end{aligned}$$

Proof of Theorem 2.2.6: The key idea of the proof is similar to that of Theorem 5 of Fan and Song (2010). The idea of this proof is to show that

$$\|\boldsymbol{\beta}_{\mathcal{I}}^M\|^2 = O(\lambda_{max}(\boldsymbol{\Sigma}_{\mathcal{I}})). \tag{2.19}$$

If so, by definition, we have

$$\begin{aligned} 0 \leq L_{ij}^\star &= E\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)\} \\ &\leq E\{l(\beta_{ij0}^M + \beta_i^M X_i + \beta_j^M X_j, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)\} \end{aligned}$$

Using Taylor's expansion, for some $D_2 > 0$, we have

$$E\{l(\beta_{ij0}^M + \beta_i^M X_i + \beta_j^M X_j, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)\} \leq D_2(\beta_{ij}^M)^2.$$

As a result, with vector from, we get

$$\|\boldsymbol{L}^\star\| \leq O(\|\boldsymbol{\beta}_{\mathcal{I}}^M\|^2) = O(\lambda_{max}(\boldsymbol{\Sigma}_{\mathcal{I}})).$$

Therefore, we can know that for any $\varepsilon > 0$, the number of $\{(i,j) : L_{ij}^\star > \varepsilon n^{-2\kappa}, 1 \leq i < j \leq p\}$ cannot exceed $O(\lambda_{max}(\Sigma_{\mathcal{I}}))$. Thus, on the set

$$B_n = \left\{ \max_{1\leq i<j\leq p} |L_{ij,n} - L_{ij}^\star| \leq \varepsilon n^{-2\kappa} \right\},$$

the number of $\{(i,j) : L_{ij,n} > 2\varepsilon n^{-2\kappa}, 1 \le i < j \le p\}$ cannot exceed the number of $\{(i,j) : L^{\star}_{ij} > \varepsilon n^{-2\kappa}, 1 \le i < j \le p\}$, which is bounded by $O(n^{2\kappa}\lambda_{max}(\Sigma_{\mathcal{I}}))$. By taking $\varepsilon = c_7/2$, we have

$$P\left(|\widehat{\mathcal{N}}_{\gamma_n}| \le O(n^{2\kappa}\lambda_{max})\right) \ge P(B_n).$$

Consequently, the conclusion can follow from Theorem 2.2.5(ii).

Now we want to prove the equation (2.19). By condition (B) and the proof of Theorem 2.2.3, $A_{22} - A_{21}A_{22}^{-1}A_{12}$ is uniformly bounded from blew, we have

$$|\beta^M_{ij}| \le D_3|Cov_L(Y, X_{ij}|\boldsymbol{X}^T_{i,j}\boldsymbol{\beta}^M_{i,j})|$$

for a positive constant D_3. Using the Lipschitz continuity of $b'(\cdot)$, we obtain

$$\begin{aligned}
|\beta^M_{ij}| &\le D_3|Cov_L(Y, X_{ij}|\boldsymbol{X}^T_{i,j}\boldsymbol{\beta}^M_{i,j})| \\
&= D_3|E(b'(\boldsymbol{X}^T\boldsymbol{\beta}^{\star}) - b'(\boldsymbol{X}^T_{i,j}\boldsymbol{\beta}^M_{i,j}))X_{ij}| \\
&\le D_4|EX_{ij}(\boldsymbol{X}^T\boldsymbol{\beta}^{\star} - \boldsymbol{X}^T_{i,j}\boldsymbol{\beta}^M_{i,j})| \\
&= D_4|[EX_{ij}(\boldsymbol{X}^T_{\mathcal{I}}\boldsymbol{\beta}^{\star}_{\mathcal{I}} + \boldsymbol{X}^T_{\mathcal{C}}\Delta\boldsymbol{\beta}_{ij})]|
\end{aligned}$$

for some constant $D_4 > 0$, where $\boldsymbol{\beta}^M_{ij-} = (\beta^M_{i,j0}, 0, \ldots, 0, \beta^M_{i,}, 0, \ldots, 0, \beta^M_{j,}, 0, \ldots, 0)^T$, $\Delta\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}^{\star}_{\mathcal{C}} - \boldsymbol{\beta}^M_{ij-}$. Let $R_{ij} = E[X_{ij}\boldsymbol{X}^T_{\mathcal{C}}\Delta\boldsymbol{\beta}_{ij}]$ and $\boldsymbol{R} = (R_{12}, R_{13}, \ldots, R_{(p-1)p})^T$, hence,

$$|\beta^M_{ij}|^2 \le D_4^2|E[X_{ij}\boldsymbol{X}^T_{\mathcal{I}}\boldsymbol{\beta}^{\star}_{\mathcal{I}}] + R_{ij}|^2$$

and

$$\|\boldsymbol{\beta}^M_{\mathcal{I}}\|^2 \le D_4^2D_5^2\|E[\boldsymbol{X}_{\mathcal{I}}\boldsymbol{X}^T_{\mathcal{I}}\boldsymbol{\beta}^{\star}_{\mathcal{I}}] + \boldsymbol{R}\|^2$$

Therefore,

$$\begin{aligned} & \|E[\boldsymbol{X}_{\mathcal{I}}\boldsymbol{X}_{\mathcal{I}}^T\boldsymbol{\beta}_{\mathcal{I}}^\star] + \boldsymbol{R}\|^2 = \|\boldsymbol{\Sigma}_{\mathcal{I}}\boldsymbol{\beta}_{\mathcal{I}}^\star + \boldsymbol{R}\|^2 \\ = \;& \boldsymbol{\beta}_{\mathcal{I}}^{\star T}\boldsymbol{\Sigma}_{\mathcal{I}}^2\boldsymbol{\beta}_{\mathcal{I}}^\star + 2\boldsymbol{R}^T\boldsymbol{\Sigma}_{\mathcal{I}}\boldsymbol{\beta}_{\mathcal{I}}^\star + \boldsymbol{R}^T\boldsymbol{R} \\ \leq \;& \lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{I}})\boldsymbol{\beta}_{\mathcal{I}}^{\star T}\boldsymbol{\Sigma}_{\mathcal{I}}\boldsymbol{\beta}_{\mathcal{I}}^\star + 2\boldsymbol{R}^T\boldsymbol{\Sigma}_{\mathcal{I}}\boldsymbol{\beta}_{\mathcal{I}}^\star + \boldsymbol{R}^T\boldsymbol{R} \\ \leq \;& \lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{I}})Var(\boldsymbol{X}^T\boldsymbol{\beta}^\star) + 2\boldsymbol{R}^T\boldsymbol{\Sigma}_{\mathcal{I}}\boldsymbol{\beta}_{\mathcal{I}}^\star + \boldsymbol{R}^T\boldsymbol{R}. \end{aligned}$$

Since $\text{Var}(\boldsymbol{X}^T\boldsymbol{\beta}^\star) = O(1)$, and by Condition (H), we can get that

$$\|\boldsymbol{\beta}_{\mathcal{I}}^M\|^2 = O(\lambda_{max}(\boldsymbol{\Sigma}_{\mathcal{I}})).$$

Proof of Theorem 2.4.1: (1)Here, the idea of proof is similar to the proof of Theorem 1 of Li et al. [2012a]. Denote that $X_k^\star = \Phi^{-1}[F_k^X(X_k)]$ and $Y^\star = \Phi^{-1}[F^Y(Y)]$, where F_k^X and F^Y are the cumulative distribution functions of X_k and Y, respectively; Φ^{-1} is the standard normal distribution function.
For the condition of Fan and Song (2010), it is equivalent to

$$|\text{Cov}(Y,\ X_k)| \geq C_1 n^{-\kappa}.$$

Indeed, $\text{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star),\ X_k) = E(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star)X_k) = E(E(Y|\boldsymbol{X})X_k) = E(YX_k) = \text{Cov}(Y, X_k)$. Since Y and X_k are standardized, we obtain that $|\rho_k| \geq C_1 n^{-\kappa}$.
Firstly, we consider the special case $l = m = 2$ and then $\widetilde{Y} = I(Y > M_d(Y))$ and $\widetilde{X}_k = \widetilde{X}_{k_2} = I(X_k > M_d(X_k))$. We only need to prove that $|\text{Cov}(\widetilde{Y},\ \widetilde{X}_k)| \geq C_2 n^{-\kappa}$ for some positive constant C_2.
Furthermore, assume that $\rho_k \geq C_1 n^{-\kappa}$ and let

$$X_{1k}^\star = \Phi^{-1}[F_k^X(X_{1k})] \;\text{ and }\; X_{2k}^\star = \Phi^{-1}[F_k^X(X_{2k})]$$

and $Y_1^\star = \Phi^{-1}[F^Y(Y_1)]$, $Y_2^\star = \Phi^{-1}[F^Y(Y_2)]$, thus, $\frac{1}{\sqrt{2}}(X_{2k}^\star - X_{1k}^\star)$ and $\frac{1}{\sqrt{2}}(Y_2^\star - Y_1^\star)$

follow the standard normal distribution. Consequently,

$$\begin{aligned}
\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) &= \mathrm{Cov}(I(Y > M_d(Y), I(X_k > M_d(X_k))) \\
&= \mathrm{Cov}(I(Y^\star > 0), I(X_k^\star > 0)) \\
&= E(I(Y^\star > 0)I(X_k^\star > 0)) - \frac{1}{4} \\
&= E\left\{ I\left(\frac{1}{\sqrt{2}}(Y_2^\star - Y_1^\star) > 0\right) I\left(\frac{1}{\sqrt{2}}(X_{2k}^\star - X_{1k}^\star) > 0\right)\right\} - \frac{1}{4} \\
&= E\left\{I(X_{2k}^\star > X_{1k}^\star)I(Y_2^\star > Y_1^\star)\right\} - \frac{1}{4}
\end{aligned}$$

Since the function $\Phi^{-1} \cdot F_k^X$ and $\Phi^{-1} \cdot F^Y$ are two increasing functions, their inverse functions are also increasing, therefore, we have

$$\begin{aligned}
\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) &= E\left\{I(X_{2k} > X_{1k})I(Y_2 > Y_1)\right\} - \frac{1}{4} \\
&= E\left\{I(X_{2k} - X_{1k} > 0)I(Y_1 - Y_2 < 0)\right\} - \frac{1}{4} \\
&= E\left\{I(X_{2k} - X_{1k} > 0)I(\Delta\varepsilon_k < \rho_k(X_{1k} - X_{2k}))\right\} - \frac{1}{4}
\end{aligned}$$

Taking into account the symmetry of $f_{\Delta\varepsilon_k|\Delta X_k}(t)$, we know that

$$1 - F_{\Delta\varepsilon_k|\Delta X_k}(-t) = F_{\Delta\varepsilon_k|\Delta X_k}(t)$$

and

$$F_{\Delta\varepsilon_k|\Delta X_k}(0) = \frac{1}{2},$$

where $F_{\Delta\varepsilon_k|\Delta X_k}(\cdot)$ is the cumulative distribution function of $\Delta\varepsilon_k$ given ΔX_k. Hence,

$$\begin{aligned}
&\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) \\
= \ & E\left\{I(X_{2k} > X_{1k})F_{\Delta\varepsilon_k|\Delta X_k}(\rho_k(X_{1k} - X_{2k}))\right\} \\
& -E\left\{I(X_{2k} > X_{1k})\right\} F\left\{\Delta\varepsilon_k < 0|\Delta X_k\right\} \\
= \ & E\left\{I(X_{2k} - X_{1k} > 0)[F_{\Delta\varepsilon_k|\Delta X_k}(\rho_k(X_{2k} - X_{1k})) - F_{\Delta\varepsilon_k|\Delta X_k}(0)]\right\} \\
= \ & E\left\{I(X_{2k} - X_{1k} > 0)\int_0^{\rho_k(X_{2k}-X_{1k})} f_{\Delta\varepsilon_k|\Delta X_k}(t)dt\right\}
\end{aligned}$$

According to Condition (M1),

$$\begin{aligned}
\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) &= E\Big\{ I(X_{2k}-X_{1k}>0) \\
&\times \int_0^{\rho_k(X_{2k}-X_{1k})} \left[\pi_{0k} f_0(t,\sigma_0^2|\Delta X_k) + (1-\pi_{0k}) f_1(t,\sigma_1^2|\Delta X_k)\right] dt \Big\} \\
&\geq \pi_{0k} E\left\{ I(X_{2k}-X_{1k}>0) \int_0^{\rho_k(X_{2k}-X_{1k})} f_0(t,\sigma_0^2|\Delta X_k)dt \right\}.
\end{aligned}$$

By the Gaussian inequality for the symmetric unimodal distribution (See Pukelsheim [1994], and Sellke and Sellke [1997]),

$$P(|X|\geq k\sigma) \leq \begin{cases} 1-\frac{k}{\sqrt{3}}, & k\leq \frac{2}{\sqrt{3}}, \\ \frac{4}{9K^2}, & k\leq \frac{2}{\sqrt{3}}, \end{cases}$$

therefore,

$$P(|X|\geq k\sigma) \leq \frac{1}{1+k/\sqrt{3}},$$

where X is a unimodal random variable with a mode at the origin zero and variance σ^2. Using this Gaussian inequality, we can get

$$\begin{aligned}
\int_0^{\rho_k(X_{2k}-X_{1k})} f_0(t,\sigma_0^2|\Delta X_k)dt &= \left\{ \int_0^{\infty} - \int_{\rho_k(X_{2k}-X_{1k})}^{\infty} \right\} f_0(t,\sigma_0^2|\Delta X_k)dt \\
&= \frac{1}{2} - P(\Delta\varepsilon_k > \rho_k(X_{2k}-X_{1k})) \\
&\geq \frac{1}{2} - \frac{1}{2}\frac{1}{1+\frac{\rho_k(X_{2k}-X_{1k})}{\sqrt{3\sigma_0^2}}} \\
&= \frac{\rho_k(X_{2k}-X_{1k})}{\sqrt{12\sigma_0^2}+2\rho_k(X_{2k}-X_{1k})}.
\end{aligned}$$

Since

$$Var(\Delta\varepsilon_k|\Delta X_k) = \pi_{0k}\sigma_0^2 + (1-\pi_{0k})\sigma_1^2 \geq \pi_{0k}\sigma_0^2 \geq \pi^\star\sigma_0^2,$$

we have

$$\int_0^{\rho_k(X_{2k}-X_{1k})} f_0(t,\sigma_0^2|\Delta X_k)dt \geq \frac{\rho_k(X_{2k}-X_{1k})}{\sqrt{12Var(\Delta\varepsilon_k|\Delta X_k)/\pi^\star}+2\rho_k(X_{2k}-X_{1k})}.$$

Define the variable $Z_k = \sqrt{Var(\Delta\varepsilon_k|\Delta X_k)} = \sqrt{Var(\Delta Y-\rho_k\Delta X_k|\Delta X_k)}$ and note

that $Var(\Delta\varepsilon_k) = Var(\Delta Y - \rho_k \Delta X_k) = 2(1-\rho_k^2)$. By Condition (M1), we know that

$$E(\Delta\varepsilon_k|\Delta X_k) = 0,$$

and then by the law of total variance,

$$Var(\Delta\varepsilon_k) = E(Var(\Delta\varepsilon_k|\Delta X_k)) + Var(E(\Delta\varepsilon_k|\Delta X_k)) = E(Var(\Delta\varepsilon_k|\Delta X_k)).$$

Hence, for a given large positive constant T, by Markov inequality,

$$P(Z_k > T) \leq \frac{E(Z_k^2)}{T^2} = \frac{E(Var(\Delta\varepsilon_k|\Delta X_k))}{T^2} = \frac{Var(\Delta\varepsilon_k)}{T^2} \leq \frac{2}{T^2},$$

that is,

$$P(Var(\Delta\varepsilon_k|\Delta X_k) > T^2) \leq \frac{2}{T^2},$$

which means that at least probability $1 - \frac{2}{T^2}$, we get

$$\begin{aligned}\int_0^{\rho_k(X_{2k}-X_{1k})} f_0(t, \sigma_0^2|\Delta X_k)dt &\geq \frac{\rho_k(X_{2k}-X_{1k})}{\sqrt{12T^2/\pi^\star} + 2\rho_k(X_{2k}-X_{1k})} \\ &\geq \frac{\rho_k(X_{2k}-X_{1k})}{4T/\sqrt{\pi^\star} + 2\rho_k(X_{2k}-X_{1k})}.\end{aligned}$$

Consequently, by $\pi_{0k} \geq \pi^\star$,

$$\begin{aligned}\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) &\geq \pi^\star E\left\{I(X_{2k}-X_{1k}>0)\frac{\rho_k(X_{2k}-X_{1k})}{4T/\sqrt{\pi^\star} + 2\rho_k(X_{2k}-X_{1k})}\right\} I(Z_k \leq T) \\ &= \pi^\star E\left\{I(X_{2k}-X_{1k}>0)\frac{\rho_k(X_{2k}-X_{1k})}{4T/\sqrt{\pi^\star} + 2\rho_k(X_{2k}-X_{1k})}\right\} \\ &\quad -\pi^\star E\left\{I(X_{2k}-X_{1k}>0)\frac{\rho_k(X_{2k}-X_{1k})}{4T/\sqrt{\pi^\star} + 2\rho_k(X_{2k}-X_{1k})}\right\} I(Z_k > T) \\ &\triangleq I_1 + I_2.\end{aligned}$$

For the term I_1,

$$\begin{aligned} I_1 &\geq \pi^\star E\left\{I(T/\rho_k > X_{2k} - X_{1k} > 0)\frac{\rho_k(X_{2k} - X_{1k})}{4T/\sqrt{\pi^\star} + 2\rho_k(X_{2k} - X_{1k})}\right\} \\ &\geq \frac{\pi^\star\rho_k}{4T/\sqrt{\pi^\star} + 2T} E\left\{(X_{2k} - X_{1k})I(T/\rho_k > X_{2k} - X_{1k} > 0)\right\} \\ &\geq \frac{\pi^\star\rho_k}{4T/\sqrt{\pi^\star} + 2T} \\ &\quad \times E\left\{(X_{2k} - X_{1k})I(X_{2k} - X_{1k} > 0) - (X_{2k} - X_{1k})I(X_{2k} - X_{1k} > T)\right\} \end{aligned}$$

Based on the inequality $E|X-Y| \geq E|X|$, where X and Y are i.i.d random variables with $E(X) = E(Y) = 0$, and Condition (M2), we know that

$$E\left\{|X_{2k} - X_{1k}|\right\} \geq E|X_{1k}| \geq c_{\mathcal{M}_\star},$$

and then, by the symmetry property of the distribution of $X_{2k} - X_{1k}$,

$$E\left\{(X_{2k} - X_{1k})I(X_{2k} - X_{1k} > 0)\right\} \geq \frac{1}{2}c_{\mathcal{M}_\star}.$$

On the other hand, according to Cauchy-Schwarz inequality,

$$E\left\{(X_{2k} - X_{1k})I(X_{2k} - X_{1k} > T)\right\} \leq \sqrt{P(X_{2k} - X_{1k} > T)E(X_{2k} - X_{1k})^2} \leq \frac{2}{T}.$$

Consequently,

$$I_1 \geq \frac{\pi^\star\rho_k c_{\mathcal{M}_\star}}{8T/\sqrt{\pi^\star} + 4T} - \frac{2\pi^\star\rho_k}{4T^2/\sqrt{\pi^\star} + 2T^2} = \frac{\pi^\star\rho_k c_{\mathcal{M}_\star}}{8T/\sqrt{\pi^\star} + 4T} - \frac{\pi^\star\rho_k}{2T^2/\sqrt{\pi^\star} + T^2}.$$

As for the term I_2, using Cauchy-Schwarz inequality again,

$$\begin{aligned} I_2 &= -\pi^\star E\left\{I(X_{2k} - X_{1k} > 0)\frac{\rho_k(X_{2k} - X_{1k})}{4T/\sqrt{\pi^\star} + 2\rho_k(X_{2k} - X_{1k})}\right\} I(Z_k > T) \\ &\geq -\frac{\pi^\star\rho_k}{4T/\sqrt{\pi^\star}} E\left\{(X_{2k} - X_{1k})I(X_{2k} - X_{1k} > 0)\right\} I(Z_k > T) \\ &\geq -\frac{\pi^\star\rho_k}{4T/\sqrt{\pi^\star}} \sqrt{E\left\{(X_{2k} - X_{1k})^2 I(X_{2k} - X_{1k} > 0)\right\}}\sqrt{P(Z_k > T)} \\ &\geq -\frac{\pi^\star\rho_k}{4T/\sqrt{\pi^\star}} \cdot \sqrt{2} \cdot \frac{\sqrt{2}}{T} = -\frac{\pi^\star\rho_k}{2T^2/\sqrt{\pi^\star}}. \end{aligned}$$

Combing the above two inequalities for the terms I_1 and I_2,

$$\begin{aligned}\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) \geq I_1 + I_2 &\geq \frac{\pi^\star \rho_k c_{\mathcal{M}_\star}}{8T/\sqrt{\pi^\star} + 4T} - \frac{\pi^\star \rho_k}{2T^2/\sqrt{\pi^\star} + T^2} - \frac{\pi^\star \rho_k}{2T^2/\sqrt{\pi^\star}} \\ &\geq \frac{(\pi^\star)^2 \rho_k c_{\mathcal{M}_\star}}{12T} - \frac{5\pi^\star \rho_k}{6T^2}.\end{aligned}$$

Taking the large positive value $T = \frac{15}{c_{\mathcal{M}_\star}\pi^\star}$, it is easy to infer that

$$\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) \geq \left[\frac{(\pi^\star)^2 c_{\mathcal{M}_\star}}{12}\frac{c_{\mathcal{M}_\star}\pi^\star}{15} - \frac{5\pi^\star}{6}\frac{(c_{\mathcal{M}_\star}\pi^\star)^2}{15^2}\right]\rho_k = \frac{(\pi^\star)^3 c_{\mathcal{M}_\star}^2}{540}\rho_k \geq C_1' n^{-\kappa},$$

where the positive constant $C_1' = C_1(\pi^\star)^3 c_{\mathcal{M}_\star}^2/540$.
If $\rho_k \leq C_1 n^{-\kappa}$, by the similar steps as above, we also can get $\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_k) \leq -C_1' n^{-\kappa}$. In summary, if the condition (M1)-(M3) hold and $|\mathrm{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star),\ X_k)| \geq C_1 n^{-\kappa}$ for any $k \in \mathcal{M}_\star$ with a constant $C_1 > 0$, and after discretizing the response and predictor, there exists a positive constant $C_2 = C_1'$ such that $|\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_{k_2})| \geq C_2 n^{-\kappa}$ in the special case $l = m = 2$. Furthermore, following the above same steps, we also obtain that $|\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_{k_1})| \geq C_2 n^{-\kappa}$ when $l = m = 2$.

In the following part, we will consider the general case: $m = 2$ and $l \geq 3$. By the above proof for the special case $l = 2$, we can know that if we only divide the predictor into two parts, actually, we must infer that for some positive constant C_1',

$$|\mathrm{Cov}(I(Y > M_d(Y), I(X > M_d(X_k))| \geq C_1' n^{-\kappa},$$

and

$$|\mathrm{Cov}(I(Y > M_d(Y), I(X < M_d(X_k))| \geq C_1' n^{-\kappa},$$

As for case $l > 2$, when l is a even number,

$$I(X_k > M_d(X_k)) = \bigcup_{i=\frac{l}{2}+1}^{l} I\left(X_k \in P_i^{X_k}\right),$$

and

$$I(X_k < M_d(X_k)) = \bigcup_{i=1}^{\frac{l}{2}} I\left(X_k \in P_i^{X_k}\right).$$

Hence,

$$\begin{aligned}
&\mathrm{Cov}(I(Y > M_d(Y), I(X > M_d(X_k)) \\
= \;& \mathrm{Cov}\left(I(Y > M_d(Y),\ \bigcup_{i=\frac{l}{2}+1}^{l} I\left(X_k \in P_i^{X_k}\right)\right) \\
= \;& \sum_{i=\frac{l}{2}+1}^{l} Cov\left(I(Y > M_d(Y),\ I\left(X_k \in P_i^{X_k}\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(I(Y > M_d(Y), I(X < M_d(X_k)) &= \mathrm{Cov}\left(I(Y > M_d(Y), \bigcup_{i=1}^{\frac{l}{2}} I\left(X_k \in P_i^{X_k}\right)\right) \\
&= \sum_{i=1}^{\frac{l}{2}} Cov\left(I(Y > M_d(Y),\ I\left(X_k \in P_i^{X_k}\right)\right)
\end{aligned}$$

which means that there exists at least one term $i = 1, \ldots, \frac{l}{2}$ or $i = \frac{l}{2} + 1, \ldots, l$, such that

$$\left|\mathrm{Cov}\left(I(Y > M_d(Y),\ I\left(X_k \in P_i^{X_k}\right)\right)\right| \geq C_2 n^{-\kappa},$$

where $C_2 = \frac{2}{l} C_1'$ is a positive constant number, that is,

$$|\mathrm{Cov}(\widetilde{Y},\ \widetilde{X}_{k_i})| \geq C_2 n^{-\kappa}.$$

When l is odd, the support set of X_k is divided into l parts and denote that $\{Q_i\}_{i=1}^{l-1}$ are a series of cutting points ($l-$quantiles), and then $P_1^{X_k} = (-\infty,\ Q_1)$, $P_l^{X_k} = [Q_{l-1},\ \infty)$, $P_i^{X_k} = [Q_i,\ Q_{i+1})$, for $1 < i < l$. Therefore,

$$\begin{aligned}
\{X_k > M_d(X_k)\} &= \left[M_d(X_k), Q_{\frac{l+1}{2}}\right) \bigcup \left[Q_{\frac{l+1}{2}}, Q_{\frac{l+3}{2}}\right) \bigcup \cdots \bigcup [Q_{l-1},\ \infty) \\
&= \left[M_d(X_k), Q_{\frac{l+1}{2}}\right) \bigcup \bigcup_{i=\frac{l+3}{2}}^{l} \left(X_k \in P_i^{X_k}\right),
\end{aligned}$$

and

$$\begin{aligned}\{X_k < M_d(X_k)\} &= (-\infty,\ Q_1) \bigcup \cdots \bigcup \left[Q_{\frac{l-3}{2}}, Q_{\frac{l-1}{2}}\right) \bigcup \left[Q_{\frac{l-1}{2}},\ M_d(X_k)\right) \\ &= \bigcup_{i=1}^{\frac{l-1}{2}} \left(X_k \in P_i^{X_k}\right) \bigcup \left[Q_{\frac{l-1}{2}},\ M_d(X_k)\right).\end{aligned}$$

Based on two results of the case $l = m = 2$, we can conclude that two cases will happen.

Case (1): There exists at least one term $i = 1, \ldots, \frac{l-1}{2}$ or $i = \frac{l+3}{2}, \ldots, l$, such that

$$\left|\mathrm{Cov}\left(I(Y > M_d(Y),\ I\left(X_k \in P_i^{X_k}\right)\right)\right| \geq \frac{2}{l+1} C_1' n^{-\kappa}.$$

Take $C_2 = \frac{2}{l+1} C_1'$, our proof will be completed.

Case (2): For all $i = 1, \ldots, \frac{l-1}{2}$ and $i = \frac{l+3}{2}, \ldots, l$, we have

$$\left|\mathrm{Cov}\left(I(Y > M_d(Y),\ I\left(X_k \in P_i^{X_k}\right)\right)\right| < \frac{2}{l+1} C_1' n^{-\kappa};$$

but

$$\left|\mathrm{Cov}\left\{I(Y > M_d(Y),\ I\left(X_k \in \left[M_d(X_k), Q_{\frac{l+1}{2}}\right)\right)\right\}\right| \geq \frac{2}{l+1} C_1' n^{-\kappa}$$

and

$$\left|\mathrm{Cov}\left\{I(Y > M_d(Y),\ I\left(X_k \in \left[Q_{\frac{l-1}{2}},\ M_d(X_k)\right)\right)\right\}\right| \geq \frac{2}{l+1} C_1' n^{-\kappa}.$$

In this case, $P_{\frac{l+1}{2}}^{X_k} = \left[Q_{\frac{l-1}{2}},\ M_d(X_k)\right) \bigcup \left[M_d(X_k), Q_{\frac{l+1}{2}}\right)$, and

$$\begin{aligned}&\mathrm{Cov}\left(I(Y > M_d(Y),\ I\left(X_k \in P_{\frac{l+1}{2}}^{X_k}\right)\right) \\ =\ &\mathrm{Cov}\left\{I(Y > M_d(Y),\ I\left(X_k \in \left[M_d(X_k), Q_{\frac{l+1}{2}}\right)\right)\right\} \\ +\ &\mathrm{Cov}\left\{I(Y > M_d(Y),\ I\left(X_k \in \left[Q_{\frac{l-1}{2}},\ M_d(X_k)\right)\right)\right\}\end{aligned}$$

It is easy to infer that

$$\left|\mathrm{Cov}\left(I(Y > M_d(Y),\ I\left(X_k \in P_{\frac{l+1}{2}}^{X_k}\right)\right)\right| > \frac{4}{l+2} C_1' n^{-\kappa}.$$

Finally, if the condition (M1)-(M2) hold and $|\mathrm{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star),\ X_k)| \geq C_1 n^{-\kappa}$ for any

$k \in \mathcal{M}_\star$ with a positive constant C_1, and after using 2-quantile and $l-$quantiles to discretize the response Y and the predictor X_k, there exists at least one $\widetilde{X}_{k_i}$ such that $|\text{Cov}(\widetilde{Y},\ \widetilde{X}_{k_i})| \geq C_2 n^{-\kappa}$ for some positive constant C_2, which is dependent on l.

(ii) Assume that $\widetilde{X}_{k_i}$ satisfies $|\text{Cov}(\widetilde{Y},\ \widetilde{X}_{k_i})| \geq C_2 n^{-\kappa}$ for some constant $C_2 > 0$ and $\widetilde{\boldsymbol{X}}_{k_i} = (1, \widetilde{X}_{k_i})^T$. The coefficient $\widetilde{\boldsymbol{\beta}}_{k_i}^M$ is denoted as the minimizer of the componentwise regression

$$\widetilde{\boldsymbol{\beta}}_{k_i}^M = (\widetilde{\beta}_{k_i,0}^M,\ \widetilde{\beta}_{k_i}^M) = \underset{\widetilde{\beta}_0,\ \widetilde{\beta}_{k_i}}{\arg\min}\, El(\widetilde{\beta}_0 + \widetilde{\beta}_{k_i}\widetilde{X}_{k_i},\ \widetilde{Y}).$$

Define that $\widetilde{\mathcal{M}}_\star = \{1 \leq j \leq \tilde{p},\ \widetilde{\boldsymbol{\beta}}_j^\star \neq 0\}$ be the true sparse model, where

$$\widetilde{\boldsymbol{\beta}}^\star = (\widetilde{\boldsymbol{\beta}}_0^\star,\ \widetilde{\boldsymbol{\beta}}_1^\star,\ \ldots,\ \widetilde{\boldsymbol{\beta}}_{\tilde{p}}^\star),$$

when we consider the new categorical predictor $\widetilde{\boldsymbol{X}} = \{1,\ \widetilde{X}_1,\ \widetilde{X}_2,\ \ldots,\ \widetilde{X}_{\tilde{p}}\}$ and response $\widetilde{Y}$. Hence,

$$\left|\text{Cov}(b'(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}}^\star),\ \widetilde{X}_{k_i})\right| = \left|\text{Cov}(\widetilde{Y},\ \widetilde{X}_{k_i})\right| \geq C_2 n^{-\kappa}.$$

Based on Theorem 3 of Fan and Song (2010), we have $\left|\widetilde{\beta}_{k_i}\right| \geq C_2' n^{-\kappa}$ for some positive constant C_2'. Consequently, by the condition (C$'$) of Fan and Song (2010), we obtain

$$\begin{aligned}
\widetilde{L}_k^\star &= E\left\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_k^T\widetilde{\boldsymbol{\beta}}_k^M, \widetilde{Y})\right\} \\
&\geq E\left\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_{k_i}^T\widetilde{\boldsymbol{\beta}}_{k_i}^M, \widetilde{Y})\right\} \\
&\geq V\left|\widetilde{\beta}_{k_i}\right|^2 \geq C_3 n^{-2\kappa}
\end{aligned}$$

where V is some positive constant and $C_3 = V(C_2')^2$.

Proof of Theorem 2.4.2: By the definition of $\text{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)$,

$$\text{Cov}_L(Y, \boldsymbol{X}_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) = E\{(Y - b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))X_{ij}\} = \text{Cov}(\zeta_{ij}, X_{ij}).$$

After discretizing Y, X_i and X_j, i.e., $Y = \sum_{k=1}^2 YI(Y \in P_k^Y)$, $X_i = \sum_{s=1}^{l_1} X_i I(X_i \in$

$P_s^{X_i}$) and $X_t = \sum_{t=1}^{l_2} X_j I(X_j \in P_t^{X_j})$, X_{ij} is transformed into

$$X_{ij} = \sum_{s,t} X_{ij} I\left(\left\{X_i \in P_s^{X_i}\right\} \bigcap \left\{X_j \in P_t^{X_j}\right\}\right), \quad 1 \le s \le l_1, \quad 1 \le t \le l_2.$$

Hence, the support set of ζ_{ij} becomes the union of several intervals. Suppose that $\zeta_{ij} = \sum_{k'} \zeta_{ij} I(\zeta_{ij} \in \Omega_{k'})$, where $1 \le k' \le 2l_1 l_2$. By taking ζ_{ij} as the response Y of Theorem 2.4.1 and X_{ij} as the predictor X_k in Theorem 2.4.1, there exists at least one term such that

$$\left|\mathrm{Cov}\Big(I(\zeta_{ij} \in \Omega_{k'}), I\left(\left\{X_i \in P_s^{X_i}\right\} \bigcap \left\{X_j \in P_t^{X_j}\right\}\right)\Big)\right| \ge a_1 n^{-\kappa},$$

for some positive constant a_1, where a_1 is related to l_1 and l_2. Therefore,

$$|\mathrm{Cov}(\widetilde{Y} - b'(\widetilde{\boldsymbol{X}}_{i,j}^T \widetilde{\boldsymbol{\beta}}_{i,j}^M), \widetilde{X}_{st}^{ij})| \ge a_2 n^{-\kappa},$$

that is, by taking $c_{10} = a_2$,

$$|\mathrm{Cov}_L(\widetilde{Y},\ \widetilde{X}_{st}^{ij} | \widetilde{\boldsymbol{X}}_{i,j}^T \widetilde{\boldsymbol{\beta}}_{i,j}^M)| \ge c_{10} n^{-\kappa}.$$

(2) By the similar proof of Theorem 2.2.3 and 2.2.4, we directly have the conclusion: for some positive constant c_{11}

$$\min_{i,j \in \mathcal{N}_\star} \widetilde{L}_{ij}^\star \ge c_{11} n^{-2\kappa}.$$

Chapter 3

Testing Markov Processes by Conditional Distance Covariance

3.1 Introduction

As mentioned in the section 1.2.2, the existing papers seldom discuss the testing for the Markov property in the past decades. In fact, the Markov property is one kind of conditional independent property, that is, when the current state is given, the future state and the past are conditional uncorrelated. Conditional independence has been widely studied by a lot of research in the literature. For instance, Su and White [2008] proposed a nonparametric test between random variables Y and Z given X based on the weighted Hellinger distance of two conditional density functions $f(y|x,z)$ and $f(y|z)$. Su and Ullah [2009] discussed the conditional uncorrelated in some multiple-equation models such as seemingly unrelated regressions (SURs), multivariate volatility models, and vector autoregressions (VARs). Su and White [2012] applied the local polynomial quantile regression to provide a nonparametric test for conditional independence. The empirical likelihood function was utilized by Su and White [2014] to achieve this test.

Recently, one new concept-conditional distance covariance (CDCov) was introduced by Wang et al. [2015], which can be used to identify the conditional independence of any two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ given the third one $\boldsymbol{z}$, where the sample data are independent and identically distributed. Based on this concept, the con-

ditional distance covariance can be generalized to detect the Markov property with the weakly dependent data. We will show that our test is more general and the test statistic follows an asymptotically normal distribution under the null hypothesis and a sequence of local alternatives. Simulation studies indicate that the test is well behaved in finite samples.

This chapter is organized as follows. In Section 3.2 we describe the methodology and main results about our test statistic. We derive the asymptotic properties of this statistics, including its null distribution, local alternative distribution and power properties in this section. In Section 3.3, we examine the corresponding finite sample properties of the statistic under various models. We give some conclusions in Section 3.4. The technical proofs will be given in Appendix 3.5. Throughout this chapter, $\|\cdot\|$ represents for the euclidean norm, $\mathbf{a}'$ as the transpose of any vector $\mathbf{a}$, and for any complex-valued function $f(x)$, the complex conjugate of f is denoted by $\overline{f}$ and $|f|^2 = f\overline{f}$.

3.2 Methodology and main results

3.2.1 Conditional Distance Covariance and Conditional Distance Correlation

Let U, V and X are three random vectors in Euclidean spaces $\mathbb{R}^{d_u}$, $\mathbb{R}^{d_v}$ and $\mathbb{R}^{d_x}$, respectively. The conditional distance covariance $\mathcal{D}(U, V|X)$ proposed by Wang et al. [2015] is defined as the square root of

$$\begin{aligned}\mathcal{D}^2(U,V|X) &= \|\phi_{U,V|X}(u,v) - \phi_{U|X}(u)\phi_{V|X}(v)\|^2 \\ &= \frac{1}{c_{d_u}c_{d_v}}\int_{\mathbb{R}^{d_u+d_v}} \frac{|\phi_{U,V|X}(u,v) - \phi_{U|X}(u)\phi_{V|X}(v)|^2}{|u|^{d_u+1}|v|^{d_v+1}}dudv \qquad (3.1)\end{aligned}$$

where $c_{d_u} = \frac{\pi^{(d_u+1)/2}}{\Gamma((d_u+1)/2)}$ and $c_{d_v} = \frac{\pi^{(d_v+1)/2}}{\Gamma((d_v+1)/2)}$; $\phi_{U,V|X}(u,v)$ is the conditional joint characteristic function of U, V given X; $\phi_{U|X}(u)$ and $\phi_{V|X}(v)$ are the conditional marginal functions of U, V given X, respectively. Therefore, the conditional variance is the square root of $\mathcal{D}^2(U|X) = \mathcal{D}^2(U,U|X)$. The corresponding conditional distance

correlation (CDCor) is expressed as the square root of

$$\rho^2(U,V|X) = \frac{\mathcal{D}^2(U,V|X)}{\sqrt{\mathcal{D}^2(U|X)\mathcal{D}^2(U|X)}},$$

when $\mathcal{D}^2(U|X)\mathcal{D}^2(U|X) > 0$; otherwise, $\rho(U,V|Z) = 0$. From the properties of CDCov and CDCor in Theorem 1 and Theorem 2 of Wang et al. [2015], U and V are conditionally independent given X if and only if $\mathcal{D}(U,V|X) = 0$ or $\rho(U,V|X) = 0$. Furthermore, let $f(x)$ be the density function of X and $S_a = E[\mathcal{D}^2(U,V|X)a(X)]$, where $a(\cdot)$ is a certain nonnegative function that has the same support as the function $f(x)$. As recommended by Wang et al. [2015], $a(X)$ can be chosen as $12f^4(X)$ from the point of computational consideration. Consequently, the conditional independence between U and V given X is equivalent to $S_a = 0$.

3.2.2 Conditional Distance Covariance and Markov property

Suppose that a stochastic process $\boldsymbol{X} = \{X_t\}_{t\geq 0}$ is a stationary time series process with dimension p, which is defined on the probability space $(\Omega, \mathcal{F}, P)$ with the filtration $\mathcal{F}_t$. Here, we consider the simple case $p = 1$. Assume that X_t is absolutely regular ($\beta-$mixing) with mixing coefficient $\beta(t) = O(t^{-(2+\delta')/\delta'})$, for some $\delta' > 0$, defined by

$$\beta(t) = E \sup_{A\in\mathcal{F}_{s+t}^{\infty}} |P(A|\mathcal{F}_0^s) - P(A)|,$$

where $\mathcal{F}_i^j$ is the sub $\sigma-$field of $\mathcal{F}$, generated by $\{X_t : i \leq t \leq j\}$.

Without loss of generality, assume that $\boldsymbol{X} = \{X_t\}_{t\geq 0}$ is a discrete time process. If not, we can sample it at the regular time points. The Markov Property tells us that: if the process $\boldsymbol{X} = \{X_t\}$ is Markovian, the conditional probability distribution of X_{t+1} given the information $\mathcal{F}_t$ is the same as the conditional distribution of X_{t+1} given X_t only. That is to say, given X_t, $(X_0, X_1, \ldots, X_{t-1})$ and X_{t+1} are conditionally independent, which implies that it is a conditionally independent problem. Here, we make use of the conditional covariance (CDCov) in subsection 3.2.1 to construct the test statistic for Markov property.

Let $U = X_{t+1}$, $V = (X_0, X_1, \ldots, X_{t-1})$ and $X = X_t$, for all $t \geq 1$. The Markov

property can be formally expressed as

$$\mathbb{H}_0 : P(U \leq u|X, V) = P(U \leq u|X) \text{ for all } t \geq 1. \tag{3.2}$$

Alternatively, if

$$\mathbb{H}_1 : P(U \leq u|X, V) \neq P(U \leq u|X) \quad \text{for some } t \geq 1, \tag{3.3}$$

$\boldsymbol{X} = \{X_t\}_{t\geq 0}$ is not a Markov process. Hence, by the properties of CDCov as mentioned in subsection 3.2.1, the Markov property is equivalent to $\mathcal{D}^2(U, V|X) = 0$, almost surely (a.s.) for all $t \geq 1$. Furthermore, it is equivalent to $S_a = 0$ for some $a(x) \geq 0$.

Now, we consider the nonparametric estimator of S_a. Firstly, we estimate the population conditional distance covariance $\mathcal{D}^2(U, V|X)$. The conditional characteristic functions are estimated by the Nadaraya-Watson kernel method in the definition (3.1), which was proposed by both Nadaraya [1964] and Watson [1964].

Suppose that there is a discretely observed sample $\{X_t\}_{t=0}^n$ of size n, and then we get the sample version for conditional characteristic functions of $(U, V|X)$, $U|X$ and $V|X$, respectively, that is,

$$\hat{\phi}_{U,V|X}(u, v) = \frac{\sum_{k=q}^{n-1} \exp(iuU_k + iv'V_k)\omega_k(x)}{\omega(x)}$$

and

$$\hat{\phi}_{U|X}(u) = \frac{\sum_{k=q}^{n-1} \exp(iuU_k)\omega_k(x)}{\omega(x)}, \quad \hat{\phi}_{V|X}(v) = \frac{\sum_{k=q}^{n-1} \exp(iv'V_k)\omega_k(x)}{\omega(x)},$$

where $\omega_k(x) = K_h(x - X_k)$ and $\omega(x) = \sum_{k=q}^{n-1} \omega_k(x)$ for some fixed $1 \leq q \leq n-1$, and $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$ with a kernel function K, h is the bandwidth. And the sample conditional distance covariance $\mathcal{D}_n(U, V|X)$ can be obtained by the plug-in method (pSCDCov):

$$\mathcal{D}_n^2(U, V|X) = \|\hat{\phi}_{U,V|X}(u, v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)\|^2.$$

Let $\xi_i = (U_i, V_i, X_i)$, $i = q, \ldots, n-1$, and then $\{\xi_i\}_{i=q}^{n-1}$ are stationary and weakly dependent samples with the sample size $N = n - q$. Denote the Euclidean distance of U_i and U_j in $\mathbb{R}^{d_u}$ as $d_{ij} = d(U_i, U_j)$, and similarly, d_{ij}^V for V. Let $d_{ijkl} = (d_{ij}^U + d_{kl}^U - d_{ik}^U - d_{jl}^U)(d_{ij}^V + d_{kl}^V - d_{ik}^V - d_{jl}^V)$. Note that d_{ijkl} is not symmetric with respect to $\{i, j, k, l\}$, and its symmetric form can be expressed as

$$d_{ijkl}^s = d_{ijkl} + d_{ijlk} + d_{ilkj}.$$

Hence, for the sample conditional distance covariance $\mathcal{D}_n^2(U, V|X)$ with stationary and weakly dependent samples, we have some properties of this covariance, which are similar to that in Wang et al. [2015].

Theorem 3.2.1 *$\mathcal{D}_n^2(U, V|X)$ can be divided into three parts D_1, D_2 and D_3, which means that*

$$\mathcal{D}_n^2(U, V|X) = D_1 + D_2 - 2D_3,$$

where

$$D_1 = \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) d_{kl}^U d_{kl}^V,$$

$$D_2 = \frac{1}{\omega^4(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) d_{kl}^U \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) d_{kl}^V$$

and

$$D_3 = \frac{1}{\omega^3(x)} \sum_{k=q}^{n-1} \sum_{l=q}^{n-1} \sum_{m=q}^{n-1} \omega_k(x)\omega_l(x)\omega_m(x) d_{km}^U d_{lm}^V.$$

Theorem 3.2.2 *(Consistency of Sample CDCov) If $E|X_t| < \infty$, for all $t \geq 0$, then almost surely*

$$\lim_{n\to\infty} \mathcal{D}_n^2(U, V|X) = \mathcal{D}^2(U, V|X).$$

Theorem 3.2.3 *$\mathcal{D}_n^2(U, V|X)$ is a V-statistic and can be rewritten as*

$$\mathcal{D}_n^2(U, V|X) = \frac{1}{N^4} \sum_{i,j,k,l} \psi_n(\xi_i, \xi_j, \xi_k, \xi_l; X)$$

where $\psi_n(\cdot)$ is the symmetric random kernel of degree 4 defined as (See Schick [1997]):

$$\psi_n(\xi_i, \xi_j, \xi_k, \xi_l; X) = \frac{N^4 \omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X)}{12\omega^4(X)} d^s_{ijkl}.$$

Theorem 3.2.3 indicates that the sample version of $\mathcal{D}^2(U, V|X)$ is a V-type statistic, but may not be unbiased. Here, we provide an unbiased sample conditional covariance by U-statistics, namely,

$$\mathcal{U}^2_n(U, V|X) = (C^4_N)^{-1} \sum_{i<j<k<l} \psi_n(\xi_i, \xi_j, \xi_k, \xi_l; X).$$

By Theorem 1 of Section 4.2 in Lee (1990), the V-type statistic and U-type statistic have the following relationship:

$$\mathcal{D}^2_n(U, V|X) = \frac{(N-1)(N-2)(N-3)}{N^3} \mathcal{U}^2_n(U, V|X) + o_p(1).$$

Consequently, as n is large enough, we know that $\lim_{n\to\infty} \mathcal{U}^2_n(U, V|X) = \mathcal{D}^2(U, V|X)$, almost surely. Based on these results of Theorem 3.2.1-3.2.3, the estimator of S_a can be defined as

$$S_n = \frac{1}{C^5_N h^4} \sum_{i<j<k<l<m} d^s_{ijkl} K_{im} K_{jm} K_{km} K_{lm},$$

where $K_{im} = K(\frac{X_i - X_m}{h})$ and $N = n - q$.

Actually, this test statistic is similar to one introduced by Wang et al. [2015]. The only difference is that our test statistic is based on $\beta-$mixing samples, and the independent and identically distributed samples are used in their test statistic. Therefore, this is a conditional distance covariance test with weak dependent condition (CDCTW). Furthermore, it is reasonable to delete the first q states to construct the sample statistics in order to efficiently estimate the population conditional distance covariance, since our sample data is $\beta-$mixing, the covariance $\mathrm{Cov}(X_{t+q}, X_t)$ will be close to zero when q increases, which means that X_{t+q} and X_t are almost uncorrelated.

3.2.3 Asymptotic Null distribution

To derive the null asymptotic distribution of the test statistic S_n, we need the following regular conditions. These conditions are usually imposed for the nonparametric research with the dependent data sets.

Assumption 1 (A1). $\{X_t\}_{t\geq 0}$ is a stationary β-mixing process with mixing coefficient $\beta(t) = O(t^{-(2+\delta')/\delta'})$, for some $\delta' > 0$.

Assumption 2 (A2). The density function of X_t is bounded and Lipschitz, and the conditional density function $f(\cdot|X_t)$ are twice differentiable, bounded and Lipschitz. And all of the derivatives are bounded.

Assumption 3 (A3). The kernel function K is symmetric, bounded and Lipschitz, such as the Gaussian Kernel $K_h(x) = \frac{1}{h}K(\frac{u}{h}) = \frac{1}{\sqrt{2\pi}h}\exp\{-\frac{u^2}{2h^2}\}$.

Assumption 4 (A4). For the kernel function K, $\int uK(u)du = 0$, $\int K(u)du = 1$, $\int |K(u)|du < \infty$, $\int K^2(u)du > 0$, $\int u^2K(u)du < \infty$.

Assumption 5 (A5). The bandwidth h converges to 0 such that $nh \to \infty$.

Assumption 1 is cited by many papers such as De Matos and Fernandes [2007], Su and White [2008] and Aït-Sahalia et al. [2009]. And also it can be satisfied by many processes such as autoregressive conditional heteroskedasticity (ARCH) models, autoregressive moving average (ARMA) models. Assumption 2 gives some conditions of the marginal density function and the conditional density function. Some regular conditions are imposed to the kernel functions in Assumption 3 and Assumption 4. The bandwidth of the kernel functions needs to satisfy Assumption 5. Now we point out the consistency and asymptotic distribution of our test statistics.

Theorem 3.2.4 *Suppose that Conditions A1-A5 hold. If the second moments of X_t exist, and $N = n - q$, where $1 \leq q \leq n-1$, then as $n \to \infty$, we have*

$$S_n \xrightarrow{P} S_a.$$

Theorem 3.2.4 illustrates that S_n is still consistent under the weak dependent condition, which is similar to that of Wang et al. [2015]. The detailed proof can be found in the Appendix of this chapter.

Theorem 3.2.5 *Assume that Conditions A1-A5 hold.*

(i) Under $\mathbb{H}_0$, that is, when the observed data $\{X_t\}_0^n$ are from a stationary Markov process and the second moments of X_t exist, $N = n - q$ for a fixed $1 \leq q \leq n-1$, we have

$$nh^{1/2}S_n \xrightarrow{d} N(0, \sigma^2),$$

where σ^2 is given in the Appendix.

(ii) If $\{X_t\}_0^n$ do not satisfy the Markov property, and $N = n - q$ for a fixed $1 \leq q \leq n-1$, then

$$nh^{1/2}S_n \xrightarrow{P} \infty.$$

From the above Theorem 3.2.4, we know clearly that the test statistics S_n with the dependent data also follows the asymptotic normal distribution. If the Markovian assumption is not satisfied, $nh^{1/2}S_n$ tends to infinity in probability and the asymptotic power of our test is all 1 in an asymptotic sense.

3.2.4 Power Study under Contiguous Alternatives

Last section, we have derived that our test is consistent against all fixed alternatives. Now, we want to know how the behavior of our test under the local alternatives. To compute the power function, classical tests usually consider the local misspecifications converging to the null at some rate. We consider a sequence of contiguous alternatives:

$$\mathbb{H}_{1n} : f(u|x,v) = (1-\delta_n)f(u|x) + \delta_n * g(v), \text{ for some } t \geq 1, \tag{3.4}$$

where $f(\cdot)$ is the density function, $g(\cdot)$ satisfies $g(\cdot) \geq 0$ and $\int g(y)dy = 1$ and a sequence $\delta_n (0 \leq \delta_n \leq 1)$ converges to 0 as $n \to \infty$. For the power performances of our test under local alternative hypothesis $\mathbb{H}_{1n}$, we obtain the following theorem.

Theorem 3.2.6 *Suppose that Conditions 1-5 hold. When the observed data $\{X_t\}_0^n$ are from a stationary process and the second moments of X_t exist, under $\mathbb{H}_{1n}$, if $\delta_n = O((n^{-1}h^{-1/2})^{1/2})$, we have that $nh^{1/2}S_n$ converges in distribution to a random variable that follows the normal distribution $N(C\mu_1, \sigma_{1A}^2)$, for some constant C, where μ_1 and σ_{1A}^2 are given in the Appendix. If $\delta_n = O((n^{-1}h^{-1/2})^a)$, $0 < a < 1/2$, then $nh^{1/2}S_n$ converges to ∞.*

From Theorem 3.2.6, we can obtain that: (1) when the local alternatives are different from the null hypothesis at the rate $n^{-1/2}h^{-1/4}$, the asymptotic powers of our test are 1 in the asymptotic sense; (2) when the alternatives converges to the null hypothesis at the rate $n^{-1/2}h^{-1/4}$, our test is also able to identify them.

3.3 Numerical Studies

In this section, we will examine the finite sample performance of out test statistics. By Theorem 3.2.5 and 3.2.6, we can get the asymptotic distributions of S_n under the different situations, but the variance of S_n seems to be difficult to be calculated and very complicated for the practical use. In order to determine the critical value or p-value, the local bootstrap procedure is considered, which was proposed by Paparoditis and Politis [2000]. The steps are listed as follows:

(a) For a given sample $\xi_n = (U_i, V_i, X_i)_{i=1}^N$, draw U_i^* from

$$\widehat{F}_{U|X=X_i} = \frac{\sum_{j=1}^N K_{ij} I_{(-\infty,U_j]}(u)}{\sum_{j=1}^N K_{ij}}$$

for $i = 1, \cdots, n$ to construct the local bootstrap sample $\xi_n^* = (U_i^*, V_i, X_i)_{i=1}^N$;

(b) Figure out the local bootstrap statistic S_n^* in the same way as S_n;

(c) Repeat steps (a) and (b) B times to get B local bootstrap statistics $\{S_{n_k}^*\}_{k=1}^B$;

(d) Compute the bootstrap p-value $p_b = B^{-1}\sum_{k=1}^B I(S_{n_k}^* > S_n)$; where $I(\cdot)$ is the indicator function.

From the statement of Chen and Hong [2012], when the method of Paparoditis and Politis [2000] is applied to the test procedure, we can similarly show that our test statistic S_n^* almost surely has the same distribution of S_n by using the analogous proof of Theorem 4.1 of Su and White [2008]. We demonstrate the finite sample performance of our test CDCTW with several simulation studies in the following part. Throughout all of the experiments in this subsection, we fix the sample size as $n = 100, 250$. For each experiment, we generate the $n + 100$ observations and delete the first 100 to avoid the influence of the initial values. To evaluate the size and power of our test, we generate 1000 random samples $\{X_t\}_{t\geq 0}$ and choose the number of bootstrap iteration as 200. And we apply the significant level of 0.05 and

0.1 to all models. For simplicity, the bandwidth h of the kernel function is taken as $h = 1.06 * S_X^2 n^{-9/2}$, where S_X^2 is the sample variance of the sample observations $\{X_t\}_{t\geq 0}$.

In order to assess the size of our test under $\mathbb{H}_0$, we consider the following three Markovian models.

Example 1. Autoregressive Model-AR(1) model: $X_t = 0.5X_{t-1} + \varepsilon_t$.

Example 2. Autoregressive Conditional Heteroskedasticity model-ARCH(1) model: $X_t = \sigma_t \varepsilon_t$ and $\sigma_t^2 = 0.1 + 0.1X_{t-1}^2$.

Example 3. The Ornstein-Uhlenbeck model: $dX_t = \kappa(\alpha - X_t)dt + \sigma dW_t$, where W_t is a Brownian motion and the parameters are chosen as $\kappa = 0.2$, $\alpha = 0.085$ and $\sigma = 0.08$, which follows the setup of Aït-Sahalia et al. [2010]. In Example 1-3, ε_t are independent and identically distributed (i.i.d.) and follow the standard normal distribution. The results are summarized in Table 3.1.

Table 3.1: The size of our test

		n=100			n=250		
	q	10	15	20	10	15	20
EX1	0.05	0.049	0.034	0.039	0.055	0.049	0.048
	0.1	0.103	0.070	0.081	0.099	0.102	0.109
EX2	0.05	0.049	0.061	0.057	0.055	0.066	0.061
	0.1	0.099	0.102	0.117	0.097	0.121	0.116
EX3	0.05	0.065	0.048	0.041	0.095	0.102	0.098
	0.1	0.118	0.107	0.090	0.048	0.057	0.056

From this table 3.1, we can find that out test can control the size well no matter what the sample size is. Furthermore, to evaluate the power of the test, we consider another three NON-Markov models, which are listed as follows:

Example 4. Moving Average model-MA(1): $X_t = 0.8\varepsilon_{t-1} + \varepsilon_t$.

Example 5. Generalized Autoregressive Conditional Heteroskedasticity model: GARCH(1,1) $X_t = h_t^{-1/2}\varepsilon_t$, $h_t = 0.1 + 0.2X_{t-1}^2 + 0.7h_{t-1}$.

Example 6. GARCH in Mean model: $X_t = 0.3 + 0.05h_t + z_t$, $z_t = h_t^{-1/2}\varepsilon_t$, and $h_t = 0.1 + 0.2X_{t-1}^2 + 0.7h_{t-1}$. And also, ε_t are i.i.d. and follow the normal distribution with mean zero and variance one.

From the results in Table 3.2, the power of our test is not good, but the power increases as the sample size becomes larger. And the choice of the lag q affects the power a little. Actually, the sample size N of ξ_i is affected by q, when n is fixed, if q

Table 3.2: The power of our test

		n=100			n=250		
	q	10	15	20	10	15	20
EX4	0.05	0.141	0.03	0.029	0.361	0.297	0.237
	0.1	0.261	0.069	0.065	0.530	0.443	0.362
EX5	0.05	0.056	0.039	0.047	0.093	0.087	0.097
	0.1	0.098	0.09	0.104	0.149	0.154	0.161
EX6	0.05	0.062	0.041	0.042	0.108	0.107	0.108
	0.1	0.119	0.085	0.099	0.196	0.190	0.190

increases, $N = n - q$ will become small, and then it will influence the power of our test. Therefore, we need to find a balance of them. This is an interesting work in the future.

3.4 Conclusion

This chapter investigates the test of Markovian assumption in one stationary β mixing process $\{X_t\}_{t\geq 0}$ with one dimensional weak dependent data. The Markov property has a wide of application in many areas such as time series, economics and finance. In the past few years, few papers have talked about this issue and the existing methods have more or less drawbacks. Since the Markov property is one kind of the conditional independence property, which means that the future information and the past information are uncorrelated with each other when the current information is fixed, we can utilize the means of testing conditional independence to detect the Markov property. The conditional distance covariance is a new concept of the conditional relationship between two random vectors give the third one. As a result, we take advantage of this concept to construct the test statistics and obtain the asymptotic null distribution and local alternative distribution. Furthermore, We will investigate the numerical performance of the proposed test statistic in some well-known models such as the Ornstein-Uhlenbeck model. Simulation studies show that our proposed test statistic can control well but its power is a little bad since it is influenced by the sample size and the lag order. How to get the balance between the lag order and the sample size will be studied in the future.

3.5 Appendix-Technical Proof of the Theorems

Lemma 3.5.1 *If $0 < \alpha < 2$, then for all $\boldsymbol{x}$ in $\mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \frac{1-\cos(\boldsymbol{t}'\boldsymbol{x})}{|\boldsymbol{t}|^{d+\alpha}} dt = C(d,\alpha)|\boldsymbol{x}|^{\alpha},$$

where

$$C(d,\alpha) = \frac{2\pi^{d/2}\Gamma(1-\frac{\alpha}{2})}{\alpha 2^{\alpha}\Gamma(\frac{d+\alpha}{2})}$$

and $\Gamma(\cdot)$ is the complete gamma function. The integrals at 0 and ∞ are meant in the principal value sense: $\lim_{\varepsilon\to 0}\int_{\mathbb{R}^d\setminus\{\varepsilon B+\varepsilon^{-1}B^c\}}$, where B is the unit ball (centered at 0) in $\mathbb{R}^d$ and B^c is the complement of B.

Proof of Theorem 3.2.1: As we know,

$$\begin{aligned}
& |\hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)|^2 \\
= \; & (\hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v))\overline{(\hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v))} \\
= \; & |\hat{\phi}_{U,V|X}(u,v)|^2 + |\hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)|^2 \\
- \; & \hat{\phi}_{U,V|X}(u,v) * \overline{\hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)} - \overline{\hat{\phi}_{U,V|X}(u,v)} * \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v) \\
\triangleq \; & T_1 + T_2 - T_3 - T_4
\end{aligned}$$

For T_1,

$$\begin{aligned}
T_1 &= |\hat{\phi}_{U,V|X}(u,v)|^2 = \hat{\phi}_{U,V|X}(u,v) * \overline{\hat{\phi}_{U,V|X}(u,v)} \\
&= \frac{\sum_{k=q}^{n-1}\exp(iuU_k + iv'V_k)\omega_k(x)}{\omega(x)} * \frac{\sum_{k=q}^{n-1}\exp(-iuU_k - iv'V_k)\omega_k(x)}{\omega(x)} \\
&= \frac{1}{\omega^2(x)}\sum_{k,l=q}^{n-1}\omega_k(x)\omega_l(x)\exp[iu(U_k-U_l) + iv'(V_k-V_l)] \\
&= \frac{1}{\omega^2(x)}\sum_{k,l=q}^{n-1}\omega_k(x)\omega_l(x)\cos[u(U_k-U_l)]\cos[v'(V_k-V_l)] + R_1
\end{aligned}$$

For T_2,

$$\begin{aligned}
T_2 &= |\hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)|^2 = \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v) * \overline{\hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)} \\
&= \hat{\phi}_{U|X}(u) * \overline{\hat{\phi}_{U|X}(u)} * \hat{\phi}_{V|X}(v) * \overline{\hat{\phi}_{V|X}(v)} \\
&= \frac{\sum_{k=q}^{n-1} \exp(iuU_k)\omega_k(x)}{\omega(x)} * \frac{\sum_{k=q}^{n-1} \exp(-iuU_k)\omega_k(x)}{\omega(x)} \\
&* \frac{\sum_{k=q}^{n-1} \exp(iv'V_k)\omega_k(x)}{\omega(x)} * \frac{\sum_{k=q}^{n} \exp(-iv'V_k)\omega_k(x)}{\omega(x)} \\
&= \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) \exp[iu(U_k - U_l)] \\
&* \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) \exp[iv'(V_k - V_l)] \\
&= \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) \cos[u(U_k - U_l)] \\
&* \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x) \cos[v'(V_k - V_l)] + R_2
\end{aligned}$$

For T_3,

$$\begin{aligned}
T_3 &= \hat{\phi}_{U,V|X}(u,v) * \overline{\hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)} \\
&= \frac{\sum_{k=q}^{n-1} \exp(iuU_k + iv'V_k)\omega_k(x)}{\omega(x)} * \frac{\sum_{k=q}^{n-1} \exp(-iuU_k)\omega_k(x)}{\omega(x)} * \frac{\sum_{k=q}^{n-1} \exp(-iv'V_k)\omega_k(x)}{\omega(x)} \\
&= \frac{1}{\omega^3(x)} \sum_{m=q}^{n-1}\sum_{k=q}^{n-1}\sum_{l=q}^{n-1} \omega_m(x)\omega_k(x)\omega_l(x) \exp[iu(U_m - U_k)] \exp[iv'(V_m - V_l)] \\
&= \frac{1}{\omega^3(x)} \sum_{m=q}^{n-1}\sum_{k=q}^{n-1}\sum_{l=q}^{n-1} \omega_m(x)\omega_k(x)\omega_l(x) \cos[u(U_m - U_k)] \cos[v'(V_m - V_l)] + R_3
\end{aligned}$$

Similarly,

$$T_4 = \frac{1}{\omega^3(x)} \sum_{m=q}^{n-1}\sum_{k=q}^{n-1}\sum_{l=q}^{n-1} * \omega_m(x)\omega_k(x)\omega_l(x) \cos[u(U_m - U_k)] \cos[v'(V_m - V_l)] + R_4$$

where, R_1, R_2, R_3 and R_4 represent terms that vanish when we calculate the integral $\|\hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)\|^2$.

To evaluate the integral $\|\hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)\|^2$, we apply Lemma 3.5.1 and

$$\cos u \cos v = 1 - (1 - \cos u) - (1 - \cos v) + (1 - \cos u)(1 - \cos v).$$

Hence, we can have

$$\begin{aligned}
T_1 &= \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[u(U_k - U_l)])(1 - \cos[v'(V_k - V_l)]) \\
&+ 1 - \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[u(U_k - U_l)]) \\
&- \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[v'(V_k - V_l)]) + R_1
\end{aligned}$$

$$\begin{aligned}
T_2 &= \frac{1}{\omega^4(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[u(U_k - U_l)]) \\
&\times \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[v'(V_k - V_l)]) \\
&+ 1 - \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[u(U_k - U_l)]) \\
&- \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1 - \cos[v'(V_k - V_l)]) + R_2
\end{aligned}$$

$$\begin{aligned}
T_3 &= \frac{1}{\omega^3(x)} \sum_{m=q}^{n-1} \sum_{k=q}^{n-1} \sum_{l=q}^{n-1} \omega_m(x)\omega_k(x)\omega_l(x) \\
&\times (1 - \cos[u(U_m - U_k)])(1 - \cos[v'(V_m - V_l)]) \\
&+ 1 - \frac{1}{\omega^2(x)} \sum_{m,k=q}^{n-1} \omega_m(x)\omega_k(x)(1 - \cos[u(U_m - U_k)]) \\
&- \frac{1}{\omega^2(x)} \sum_{m,l=q}^{n-1} \omega_m(x)\omega_l(x)(1 - \cos[v'(V_m - V_l)]) + R_3
\end{aligned}$$

$$
\begin{aligned}
T_4 &= \frac{1}{\omega^3(x)} \sum_{m=q}^{n-1} \sum_{k=q}^{n-1} \sum_{l=q}^{n-1} \omega_m(x)\omega_k(x)\omega_l(x) \\
&\times (1-\cos[u(U_m - U_k)])(1-\cos[v'(V_m - V_l)]) \\
&+ 1 - \frac{1}{\omega^2(x)} \sum_{m,k=q}^{n-1} \omega_m(x)\omega_k(x)(1-\cos[u(U_m - U_k)]) \\
&- \frac{1}{\omega^2(x)} \sum_{m,l=q}^{n-1} \omega_m(x)\omega_l(x)(1-\cos[v'(V_m - V_l)]) + R_4
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
& |\hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v)|^2 \\
= \;& T_1 + T_2 - T_3 - T_4 \\
= \;& \frac{1}{\omega^2(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1-\cos[u(U_k - U_l)])(1-\cos[v'(V_k - V_l)]) \\
+ \;& \frac{1}{\omega^4(x)} \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1-\cos[u(U_k - U_l)]) \\
& \sum_{k,l=q}^{n-1} \omega_k(x)\omega_l(x)(1-\cos[v'(V_k - V_l)]) \\
- \;& 2\frac{1}{\omega^3(x)} \sum_{m=q}^{n-1} \sum_{k=q}^{n-1} \sum_{l=q}^{n-1} \omega_m(x)\omega_k(x)\omega_l(x) \\
& (1-\cos[u(U_m - U_k)])(1-\cos[v'(V_m - V_l)]) \\
+ \;& R_1 + R_2 - R_3 - R_4.
\end{aligned}
$$

And then, by Lemma 3.5.1 we only need to evaluate the integrals of the type

$$
\begin{aligned}
& \int_{\boldsymbol{R}^{d_u+d_v}} \frac{(1-\cos u(U_k - U_l))(1-\cos v'(V_k - V_l))}{|u|^{d_u+1}|v|^{d_v+1}} du dv \\
= \;& \int_{\boldsymbol{R}^{d_u}} \frac{(1-\cos u(U_k - U_l))}{|u|^{d_u+1}} du * \int_{\boldsymbol{R}^{d_v}} \frac{(1-\cos v'(V_k - V_l))}{|v|^{d_v+1}} dv \\
= \;& C_{d_u}|U_k - U_l| * C_{d_v}|V_k - V_l| \\
= \;& C_{d_u} d_{kl}^U * C_{d_v} d_{kl}^V
\end{aligned}
$$

Finally, we can get

$$\mathcal{D}_n^2(U,V|X) = D_1 + D_2 - 2D_3.$$

Proof of Theorem 3.2.2: Denote that $u_k = \exp\{iuU_k\} - \phi_{U|X}(u)$ and $v_k = \exp\{iv'V_k\} - \phi_{V|X}(v)$, and then

$$\begin{aligned}\xi_n(u,v) &= \hat{\phi}_{U,V|X}(u,v) - \hat{\phi}_{U|X}(u)\hat{\phi}_{V|X}(v) \\ &= \frac{\sum_{k=q}^{n-1} u_k v_k \omega_k(x)}{\omega(x)} - \frac{\sum_{k=q}^{n-1} u_k \omega_k(x)}{\omega(x)} \frac{\sum_{k=q}^{n-1} v_k \omega_k(x)}{\omega(x)}\end{aligned}$$

For each $\delta > 0$, define the region

$$O(\delta) = \{(u,v) : \delta \le |u| \le 1/\delta,\ \delta \le |v| \le 1/\delta\}$$

and random variables

$$\mathcal{D}_{n,\delta}^2(U,V|X) = \int_{O(\delta)} \frac{1}{c_{d_u} c_{d_v}} \frac{|\xi_n(u,v)|^2}{|u|^{d_u+1}|v|^{d_v+1}} dudv$$

For any fixed $0 < \delta < 1$, the function $\frac{1}{c_{d_u} c_{d_v} |u|^{d_u+1}|v|^{d_v+1}}$ is bounded on $O(\delta)$. By the strong law of large numbers(SLLN), it follows that almost surely

$$\begin{aligned}\lim_{n\to\infty} \mathcal{D}_{n,\delta}^2(U,V|X) &= \mathcal{D}_{\cdot,\delta}^2(U,V|X) \\ &= \int_{O(\delta)} \frac{1}{c_{d_u} c_{d_v}} \frac{|\phi_{U,V|X}(u,v) - \phi_{U|X}(u)\phi_{V|X}(v)|^2}{|u|^{d_u+1}|v|^{d_v+1}} dudv.\end{aligned}$$

Clearly, $\lim_{\delta\to 0} \mathcal{D}_{\cdot,\delta}^2(U,V|X) = \mathcal{D}^2(U,V|X)$. Now it remains to prove that almost surely,

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} |\mathcal{D}_{n,\delta}^2(U,V|X) - \mathcal{D}_n^2(U,V|X)| = 0.$$

Given $\delta > 0$,

$$\begin{aligned}|\mathcal{D}_{n,\delta}^2 - \mathcal{D}_n^2| &\le \Big(\int_{|u|<\delta} + \int_{|u|>1/\delta} + \int_{|v|<\delta} + \int_{|v|>1/\delta}\Big) \frac{1}{c_{d_u} c_{d_v}} \frac{|\xi_n(u,v)|^2}{|u|^{d_u+1}|v|^{d_v+1}} dudv. \\ &\triangleq I_1 + I_2 + I_3 + I_4\end{aligned}$$

For $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$ in $\boldsymbol{R}^p$ and $|\boldsymbol{z}| = 1$, define the function

$$G(y) = \int_{|\boldsymbol{z}|<y} \frac{1 - \cos t'\boldsymbol{z}}{|\boldsymbol{z}|^{1+p}} d\boldsymbol{z}.$$

According to Lemma 3.5.1, $G(y)$ is bounded by c_p and $\lim_{y\to\infty} G(y) = 0$. Applying the inequality $|x+y|^2 < 2|x|^2 + 2|y|^2$ and the Cauchy-Schwarz inequality for sums, we can obtain that

$$\begin{aligned}
|\xi_n(u,v)|^2 &\leq 2\left|\frac{\sum_{k=q}^{n-1} u_k v_k \omega_k(x)}{\omega(x)}\right|^2 + 2\left|\frac{\sum_{k=q}^{n-1} u_k \omega_k(x)}{\omega(x)}\right|^2 \left|\frac{\sum_{k=q}^{n-1} v_k \omega_k(x)}{\omega(x)}\right|^2 \\
&= 2\frac{|\sum_{k=q}^{n-1} u_k \sqrt{\omega_k(x)} v_k \sqrt{\omega_k(x)}|}{(\omega(x))^2} \\
&+ 2\frac{|\sum_{k=q}^{n-1} u_k \sqrt{\omega_k(x)} * \sqrt{\omega_k(x)}|}{(\omega(x))^2} \frac{|\sum_{k=q}^{n-1} v_k \sqrt{\omega_k(x)} * \sqrt{\omega_k(x)}|}{(\omega(x))^2} \\
&\leq 2\frac{\sum_{k=q}^{n-1} |u_k|^2 \omega_k(x) \sum_{k=q}^{n-1} |v_k|^2 \omega_k(x)}{(\omega(x))^2} \\
&+ 2\frac{\sum_{k=q}^{n-1} |u_k|^2 \omega_k(x) * \sum_{k=q}^{n-1} \omega_k(x)}{(\omega(x))^2} \frac{\sum_{k=q}^{n-1} |v_k|^2 \omega_k(x) * \sum_{k=q}^{n-1} \omega_k(x)}{(\omega(x))^2} \\
&= 4\frac{\sum_{k=q}^{n-1} |u_k|^2 \omega_k(x) \sum_{k=q}^{n-1} |v_k|^2 \omega_k(x)}{(\omega(x))^2}.
\end{aligned}$$

Thus,

$$I_1 \leq 4\frac{\sum_{k=q}^{n-1} \int_{|u|<\delta} \frac{|u_k|^2}{c_{d_u}|u|^{d_u+1}} du \omega_k(x)}{\omega(x)} \frac{\sum_{k=q}^{n-1} \int_{\boldsymbol{R}^{d_v}} \frac{|v_k|^2}{c_{d_v}|v|^{d_v+1}} dv \omega_k(x)}{\omega(x)}.$$

Here,

$$\begin{aligned}
|v_k|^2 &= |\exp\{iv'V_k\} - \phi_{V|X}(v)|^2 \\
&= 1 - \exp\{iv'V_k\}\overline{\phi_{V|X}(v)} - \exp\{-iv'V_k\}\phi_{V|X}(v) + |\phi_{V|X}(v)|^2 \\
&= 1 - \exp\{iv'V_k\}E(\exp\{-iv'V\}|X) - \exp\{-iv'V_k\}E(\exp\{iv'V\}|X) \\
&\quad +|\phi_{V|X}(v)|^2 \\
&= 1 - E_V(\exp\{iv'(V_k - V)\}|X) - E_V(\exp\{iv'(V - V_k)\}|X) + |\phi_{V|X}(v)|^2 \\
&= E_V(1 - \exp\{iv'(V_k - V)\}|X) + E_V(1 - \exp\{iv'(V - V_k)\}|X) \\
&\quad -(1 - |\phi_{V|X}(v)|^2) \\
&= E_V(1 - \exp\{iv'(V_k - V)\}|X) + E_V(1 - \exp\{iv'(V - V_k)\}|X) \\
&- E(\exp\{iv'(V - V')\}|X)
\end{aligned}$$

where the expectation E_V is taken with respect to V, and we take $V' \stackrel{D}{=} V$ such that they are independent. Hence,

$$\int_{\boldsymbol{R}^{d_v}} \frac{|v_k|^2}{c_{d_v}|v|^{d_v+1}}dv = 2E_V(|V_k - V||X) - E(|V - V'||X) \leq 2[|V_k| + E_V(|V||X)]$$

Further, take a suitable change of variables: $u = \frac{\boldsymbol{z}}{|U_k - U|}$,

$$\begin{aligned}
&\int_{|u|<\delta} \frac{1 - \cos u'(U_k - U)}{c_{d_u}|u|^{d_u+1}}du \\
= &\int_{|z|<|U_k-U|\delta} \frac{1 - \cos(z'\frac{U_k-U}{|U_k-U|})}{c_{d_z}|z|^{d_z+1}}|U_k - U|^{d_z+1} * \frac{dz}{|U_k - U|^{d_z}} \\
= &|U_k - U| * \int_{|z|<|U_k-U|\delta} \frac{1 - \cos(z'\frac{U_k-U}{|U_k-U|})}{c_{d_z}|z|^{d_z+1}}dz \\
= &|U_k - U| * G(|U_k - U|\delta)
\end{aligned}$$

Thus, we can get

$$\begin{aligned}
\int_{|u|<\delta} \frac{|u_k|^2}{c_{d_u}|u|^{d_u+1}}du &= 2E_U(|U_k - U| * G(|U_k - U|\delta)|X) \\
&\quad -E(|U - U'| * G(|U - U'|\delta)|X) \\
&\leq 2E_U(|U_k - U| * G(|U_k - U|\delta)|X)
\end{aligned}$$

Therefore,

$$\begin{aligned} I_1 &\leq 4\frac{\sum_{k=q}^{n-1}\int_{|u|<\delta}\frac{|u_k|^2}{c_{d_u}|u|^{d_u+1}}du\omega_k(x)}{\omega(x)}\frac{\sum_{k=q}^{n-1}\int_{\boldsymbol{R}^{d_v}}\frac{|v_k|^2}{c_{d_v}|v|^{d_v+1}}dv\omega_k(x)}{\omega(x)} \\ &\leq 4\frac{\sum_{k=q}^{n-1}2E_U(|U_k-U|*G(|U_k-U|\delta)|X)\omega_k(x)}{\omega(x)} \\ &\quad\times\frac{\sum_{k=q}^{n-1}2(|V_k|+E_V(|V||X))\omega_k(x)}{\omega(x)} \\ &= 16\frac{\sum_{k=q}^{n-1}E_U(|U_k-U|*G(|U_k-U|\delta)|X)\omega_k(x)}{\omega(x)} \\ &\quad\times\left(\frac{\sum_{k=q}^{n-1}|V_k|\omega_k(x)}{\omega(x)}+E_V(|V||X)\right). \end{aligned}$$

Since $\omega(x)/(n-q-1)$ is a consistent density function estimator of random variable X,

$$\limsup_{n\to\infty} I_1 \leq 16E_U(|U_k-U|G(|U_k-U|\delta)|X)*2E(|V||X)$$

almost surely. By the Lebesgue bounded convergence theorem for integrals and expectations,

$$\limsup_{\delta\to 0}\limsup_{n\to\infty} I_1 = 0$$

almost surely.

Next, consider the second term I_2. Since $|u_k|^2 \leq 4$,

$$\frac{1}{\omega(x)}\sum_{k=q}^{n-1}\omega_k(x)\int_{|u|>1/\delta}\frac{|u_k|^2}{c_{d_u}|u|^{d_u+1}}du \leq 4\int_{|u|>1/\delta}\frac{1}{c_{d_u}|u|^{d_u+1}}du$$

and

$$\frac{1}{\omega(x)}\sum_{k=q}^{n-1}\omega_k(x)\int_{\boldsymbol{R}^{d_v}}\frac{|v_k|^2}{c_{d_v}|v|^{d_v+1}}dv \leq 2\frac{1}{\omega(x)}\sum_{k=q}^{n-1}\omega_k(x)[|V_k|+E_V(|V||X)]$$

Therefore,

$$\begin{aligned} I_2 &\leq 4\frac{\sum_{k=q}^{n-1}\int_{|u|>1/\delta}\frac{|u_k|^2}{c_{d_u}|u|^{d_u+1}}du\omega_k(x)}{\omega(x)}\frac{\sum_{k=q}^{n-1}\int_{\boldsymbol{R}^{d_v}}\frac{|v_k|^2}{c_{d_v}|v|^{d_v+1}}dv\omega_k(x)}{\omega(x)} \\ &\leq 16\int_{|u|>1/\delta}\frac{1}{c_{d_u}|u|^{d_u+1}}du * 2\frac{1}{\omega(x)}\sum_{k=q}^{n-1}\omega_k(x)[|V_k| + E_V(|V||X)] \\ &= 16\delta * 2\frac{1}{\omega(x)}\sum_{k=q}^{n-1}\omega_k(x)[|V_k| + E_V(|V||X)] \end{aligned}$$

And then, we have almost surely $\limsup_{\delta\to 0}\limsup_{n\to\infty} I_2 = 0$.
The remaining two terms I_3 and I_4 can be dealt with similar method of the first two summands. Finally, we get

$$\lim_{n\to\infty}\mathcal{D}_n^2(U,V|X) = \mathcal{D}^2(U,V|X).$$

Proof of Theorem 3.2.3: Since $d^s_{ijkl} = d_{ijkl} + d_{ijlk} + d_{ilkj}$, we have

$$\begin{aligned} &\sum_{i,j,k,l}\omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X)d^s_{ijkl} \\ =\ & 3\sum_{i,j,k,l}\omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X)d_{ijkl} \\ =\ & 3[\sum_{i,j,k,l}(d^U_{ij}d^V_{ij} + d^U_{kl}d^V_{kl} + d^U_{ik}d^V_{ik} + d^U_{jl}d^V_{jl})\omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X) \\ &+\sum_{i,j,k,l}(d^U_{ij}d^V_{kl} + d^U_{kl}d^V_{ij} + d^U_{ik}d^V_{jl} + d^U_{jl}d^V_{ik})\omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X) \\ &-\sum_{i,j,k,l}(d^U_{ij}d^V_{ik} + d^U_{ij}d^V_{jl} + d^U_{kl}d^V_{ik} + d^U_{kl}d^V_{jl} \\ &+(d^U_{ik}d^V_{ij} + d^U_{ik}d^V_{kl} + d^U_{jl}d^V_{ij} + d^U_{jl}d^V_{kl})\omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X)] \\ \triangleq\ & K_1 + K_2 - K_3 \end{aligned}$$

Using the symmetry of distance $d^U_{ij} = d^U_{ji}$ and $d^V_{ij} = d^V_{ji}$, we ca easily get the following equations:

$$K_1 = 12\omega^2(X)\sum_{i,j}d^U_{ij}d^V_{ij}\omega_i(X)\omega_j(X),$$

$$K_2 = 12 \sum_{i,j} d_{ij}^U \omega_i(X)\omega_j(X) \sum_{k,l} d_{kl}^V \omega_k(X)\omega_l(X),$$

$$K_3 = 24\omega(X) \sum_i \sum_{k,l} d_{ij}^U d_{ik}^V \omega_i(X)\omega_j(X)\omega_k(X).$$

Hence, based on the proof of Theorem 3.2.1,

$$\begin{aligned}
&\sum_{i,j,k,l} \omega_i(X)\omega_j(X)\omega_k(X)\omega_l(X) d_{ijkl}^s \\
= \;& 12\omega^2(X) \sum_{i,j} d_{ij}^U d_{ij}^V \omega_i(X)\omega_j(X) + 12 \sum_{i,j} d_{ij}^U \omega_i(X)\omega_j(X) \sum_{k,l} d_{kl}^V \omega_k(X)\omega_l(X) \\
& -24\omega(X) \sum_i \sum_{k,l} d_{ij}^U d_{ik}^V \omega_i(X)\omega_j(X)\omega_k(X) \\
= \;& 12\omega^4(X)(D_1 + D_2 - 2D_3) \\
= \;& 12\omega^4(X)\mathcal{D}_n^2(U, V|X),
\end{aligned}$$

which implies that

$$\mathcal{D}_n^2(U, V|X) = \frac{1}{n^4} \sum_{i,j,k,l} \psi_n(\xi_i, \xi_j, \xi_k, \xi_l; X).$$

In order to prove Theorem 3.2.4, we need the following two theorems in the book Lee [1990].

Lemma 3.5.2 (Lee [1990]) *Let $t_1 < t_2 < \cdots < t_k$ be integers, let F, G_j and H_j be the distribution functions of $(X_{t_1}, \ldots, X_{t_k})$, $(X_{t_1}, \ldots, X_{t_j})$ and $(X_{t_{j+1}}, \ldots, X_{t_k})$ respectively and let μ be the signed measure corresponding to the function*

$$F(x_1, \ldots, x_k) - G_j(x_1, \ldots, x_j)H_j(x_{j+1}, \ldots, x_k)$$

which is bounded variation. Then $|\mu| = \beta(t_{j+1} - t_j)$.

Lemma 3.5.3 (Lee [1990]) *Let $t_1 < t_2 < \cdots < t_k$, F, G_j and H_j be as in Lemma 3.5.2. Let h be a measurable function such that*

$$M = \max\left(\int |h|^{1+\delta} dF, \int\int |h|^{1+\delta} dG_j dH_j \right)$$

is finite for some $\delta > 0$. Then

$$\left| \int h dF - \int \int h dG_j dH_j \right| \leq 3M^{\frac{1}{1+\delta}} \beta^{\frac{\delta}{1+\delta}} (t_{j+1} - t_j).$$

Proof of Theorem 3.2.4: Let $P_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5) = d^s_{1234} K_{15} K_{25} K_{35} K_{45}$ and express S_N as a U-statistic with random kernel,

$$S_N = \frac{1}{C_N^5 h^4} \sum_{i<j<k<l<m} \Psi_N(\xi_i, \xi_j, \xi_k, \xi_l, \xi_m)$$

where

$$\begin{aligned} & \Psi_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5) \\ = \; & \frac{1}{5}[P_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5) + P_N(\xi_5, \xi_2, \xi_3, \xi_4, \xi_1) \\ & + P_N(\xi_1, \xi_5, \xi_3, \xi_4, \xi_2) + P_N(\xi_1, \xi_2, \xi_5, \xi_4, \xi_3) \\ & + P_N(\xi_1, \xi_2, \xi_3, \xi_5, \xi_4) \end{aligned}$$

By Considering the H-decomposition in Lee(1990), we denote that for $c = 1, 2, 3, 4, 5$,

$$P_{Nc}(w_1, \cdots, w_c) = E(P_N(\xi_1, \cdots, \xi_5) | (\xi_1, \cdots, \xi_c) = (w_1, \cdots, w_c)),$$

$$\Psi_{Nc}(w_1, \cdots, w_c) = E(\Psi_N(\xi_1, \cdots, \xi_5) | (\xi_1, \cdots, \xi_c) = (w_1, \cdots, w_c)).$$

Further, let $h^{(1)}(w_1) = \Psi_{N1}(w_1)/h^4 - \theta$ and

$$h^{(c)}(w_1, w_2, \cdots, w_c) = \Psi_{Nc}(w_1, \cdots, w_c)/h^4 - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(w_{i_1}, \cdots, w_{i_j}) - \theta.$$

And then,

$$S_N = \theta + \sum_{j=1}^{5} \binom{5}{j} H_N^{(j)},$$

where $H_N^{(j)} = \binom{N}{j}^{-1} \sum_{(N,j)} h^{(j)}(w_{i_1}, \cdots, w_{i_j})$ and $\theta = \frac{1}{h^4} \int \Psi_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5) \prod_{i=1}^{5} dF(\xi_i)$.

Step1: $S_N = \frac{1}{C_N^5 h^4} \sum_{i<j<k<l<m} E[\Psi_N(\xi_i, \xi_j, \xi_k, \xi_l, \xi_m)] + o_p(1)$.

Based on Markov's law of large numbers, we only need to prove that $Var(S_N) \to 0$. From the H-decomposition of S_N,

$$Var(S_N) = \sum_{j=1}^{5} \left(\binom{5}{j} \right)^2 Var(H_N^{(j)}) + 2 \sum_{1 \le k \ne l \le 5} Cov\left(\binom{5}{k} H_N^{(k)}, \binom{5}{l} H_N^{(l)} \right)$$

Firstly, we consider the variance terms. For the term $Var(H_N^{(1)})$, using the stationary property, we obtain

$$\begin{aligned} Var(H_N^{(1)}) &= Var(\frac{1}{N} \sum_{i=q}^{n-1} h^{(1)}(\xi_i)) \\ &= \frac{1}{N^2} \sum_{i=q}^{n-1} \sum_{j=q}^{n-1} Cov(h^{(1)}(\xi_i), h^{(1)}(\xi_j)) \\ &= \frac{1}{N^2} [N Var(h^{(1)}(\xi_q)) + 2 \sum_{j=1}^{N-1} (N-j) Cov(h^{(1)}(\xi_q), h^{(1)}(\xi_{q+j}))] \end{aligned}$$

and

$$\begin{aligned} Var(h^{(1)}(\xi_q)) &= \frac{1}{h^8} Var(\Psi_{N1}(\xi_q)) \\ &= \frac{1}{h^8} E[\Psi_{N1}(\xi_q)]^2 \\ &= \frac{1}{h^8} E[E(\Psi_N(\xi_1, \cdots, \xi_5) | \xi_1 = \xi_q)]^2 \end{aligned}$$

Note that $E[\Psi_{N1}(\xi_q)]^2$ can be expanded into several terms, and each of these terms can be shown to be of order h^8. The proof of the first term is listed as follows:

$$\begin{aligned} & E[P_{N1}(\xi_q)]^2 \\ = & \int [E(P_N(\xi_1, \cdots, \xi_5) | \xi_1 = \xi_q)]^2 * f(\xi_q) d\xi_q \\ = & \int [\int d^s_{q234} K_{q2} K_{q3} K_{q4} K_{q5} f(\xi_2, \xi_3, \xi_4, \xi_5) d\xi_2 d\xi_3 d\xi_4 d\xi_5]^2 f(\xi_q) d\xi_q \\ = & \int [\int d^s_{q234} K_{q2} K_{q3} K_{q4} K_{q5} f(u_2, v_2, x_2, u_3, v_3, x_3, u_4, v_4, x_4, u_5, v_5, x_5) \\ & du_2 dv_2 dx_2 du_3 dv_3 dx_3 du_4 dv_4 dx_4 du_5 dv_5 dx_5]^2 f(u_q, v_q, x_q) du_q dv_q dx_q \end{aligned}$$

Denote that $x_{5q} = \frac{x_5 - x_q}{h}$, $x_{52} = \frac{x_5 - x_2}{h}$, $x_{53} = \frac{x_5 - x_3}{h}$ and $x_{54} = \frac{x_5 - x_4}{h}$ and using the

Taylor Expansion,

$$\begin{aligned} & E[P_{N1}(\xi_q)]^2 \\ = \; & h^8 \int [\int d^s_{q234} K(x_{5q})K(x_{52})K(x_{53})K(x_{54}) f(u_2, v_2, x_q, u_3, v_3, x_q, u_4, v_4, x_q, \\ & u_5, v_5, x_q) du_2 dv_2 du_3 dv_3 du_4 dv_4 du_5 dv_5 dx_{5q} dx_{52} dx_{53} dx_{54}]^2 f(\xi_q) d\xi_q + o(h^8) \\ = \; & O(h^8) \end{aligned}$$

Therefore, $Var(h^{(1)}(\xi_q)) = O(1)$. And then, by Lemma 3.5.3, we can replace δ by $\delta/2$ and $\delta > \delta'$. Let $\lambda = \frac{\delta(2+\delta')}{\delta'(2+\delta)} > 1$,

$$\begin{aligned} \sum_{j=1}^{N-1}(N-j)Cov(h^{(1)}(\xi_q), h^{(1)}(\xi_{q+j})) &= \sum_{j=1}^{N-1}(N-j)E(h^{(1)}(\xi_q)h^{(1)}(\xi_{q+j})) \\ &\leq \sum_{j=1}^{N-1}(N-j)C\beta^{\delta/2+\delta}(j) \\ &\leq C\sum_{j=1}^{N-1}(N-j)j^{-\frac{\delta(2+\delta')}{\delta'(2+\delta)}} \\ &= O(N^{2-\lambda}). \end{aligned}$$

Therefore, $Var(H_N^{(1)}) = O(\frac{1}{N}) + O(\frac{1}{N^\lambda}) = O(\frac{1}{N})$.

Similarly, for other terms $Var(H_N^{(j)})$, $2 \leq j \leq 5$, we obtain $\lim_{N\to\infty} Var(H_N^{(j)}) = 0$. For the convariance terms, $i \neq j$,

$$|Cov(H_N^{(i)}, H_N^{(j)})| \leq (Var(H_N^{(i)}), Var(H_N^{(j)}))^{\frac{1}{2}} \to 0.$$

Thus, by Markov's law of large numbers,

$$S_N = \frac{1}{C_N^5 h^4} \sum_{i<j<k<l<m} E[\Psi_N(\xi_i, \xi_j, \xi_k, \xi_l, \xi_m)] + o_p(1)$$

when $nh \to 0$.

Step2: $E(S_N) = E[\mathcal{D}^2(U, V|X)a(X)] + O(h^2)$.

Actually, we only need to prove that $|E(S_N) - E[\mathcal{D}^2(U, V|X)a(X)]| \to 0$. Here, $a(X) = 12f^4(X)$. By using Lemma 3.5.3 repeatedly and replacing δ by $\delta/2$, we have

for different $i<j<k<l<m$,

$$\begin{aligned}
&\left|E[\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)]-\int\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)dF(\xi_i)dF(\xi_j)dF(\xi_k)dF(\xi_l)dF(\xi_m)\right|\\
\leq\ & 3M_1^{\frac{2}{2+\delta}}\beta^{\frac{\delta}{2+\delta}}(j-i)+3M_2^{\frac{2}{2+\delta}}\beta^{\frac{\delta}{2+\delta}}(k-j)\\
+\ & 3M_3^{\frac{2}{2+\delta}}\beta^{\frac{\delta}{2+\delta}}(l-k)+3M_4^{\frac{2}{2+\delta}}\beta^{\frac{\delta}{2+\delta}}(m-l)
\end{aligned}$$

Let $M=\max\{3M_1^{\frac{2}{2+\delta}},3M_2^{\frac{2}{2+\delta}},3M_3^{\frac{2}{2+\delta}},3M_4^{\frac{2}{2+\delta}}\}$, and then,

$$\begin{aligned}
&\left|\frac{1}{C_N^5h^4}\sum_{i<j<k<l<m}E[\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)]-\frac{1}{h^4}\int\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)\prod_{t=i,j,k,l,m}dF(\xi_t)\right|\\
=\ & \frac{1}{h^4}\left|\frac{1}{C_N^5}\sum_{i<j<k<l<m}[E[\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)]-\int\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)\prod_{t=i,j,k,l,m}dF(\xi_t)]\right|\\
\leq\ & \frac{1}{h^4C_N^5}\sum_{i<j<k<l<m}\left|E[\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)]-\int\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)\prod_{t=i,j,k,l,m}dF(\xi_t)\right|\\
\leq\ & \frac{1}{h^4C_N^5}\sum_{t=1}^{N-4}M\beta^{\frac{\delta}{2+\delta}}(t)\leq\frac{M}{h^4C_N^5}\sum_{t=1}^{N-4}t^{-\frac{\delta(2+\delta')}{\delta'(2+\delta)}}=\frac{M}{h^4C_N^5}\sum_{t=1}^{N-4}t^{-\lambda}\\
=\ & O(\frac{1}{h^4N^{4+\lambda}})=o(\frac{1}{h^4N^4})
\end{aligned}$$

That is to say,

$$\begin{aligned}
&\frac{1}{C_N^5h^4}\sum_{i<j<k<l<m}E[\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)]\\
=\ & \frac{1}{h^4}\int\Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)\prod_{i=1}^{5}dF(\xi_i)+o(\frac{1}{h^4N^4})
\end{aligned}$$

Based on the proof of Theorem 6 introduced by Wang et al. [2015], we obtain that

$$\frac{1}{h^4}\int\Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)\prod_{i=1}^{5}dF(\xi_i)=E[\mathcal{D}^2(U,V|X)*12f^4(X)]+O(h^2)$$

Finally, we can get

$$\frac{1}{C_N^5h^4}\sum_{i<j<k<l<m}E[\Psi_N(\xi_i,\xi_j,\xi_k,\xi_l,\xi_m)]=E[\mathcal{D}^2(U,V|X)*12f^4(X)]+o(\frac{1}{h^4N^4})+O(h^2).$$

Combining the results in Step 1 and Step 2, we can eventually get that

$$S_N \xrightarrow{P} S_a.$$

Proof of Theorem 3.2.5:(i) Firstly, we want to show that

$$E[\Psi_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)|\xi_1] = 0, \text{ almost surely.}$$

By the definition of $\Psi_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$, we need to prove each of five terms converge to zero almost surely. Now, we give the proof of the first term as follows:

$$\begin{aligned}
& E[P_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)|\xi_1] \\
= \; & E[d^s_{1234} K_{15} K_{25} K_{35} K_{45}|\xi_1] \\
= \; & 12E[(d^U_{12} d^V_{12} + d^U_{12} d^V_{34} - 2d^U_{12} d^V_{13}) K_{15} K_{25} K_{35} K_{45}|\xi_1] \\
= \; & 12E[E[(d^U_{12} d^V_{12} + d^U_{12} d^V_{34} - 2d^U_{12} d^V_{13})|X_1, X_2, X_3, X_4] K_{15} K_{25} K_{35} K_{45}|\xi_1]
\end{aligned}$$

By the conditionally independent property between U and V, we easily know that

$$E[P_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)|\xi_1] = 0.$$

Consequently, $E[\Psi_N(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)|\xi_1] = 0$, almost surely. Therefore, we can say that S_N is degenerated U-statistics with random kernel.

Next, since S_N is degenerated, $h^{(1)}(w_1) = 0$, almost surely, and then the H-decomposition of S_N becomes

$$S_N = \frac{\binom{5}{2}}{\binom{N}{2} h^4} \sum_{(N,2)} \Psi_{N2}(\xi_{i_1}, \xi_{i_2}) + R^{(2)}_N,$$

where $R^{(2)}_N = \sum_{j=3}^{5} \binom{5}{j} H^{(j)}_N$.

By Theorem 2 of Section 1.6 (Lee [1990]), we can know that

$$\int \Psi_{N2}(\xi_i, y) dF(\xi_i) = E\Psi_{N2}(\xi_i, y) = 0$$

and

$$\int \Psi_{N2}(x,\xi_j)dF(\xi_j) = E\Psi_{N2}(x,\xi_j) = 0,$$

for any fixed x and y. Noting that

$$E[\Psi_{N2}(\xi_i,\xi_j)|\xi_i] = E[\Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)|\xi_1 = \xi_i] = 0.$$

Furthermore, by the Markovian property, $E\{\Psi_{N2}(\xi_i,\xi_j)|\xi_1,\xi_2,\ldots,\xi_{j-1}\} = 0$, for any $i < j$. Apart from a constant factor, the first term of S_N is a degenerate U-statistic of second order. In the following proof, we will use Theorem A in Hjellvik *et al.* (1998) to obtain the asymptotical distribution of S_N. Now we check the conditions of Theorem A.

$$\begin{aligned} E[\Psi_{N2}(\xi_{i_1},\xi_{i_2})] &= E[E[\Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)|\xi_1 = \xi_{i_1},\xi_2 = \xi_{i_2}]] \\ &= E[E[\Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)|\xi_1]] \\ &= 0 \end{aligned}$$

Denote that $\sigma_N^2 = \sum_{q\le i<j\le n-1} Var(\Psi_{N2}(\xi_i,\xi_j)) = \sum_{q\le i<j\le n-1} E[\Psi_{N2}(\xi_i,\xi_j)]^2$.

$$\begin{aligned} \Psi_{N2}(\xi_i,\xi_j) &= E[\Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)|\xi_1 = \xi_i,\xi_2 = \xi_j] \\ &= \frac{1}{5}E[P_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5) + P_N(\xi_5,\xi_2,\xi_3,\xi_4,\xi_1) \\ &\quad + P_N(\xi_1,\xi_5,\xi_3,\xi_4,\xi_2) + P_N(\xi_1,\xi_2,\xi_5,\xi_4,\xi_3) \\ &\quad + P_N(\xi_1,\xi_2,\xi_3,\xi_5,\xi_4)|\xi_1 = \xi_i,\xi_2 = \xi_j] \end{aligned}$$

For each term of $\Psi_{N2}(\xi_i,\xi_j)$, using transformations and Taylor expansions, we can obtain

$$\begin{aligned} & E[P_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)|\xi_1 = \xi_i,\xi_2 = \xi_j] \\ = & \int d_{1234}^s K_{15}K_{25}K_{35}K_{45}f(\xi_3,\xi_4,\xi_5)\prod_{i=3}^{5} d\xi_i \\ = & h^3\int d_{1234}^s K(\frac{x_2+hx_{52}-x_1}{h})K(x_{52})K(x_{53})K(x_{54})f(u_3,v_3,x_2+h(x_{52}-x_{53}), \\ & u_4,v_4,x_2+h(x_{52}-x_{54}),u_5,v_5,x_2+hx_{52})du_3dv_3dx_{53}du_4dv_4dx_{54}du_5dv_5dx_{52} \end{aligned}$$

$$
\begin{aligned}
& E[P_N(\xi_5,\xi_2,\xi_3,\xi_4,\xi_1)|\xi_1=\xi_i,\xi_2=\xi_j] \\
= & \int d^s_{5234}K_{51}K_{21}K_{31}K_{41}f(\xi_3,\xi_4,\xi_5)\prod_{i=3}^{5}d\xi_i \\
= & h^3\int d^s_{5234}K(x_{51})K(\frac{x_2-x_1}{h})K(x_{31})K(x_{41})f(u_3,v_3,x_1+hx_{31}, \\
& u_4,v_4,x_1+hx_{41},u_5,v_5,x_1+hx_{51})du_3dv_3dx_{31}du_4dv_4dx_{41}du_5dv_5dx_{51}
\end{aligned}
$$

$$
\begin{aligned}
& E[P_N(\xi_1,\xi_5,\xi_3,\xi_4,\xi_2)|\xi_1=\xi_i,\xi_2=\xi_j] \\
= & \int d^s_{1534}K_{12}K_{52}K_{32}K_{42}f(\xi_3,\xi_4,\xi_5)\prod_{i=3}^{5}d\xi_i \\
= & h^3\int d^s_{1534}K(\frac{x_2-x_1}{h})K(x_{52})K(x_{32})K(x_{42})f(u_3,v_3,x_2+hx_{32}, \\
& u_4,v_4,x_2+hx_{42},u_5,v_5,x_2+hx_{52})du_3dv_3dx_{32}du_4dv_4dx_{42}du_5dv_5dx_{52}
\end{aligned}
$$

$$
\begin{aligned}
& E[P_N(\xi_1,\xi_2,\xi_5,\xi_4,\xi_3)|\xi_1=\xi_i,\xi_2=\xi_j] \\
= & \int d^s_{1254}K_{13}K_{23}K_{53}K_{43}f(\xi_3,\xi_4,\xi_5)\prod_{i=3}^{5}d\xi_i \\
= & h^3\int d^s_{1254}K(\frac{x_2+hx_{32}-x_1}{h})K(x_{32})K(x_{53})K(x_{43})f(u_3,v_3,x_2+hx_{32}, \\
& u_4,v_4,x_2+h(x_{32}+x_{43}),u_5,v_5,x_2+h(x_{32}+x_{53}))du_3dv_3dx_{32}du_4dv_4dx_{43}du_5dv_5dx_{53}
\end{aligned}
$$

$$
\begin{aligned}
& E[P_N(\xi_1,\xi_2,\xi_3,\xi_5,\xi_4)|\xi_1=\xi_i,\xi_2=\xi_j] \\
= & \int d^s_{1235}K_{14}K_{24}K_{34}K_{54}f(\xi_3,\xi_4,\xi_5)\prod_{i=3}^{5}d\xi_i \\
= & h^3\int d^s_{1235}K(\frac{x_2+hx_{42}-x_1}{h})K(x_{42})K(x_{43})K(x_{54})f(u_3,v_3,x_2+h(x_{42}-x_{43}), \\
& u_4,v_4,x_2+hx_{42},u_5,v_5,x_2+h(x_{42}+x_{54}))du_3dv_3dx_{42}du_4dv_4dx_{43}du_5dv_5dx_{54}
\end{aligned}
$$

Note that $E[\Psi_{N2}(\xi_i,\xi_j)]^2$ can be expanded into several terms, and each of these terms

can be shown to be of order h^7. Now we give the proof for the first term as follows:

$$
\begin{aligned}
& E[E[P_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5)|\xi_1=\xi_i,\xi_2=\xi_j]^2] \\
= & \int [\int d^s_{1234} K_{15}K_{25}K_{35}K_{45} f(\xi_3,\xi_4,\xi_5)\prod_{i=3}^{5} d\xi_i]^2 f(\xi_1,\xi_2)d\xi_1 d\xi_2 \\
= & h^6 \int [\int d^s_{1234} K(\frac{x_2+hx_{52}-x_1}{h})K(x_{52})K(x_{53})K(x_{54}) f(u_3,v_3,x_2+h(x_{52}-x_{53}), \\
& u_4,v_4,x_2+h(x_{52}-x_{54}),u_5,v_5,x_2+hx_{52})du_3dv_3dx_{53}du_4dv_4dx_{54}du_5dv_5dx_{52}]^2 \\
& f(u_1,v_1,x_1,u_2,v_2,x_2)du_1dv_1dx_1du_2dv_2dx_2 \\
= & h^7 \int [\int d^s_{1234} K(x_{21}+x_{52})K(x_{52})K(x_{53})K(x_{54}) f(u_3,v_3,x_1+h(x_{21}+x_{52}-x_{53}), \\
& u_4,v_4,x_1+h(x_{21}+x_{52}-x_{54}),u_5,v_5,x_1+h(x_{21}+x_{52})) \\
& du_3dv_3dx_{53}du_4dv_4dx_{54}du_5dv_5dx_{52}]^2 \\
& f(u_1,v_1,x_1,u_2,v_2,x_1+hx_{21})du_1dv_1dx_1du_2dv_2dx_{21} \\
= & O(h^7)
\end{aligned}
$$

Finally, we can get $E[\Psi_{N2}(\xi_i,\xi_j)]^2 = O_p(h^7)$, which implies that

$$\sigma_N^2 = \sum_{q\le i<j\le n-1} E[\Psi_{N2}(\xi_i,\xi_j)]^2 = O(N^2h^7).$$

Now we will verify the conditions one by one. For some small constant $0<\delta<1$,

$$
\begin{aligned}
& E|\Psi_{N2}(\xi_1,\xi_j)\Psi_{N2}(\xi_i,\xi_j)|^{1+\delta} \\
= & E|E[\Psi_N(\xi_1,\xi_j,\xi_3,\xi_4,\xi_5)] * E[\Psi_N(\xi_i,\xi_j,\xi_3,\xi_4,\xi_5)]|^{1+\delta} \\
= & \int |E[\Psi_N(\xi_1,\xi_j,\xi_3,\xi_4,\xi_5)] * E[\Psi_N(\xi_i,\xi_j,\xi_3,\xi_4,\xi_5)]|^{1+\delta} f(\xi_1,\xi_i,\xi_j)d\xi_1 d\xi_i d\xi_j
\end{aligned}
$$

Similar to the method of evaluating the order of σ_N^2, with one more transformation $\xi_j = \xi_1 + h\xi_{j1}$ and $\xi_j = \xi_i + h\xi_{ji}$ in the integral, we can obtain

$$E|\Psi_{N2}(\xi_i,\xi_k)\Psi_{N2}(\xi_j,\xi_k)|^{1+\delta} = O(h^{6(1+\delta)+2}).$$

Consequently, $M_{N1} = O(h^{6(1+\delta)+2})$.

Therefore,

$$\lim_{N\to\infty} \frac{N^2 M_{N1}^{1/(1+\delta)}}{\sigma_N^2} = \lim_{h\to 0} O(h^{\frac{1-\delta}{1+\delta}}) = 0.$$

Based on the same procedure, we easily get $M_{N2} = O(h^{12(1+\delta)+2})$. And then,

$$\lim_{N\to\infty} \frac{N^{3/2} M_{N2}^{1/2(1+\delta)}}{\sigma_N^2} = \lim_{N\to\infty} O\left(\frac{h^{\frac{1-\delta}{2(1+\delta)}}}{(Nh)^{\frac{1}{2}}}\right) = 0.$$

For M_{N3}, we have $M_{N3} = O(h^{14})$, which infers that

$$\lim_{N\to\infty} \frac{N^{3/2} M_{N3}^{1/2}}{\sigma_N^2} = \lim_{N\to\infty} O\left(\frac{1}{N^{\frac{1}{2}}}\right) = 0.$$

For M_{N4}, we have $M_{N4} = O(h^{12(1+\delta)+2})$, which indicates that

$$\lim_{N\to\infty} \frac{N^{3/2} M_{N4}^{1/2(1+\delta)}}{\sigma_N^2} = \lim_{N\to\infty} O\left(\frac{h^{\frac{1-\delta}{2(1+\delta)}}}{(Nh)^{\frac{1}{2}}}\right) = 0.$$

Next, we continue to consider the order of M_{N5} and M_{N6}.

For M_{N5},

$$\begin{aligned} & \int \Psi_{N2}(\xi_1, \xi_i)\Psi_{N2}(\xi_1, \xi_j) f(\xi_1) d\xi_1 \\ = & \int (E[\Psi_N(\xi_1, \xi_i, \xi_3, \xi_4, \xi_5)]) * (E[\Psi_N(\xi_1, \xi_j, \xi_3, \xi_4, \xi_5)]) f(\xi_1) d\xi_1 \end{aligned}$$

and

$$\begin{aligned} & \int (E[P_N(\xi_1, \xi_i, \xi_3, \xi_4, \xi_5)]) * (E[P_N(\xi_1, \xi_j, \xi_3, \xi_4, \xi_5)]) f(\xi_1) d\xi_1 \\ = & \int (h^3 \int d_{1i34}^s K(\frac{x_i + hx_{5i} - x_1}{h}) K(x_{5i}) K(x_{53}) K(x_{54}) f(u_3, v_3, x_i + h(x_{5i} - x_{53}), \\ & u_4, v_4, x_i + h(x_{5i} - x_{54}), u_5, v_5, x_i + hx_{5i}) du_3 dv_3 dx_{53} du_4 dv_4 dx_{54} du_5 dv_5 dx_{5i}) \\ & \cdot (h^3 \int d_{1j34}^s K(\frac{x_j + hx_{5j} - x_1}{h}) K(x_{5j}) K(x_{53}) K(x_{54}) f(u_3, v_3, x_j + h(x_{5j} - x_{53}), \\ & u_4, v_4, x_j + h(x_{5j} - x_{54}), u_5, v_5, x_j + hx_{5j}) du_3 dv_3 dx_{53} du_4 dv_4 dx_{54} du_5 dv_5 dx_{5j}) \\ & \cdot f(u_1, v_1, x_1) du_1 dv_1 dx_1 \\ = & h^7 \int (\int d_{1i34}^s K(x_{i1} + x_{5i}) K(x_{5i}) K(x_{53}) K(x_{54}) f(u_3, v_3, x_i + h(x_{5i} - x_{53}), \\ & u_4, v_4, x_i + h(x_{5i} - x_{54}), u_5, v_5, x_i + hx_{5i}) du_3 dv_3 dx_{53} du_4 dv_4 dx_{54} du_5 dv_5 dx_{5i}) \\ & \cdot (\int d_{1j34}^s K(\frac{x_j - x_i + hx_{5j} + hx_{i1}}{h}) K(x_{5j}) K(x_{53}) K(x_{54}) f(u_3, v_3, x_j + h(x_{5j} - x_{53}), \\ & u_4, v_4, x_j + h(x_{5j} - x_{54}), u_5, v_5, x_j + hx_{5j}) du_3 dv_3 dx_{53} du_4 dv_4 dx_{54} du_5 dv_5 dx_{5j}) \\ & \cdot f(u_1, v_1, x_i - hx_{i1}) du_1 dv_1 dx_{i1} \end{aligned}$$

With one more transformation $x_j = x_i + hx_{ji}$, we can verify $M_{N5} = O(h^{14(1+\delta)+1})$. And then,

$$\lim_{N\to\infty} \frac{N^2 M_{N5}^{1/2(1+\delta)}}{\sigma_N^2} = \lim_{h\to 0} O(h^{\frac{1}{2(1+\delta)}}) = 0.$$

For M_{N6}, similarly, we can obtain $M_{N6} = O(h^{15})$ and

$$\lim_{N\to\infty} \frac{N^2 M_{N6}^{1/2}}{\sigma_N^2} = \lim_{h\to 0} O(h^{\frac{1}{2}}) = 0.$$

Now, we choose some $0 < \delta, \delta' < 1$ satisfying the condition $\delta' < \frac{2\delta}{2\delta+3}$, which infers that

$$\sum_{k=1}^{\infty} k^2 \{\beta(k)\}^{\frac{\delta}{1+\delta}} \leq \sum_{k=1}^{\infty} k^2 * Ck^{-\frac{(2+\delta')\delta}{\delta'(1+\delta)}} \leq C \sum_{k=1}^{\infty} \frac{1}{k^{\frac{(2+\delta')\delta}{\delta'(1+\delta)}-2}} < \infty,$$

where $\frac{(2+\delta')\delta}{\delta'(1+\delta)} - 2 > 1$ and C is a generic constant. Hence, we have

$$\max \frac{1}{\sigma_N^2} \left\{ N^2 \{M_{N1}^{\frac{1}{1+\delta}} + M_{N5}^{\frac{1}{2(1+\delta)}} + M_{N6}^{\frac{1}{2}}\}, N^{\frac{3}{2}} \{M_{N2}^{\frac{1}{2(1+\delta)}} + M_{N3}^{\frac{1}{2}} + M_{N4}^{\frac{1}{2(1+\delta)}}\} \right\} \to 0$$

as $N \to \infty$. By Lemma A, we get

$$\frac{1}{\sigma_N} \sum_{(N,2)} \Psi_{N2}(\xi_i, \xi_j)) \to N(0,1).$$

Finally, we want to show that the remainder term $R_N^{(2)}$ is of a smaller order than the first term, where

$$R_N^{(2)} = \sum_{j=3}^{5} \frac{\binom{5}{j}}{\binom{N}{j}} \sum_{(N,j)} h_N^{(j)}(\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_j}) \triangleq S_3 + S_4 + S_5.$$

For S_3,

$$S_3 = \frac{\binom{5}{3}}{\binom{N}{3}} \sum_{(N,3)} h_N^{(3)}(\xi_1, \xi_2, \xi_3) = \frac{\binom{5}{3}}{\binom{N}{3} h^4} \sum_{(N,3)} [\Psi_{N3}(\xi_1, \xi_2, \xi_3) - \Psi_{N2}(\xi_{i_1}, \xi_{i_2})].$$

and

$$Var[\Psi_{N3}(\xi_{i_1}, \xi_{i_2}, \xi_{i_3})] = E[\Psi_{N3}(\xi_{i_1}, \xi_{i_2}, \xi_{i_3})]^2$$

Similarly, $E[\Psi_{N3}(\xi_{i_1}, \xi_{i_2}, \xi_{i_3})]^2$ can also be expanded into several terms, and each of

these terms be shown to be of order h^6. We only give the proof of the first term.

$$
\begin{aligned}
& E[P_{N3}(\xi_1,\xi_2,\xi_3)]^2 \\
= & \int [\int d^s_{1234} K_{15}K_{25}K_{35}K_{45} f(\xi_4,\xi_5) d\xi_4 d\xi_5]^2 f(\xi_1,\xi_2,\xi_3) d\xi_1 d\xi_2 d\xi_3 \\
= & \int [h^2 \int d^s_{1234} K(\frac{x_3+hx_{53}-x_1}{h}) K(\frac{x_3+hx_{53}-x_2}{h}) K(x_{53}) \\
& \cdot K(x_{54}) f(u_4,v_4,x_3+hx_{53}-hx_{54},u_5,v_5,x_3+hx_{53}) du_4 dv_4 dx_{54} du_5 dv_5 dx_{53}]^2 \\
& f(u_1,v_1,x_1,u_2,v_2,x_2,u_3,v_3,x_3) du_1 dv_1 dx_1 du_2 dv_2 dx_2 du_3 dv_3 dx_3 \\
= & h^4 \int [\int d^s_{1234} K(\frac{x_3+hx_{53}-x_1}{h}) K(\frac{x_3+hx_{53}-x_2}{h}) K(x_{53}) \\
& \cdot K(x_{54}) f(u_4,v_4,x_3+hx_{53}-hx_{54},u_5,v_5,x_3+hx_{53}) du_4 dv_4 dx_{54} du_5 dv_5 dx_{53}]^2 \\
& f(u_1,v_1,x_1,u_2,v_2,x_2,u_3,v_3,x_3) du_1 dv_1 dx_1 du_2 dv_2 dx_2 du_3 dv_3 dx_3 \\
= & h^6 \int [\int d^s_{1234} K(x_{31}+x_{53}) K(x_{32}+x_{53}) K(x_{53}) K(x_{54}) \\
& \cdot f(u_4,v_4,x_1+hx_{31}+hx_{53}-hx_{54},u_5,v_5,x_1+hx_{31}+hx_{53}) du_4 dv_4 dx_{54} du_5 dv_5 dx_{53}]^2 \\
& f(u_1,v_1,x_1,u_2,v_2,x_1+hx_{31}-hx_{32},u_3,v_3,x_1+hx_{31}) \\
& \cdot du_1 dv_1 dx_1 du_2 dv_2 dx_{32} du_3 dv_3 dx_{31} \\
= & O(h^6)
\end{aligned}
$$

Therefore, $E[\Psi_{N3}(\xi_1,\xi_2,\xi_3)]^2 = O_p(h^6)$. Furthermore,

$$
\begin{aligned}
& |Cov(\Psi_{N3}(\xi_{i_1},\xi_{i_2},\xi_{i_3}), \Psi_{N3}(\xi_{j_1},\xi_{j_2},\xi_{j_3}))| \\
\leq & \sqrt{E(\Psi_{N3}(\xi_{i_1},\xi_{i_2},\xi_{i_3}))} * \sqrt{E(\Psi_{N3}(\xi_{j_1},\xi_{j_2},\xi_{j_3}))} \\
= & O(h^6)
\end{aligned}
$$

and

$$
\begin{aligned}
& |Cov(\Psi_{N3}(\xi_{i_1},\xi_{i_2},\xi_{i_3}), \Psi_{N2}(\xi_{j_1},\xi_{j_2}))| \\
\leq & \sqrt{E(\Psi_{N3}(\xi_{i_1},\xi_{i_2},\xi_{i_3}))} * \sqrt{E(\Psi_{N2}(\xi_{j_1},\xi_{j_2}))} \\
= & O(h^{6.5})
\end{aligned}
$$

Consequently, $Var[\sum_{(N,3)} [\Psi_{N3}(\xi_1,\xi_2,\xi_3) - \Psi_{N2}(\xi_{i_1},\xi_{i_2})]] = O(N^3h^6)$ as $Nh \to \infty$. And then, $Var(S_3) = O(\frac{1}{N^6h^8} N^3h^6) = O_p(\frac{1}{N^3h^2})$. Using the same procedure, we can obtain

that

$$Var(S_4) = O\left(\frac{1}{N^8h^8}N^4h^5\right) = O\left(\frac{1}{N^4h^3}\right)$$

and

$$Var(S_5) = O\left(\frac{1}{N^{10}h^8}N^5h^4\right) = O\left(\frac{1}{N^5h^4}\right).$$

Let $S_2 = \frac{\binom{5}{2}}{\binom{N}{2}h^4}\sum_{(N,2)}\Psi_{N2}(\xi_{i_1},\xi_{i_2})$, and then we can find that

$$Var(S_2) = O\left(\frac{1}{N^4h^8}\sigma_N^2\right) = O\left(\frac{1}{N^2h}\right).$$

Furthermore, we can easily illustrate that as $Nh \to \infty$,

$$\frac{Var(S_3)}{Var(S_2)} = O\left(\frac{1}{Nh}\right) = o(1),$$

$$\frac{Var(S_4)}{Var(S_2)} = O\left(\frac{1}{(Nh)^2}\right) = o(1),$$

$$\frac{Var(S_5)}{Var(S_2)} = O\left(\frac{1}{(Nh)^3}\right) = o(1),$$

$$\frac{|Cov(S_3,S_4)|}{Var(S_2)} \le \frac{\sqrt{Var(S_3)Var(S_4)}}{S_2} = O\left(\frac{1}{(Nh)^{3/2}}\right) = o(1),$$

$$\frac{|Cov(S_4,S_5)|}{Var(S_2)} \le \frac{\sqrt{Var(S_4)Var(S_5)}}{S_2} = O\left(\frac{1}{(Nh)^{5/2}}\right) = o(1),$$

$$\frac{|Cov(S_3,S_5)|}{Var(S_2)} \le \frac{\sqrt{Var(S_3)Var(S_5)}}{S_2} = O\left(\frac{1}{(Nh)^2}\right) = o(1);$$

$$Var(Nh^{1/2}S_3) = O\left(\frac{1}{Nh}\right) = o(1),$$

$$Var(Nh^{1/2}S_4) = O\left(\frac{1}{(Nh)^2}\right) = o(1),$$

$$Var(Nh^{1/2}S_5) = O\left(\frac{1}{(Nh)^3}\right) = o(1).$$

Consequently, based on these results, we can know that

$$Nh^{1/2}S_N \xrightarrow{d} N(0, \frac{400}{(N-1)^2h^7}\sigma_N^2).$$

Actually, by the definition of σ_N^2, we have $\sigma_N^2 = \sum_{1\le i<j\le N} E[\Psi_{N2}(\xi_i,\xi_j)]^2$. And from the proof above, we can know that for each pair $1 \le i < j \le N = n-q$, there exists one constant C_{ij} such that $E[\Psi_{N2}(\xi_i,\xi_j)]^2 = C_{ij}h^7$. Let $\sigma^2 = \frac{400}{(n-q-1)^2}\sum_{i<j} C_{ij}$, and then

$$nh^{1/2}S_n \xrightarrow{d} N(0,\sigma^2).$$

(ii) If $\{X_t\}_0^n$ dose not satisfy the Markov property, then $S_a > 0$. By Theorem 2.2.4, we easily know that $S_n \xrightarrow{P} S_a > 0$, which implies that

$$nh^{1/2}S_n = \sqrt{n}\sqrt{nh} * S_n \xrightarrow{P} \infty.$$

Proof of Theorem 3.2.6: Under H_{1n}, we have

$$f(u|x,v) = (1-\delta_n)f(u|x) + \delta_n g(v),$$

hence,

$$f(u,v|x) = f(u|x,v)f(v|x) = (1-\delta_n)f(u|x)f(v|x) + \delta_n g(v)f(v|x),$$

and then

$$\begin{aligned}
& |\phi_{U,V|X}(u,v) - \phi_{U|X}(u)\phi_{V|X}(v)|^2 \\
= & \Big| \int \exp\{iu'U + iv'V\} f(U,V|X)dUdV \\
& - \int \exp\{iu'U\} f(U|X)dU \int \exp\{iv'V\} f(V|X)dV \Big|^2 \\
= & \left| \int \exp\{iu'U + iv'V\} \delta_n [g(V) - f(U|X)] f(V|X)dUdV \right|^2 \\
= & \delta_n^2 \left| \int \exp\{iu'U + iv'V\} [g(V) - f(U|X)] f(V|X)dUdV \right|^2
\end{aligned}$$

and

$$\begin{aligned}
S_a &= E[\mathcal{D}^2(U,V|X)a(X)] \\
&= \int \mathcal{D}^2(U,V|X)a(X)dF(X) \\
&= \int \frac{1}{c_{d_u}c_{d_v}} \int_{\mathbb{R}^{d_u+d_v}} \frac{|\phi_{U,V|X}(u,v) - \phi_{U|X}(u)\phi_{V|X}(v)|^2}{|u|^{d_u+1}|v|^{d_v+1}} dudvdF(X) \\
&= \int \frac{1}{c_{d_u}c_{d_v}} \int_{\mathbb{R}^{d_u+d_v}} \frac{\delta_n^2 \left|\int \exp\{iu'U + iv'V\}[g(V) - f(U|X)]f(V|X)dUdV\right|^2}{|u|^{d_u+1}|v|^{d_v+1}} dudvdF(X) \\
&= \delta_n^2 \int \frac{1}{c_{d_u}c_{d_v}} \int_{\mathbf{R^{d_u+d_v}}} \frac{\left|\int \exp\{iu'U + iv'V\}[g(V) - f(U|X)]f(V|X)dUdV\right|^2}{|u|^{d_u+1}|v|^{d_v+1}} dudvdF(X) \\
&\triangleq \delta_n^2 \mu_1
\end{aligned}$$

If $\delta_n = O(n^{-1/2}h^{-1/4})$, then $Nh^{1/2}S_N \to Nh^{1/2}\delta_n^2\mu_1 = C\mu_1$ for some constant C.

Next, we still consider the Hoeffding's decomposition of S_N in the Proof of Theorem 2.2.4,

$$S_N = \theta + \sum_{j=1}^{5} \binom{5}{j} H_N^{(j)},$$

where $H_N^{(j)} = \binom{N}{j}^{-1} \sum_{(N,j)} h^{(j)}(w_{i_1}, \cdots, w_{i_j})$ and $\theta = \frac{1}{h^4} \int \Psi_N(\xi_1,\xi_2,\xi_3,\xi_4,\xi_5) \prod_{i=1}^{5} dF(\xi_i)$.

Let $\varphi(\xi_i,\xi_j) = \Psi_{N2}(\xi_i,\xi_j) - h^4\theta$, then $H_N^{(2)} = \frac{1}{\binom{N}{2}} \sum_{(N,2)} \varphi(\xi_i,\xi_j)$, which implies that $H_N^{(2)}$ is a U-statistic of degree two based on samples $\{\xi_q, \xi_{q+1}, \ldots, \xi_{n-1}\}$. Based on Theorem 2 in Section 3.7.3 of Lee [1990], we can get

$$\frac{\sqrt{N}}{2\sigma_{1A}} \frac{1}{\binom{N}{2}} \sum_{(N,2)} \varphi(\xi_i,\xi_j) \to N(0,1),$$

where $\sigma_{1AN}^2 = nVar(\varphi_{N1}(\xi_i))$. Consequently, by the relationship between θ and S_a from the proof of Theorem 3.2.4 and let $\sigma_{1A}^2 = \frac{400\sigma_{1AN}^2}{h^7}$, and $N = n - q$, we can know that

$$nh^{1/2}S_n \xrightarrow{d} N(C\mu_1, \sigma_{1A}^2).$$

Chapter 4

Discussion

This thesis has discussed two important problems–interaction pursuit and hypothesis testing of Markov property in the two kinds of complex models–generalize linear models and time series models, respectively.

In Chapter 2, we talk about the interaction screening problem and provide the efficient algorithm DSSI. It depends on the Boolean representation and discretization. In fact, Discretization plays essential roles in various algorithms of many fields such as decision tree in data mining although it can lose some information compared to the original data set. It can make some methods more efficient and powerful. Hence, the idea of discretization should be extended to other methods or algorithms. It is one greatly interesting research topic in the future.

How to choose the threshold value in the second step of DSSI is an interesting and important problem. If the threshold value is very large, we would miss many important interaction terms; if it is very small, the false selection rate will increase. As discussed in Wan et al. [2010a], the threshold value can be taken by the Bonferroni correction, but sometimes signals are a little weaker, more than n terms would pass the threshold value. In this case, We need to take another threshold value. In practice, our favor is to pick out sufficiently many terms such that $|\mathcal{N}_{\gamma_n}| = n$, $n-1$ or $n/\log(n)$. The discretization parameter l is also important. In our method, $l = 3$ is available to ensure that the method "DSSI" is powerful to select the significant interaction terms.

The key idea of SSI and DSSI is to use marginal utilities to filter out the insignifi-

cant interaction effects, therefore, it will encounter the same issue as the SIS. Firstly, SSI and DSSI may miss some important interaction terms that are jointly correlated but marginally uncorrelated with the response after screening. Secondly, some unimportant interaction terms are highly correlated with the significant ones, they have higher priority to be selected than other important ones that are relatively weakly related to the response. To address these issues, we propose to extend SSI and DSSI to iterative SSI (ISSI) and iterative DSSI (IDSSI), which is motivated by the idea of the iterative SIS (ISIS) proposed by Fan and Lv [2008] and Fan et al. [2009]. In the first step, we apply SIS and SSI (DSSI) to the original data set and select the top d_1 main effects and d_2 interaction terms, respectively, and then use a regularization method such as LASSO or SCAD methods, thus we can obtain a subset A_1 with size k_{a_1} of main effects and one subset B_1 with size k_{b_1} of interactions. Let $C_1 = A_1 \bigcup B_1$, then we have a $n \times 1$ residual vector by regressing the response Y and the variables in the set C_1. In the next step, by treating that residual vector as the new response and applying the same method as in the first step to the remaining $p - k_{a_1}$ main variables and $q - k_{b_1}$ interaction terms, where $q = p(p-1)/2$, we will have a subset A_2 of k_{a_2} main effects and one subset B_2 of k_{b_2} interaction terms. Thirdly, we iteratively repeat the previous step until the union $A = \bigcup A_i$ and $B = \bigcup B_j$ reach a given sized d_1 and d_2, which are less than n. Finally, we can select the important main effects and interactions by using a moderate method such as LASSO, SCAD to the final two sets A and B. In fact, this procedure combines SIS and SSI (DSSI) and can simultaneously select the important main effects and interactions.

SSI (DSSI) focuses on the screening of two-way interaction effects in this paper. Sometimes more than two variables have simultaneous influence on another one variable in the models, this influence is called as higher-order interactions such as three-way interactions, four-way interactions. Higher order interactions also play an important role in the complex diseases. For instance, Ritchie et al. [2001] claimed that they reported firstly the four-locus interactions associated with the complex disease and indicated that the four-locus interactions of the four genes COMT,CYP1A1, CYP1B1 and GSTM1 is significantly related to the risk for sporadic breast cancer by using the method MDR proposed in Ritchie et al. [2001]. Therefore, another

extension of SSI (DSSI) is to exploit the same idea to identify the higher order interactions. For example, for three interactions, we can consider the increments of the log-likelihood functions between these two generalized linear models:

$$g(E(Y|x)) = \beta_0 + X_i\beta_i + X_j\beta_j + X_k\beta_k + X_iX_j\beta_{ij} + X_iX_k\beta_{ik} + X_jX_k\beta_{jk}$$

and

$$g(E(Y|x)) = \beta_0 + X_i\beta_i + X_j\beta_j + X_k\beta_k + X_iX_j\beta_{ij} + X_iX_k\beta_{ik} + X_jX_k\beta_{jk} + X_iX_jX_k\beta_{ijk}.$$

Since the number of features in the models increases exponentially with the order of interactions, an exhaustive search may be restrained, a more efficient algorithm is needed. This leads to another interesting issue of future work and is beyond the scope of this thesis.

In Chapter 3, we consider another essential testing problem about Markovian assumption in time series models. The conditional distance covariance is utilized to construct the test statistic. And there are some remaining problems in this chapter. One of the most fundamental things is to find the appropriate estimator of the lag order q in each model since the power of our proposed test may be affected by it. When q is larger, $\mathbb{H}_0^*$ is more close to original null hypothesis $\mathbb{H}_0$, but the sample size $N = n - q$ will be smaller. Hence, to find the balance between q and N is one important thing in the future. And also, we will utilize some real data sets such as Chicago Board Options Exchange's Volatility Index (VIX) and 3-month Treasury Bill data, to identify the power of our test.

Another thing is that we only consider one dimensional stochastic process, therefore, we will extend our one dimension test to d dimension test in the future work and make the proposed test become more general. Furthermore, the computational efficiency is not excellent. We will try to offer a method to simplify the null hypothesis distribution approximation and then provide the more efficient algorithm to carry out this test procedure. This also gives rise to one interesting topic for future work.

Bibliography

A. Agresti. *Categorical data analysis*, volume 2. Wiley New York:, 2002.

Y. Aït-Sahalia, J. Fan, and H. Peng. Nonparametric transition-based tests for jump diffusions. *Journal of the American Statistical Association*, 104(487):1102–1116, 2009.

Y. Aït-Sahalia, J. Fan, and J. Jiang. Nonparametric tests of the markov hypothesis in continuous-time models. *The Annals of Statistics*, 38(5):3129–3163, 2010.

Y. Aitï-Sahalia. Do interest rates really follow continuous-time markov diffusions? *Manuscript, Graduate School of Business, University of Chicago*, 1996.

H. Akaike. Information theory and an extension of the maximum likelihood principle. pages 267–281, 1973.

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

D. M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

W. Bateson. Mendel's principles of heredity. *University Press, Cambridge*, 1909.

R. Bellman. A markovian decision process. Technical report, DTIC Document, 1957.

J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111–1141, 2013.

B. Chen and Y. Hong. Characteristic function–based testing for multifactor continuous-time markov models via nonparametric regression. *Econometric Theory*, 26(04):1115–1179, 2010.

B. Chen and Y. Hong. Testing for the markov property in time series. *Econometric Theory*, 28(01):130–178, 2012.

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

H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.

N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489): 354–364, 2010.

H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

D. R. Cox. Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24, 1984.

N. J. Cox, M. Frigge, D. L. Nicolae, P. Concannon, C. L. Hanis, G. I. Bell, and A. Kong. Loci on chromosomes 2 (niddm1) and 15 interact to increase susceptibility to diabetes in mexican americans. *Nature genetics*, 21:213–215, 1999.

J. A. De Matos and M. Fernandes. Testing the markov property with high frequency data. *Journal of Econometrics*, 141(1):44–64, 2007.

F. Demichelis, K. Fall, S. Perner, O. Andrén, F. Schmidt, S. Setlur, Y. Hoshida, J. Mosquera, Y. Pawitan, C. Lee, et al. Tmprss2: Erg gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene*, 26(31):4596–4599, 2007.

W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–826, 2001.

C. Dong, S. Wang, W.-D. Li, D. Li, H. Zhao, and R. A. Price. Interacting genetic loci on chromosomes 20 and 10 influence extreme human obesity. *The American Journal of Human Genetics*, 72(1):115–124, 2003.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

J. Fan. Comments on "wavelets in statistics: A review" by A. Antoniadis. *Statistical Methods and Applications*, 6(2):131–138, 1997.

J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

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

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

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

J. Fan and R. Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

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

J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.

J. Fan, F. Han, and H. Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014a.

J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284, 2014b.

Y. Fan and C. Y. Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, 2013.

Y. Fan, Y. Kong, D. Li, and J. Lv. Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*, 2016.

W. Feller. Non-markovian processes with the semi-group property. *The Annals of Mathematical Statistics*, 30:1252–1253, 1959.

S. E. Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, pages 907–917, 1970.

R. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edin.*, 52:399–433, 1918.

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

J. H. Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722–738, 2012.

J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

C. Gao, N. Wang, Q. Yu, and Z. Zhang. A feasible nonconvex relaxation approach to feature selection. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 356–361. AAAI Press, 2011.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

S. J. Haberman. *The analysis of frequency data*. University of Chicago Press Chicago, 1974.

P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18 (3):533–550, 2009.

P. Hall and J.-H. Xue. On selecting interacting features from high-dimensional data. *Computational Statistics & Data Analysis*, 71:694–708, 2014.

P. Hall, D. Titterington, and J.-H. Xue. Tilting methods for assessing the influence of components in a classifier. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):783–803, 2009.

N. Hao and H. H. Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014.

N. Hao, Y. Feng, and H. H. Zhang. Model selection for high dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, (just-accepted), 2016.

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

J. G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, and S. T. Mayne. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

O. Klezovitch, M. Risk, I. Coleman, J. M. Lucas, M. Null, L. D. True, P. S. Nelson, and V. Vasioukhin. A causal role for erg in neoplastic transformation of prostate epithelium. *Proceedings of the National Academy of Sciences*, 105(6):2105–2110, 2008.

Y. Kong, D. Li, Y. Fan, and J. Lv. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *arXiv preprint arXiv:1605.03315*, 2016.

A. J. Lee. U-statistics: Theory and practice. *Statistics: Textbooks and Monographs, Marcel Dekker, Inc.*, 1990.

P. Lees, F. Cunningham, and J. Elliott. Principles of pharmacodynamics and their applications in veterinary pharmacology. *Journal of veterinary pharmacology and therapeutics*, 27(6):397–414, 2004.

G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, pages 1846–1877, 2012a.

J. Li, W. Zhong, R. Li, and R. Wu. A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *The annals of applied statistics*, 8(4): 2292–2318, 2014.

R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012b.

H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data mining and knowledge discovery*, 6(4):393–423, 2002.

J. Liu, R. Li, and R. Wu. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274, 2014.

R. E. Lucas, Jr. Asset prices in an exchange economy. *Econometrica: Journal of the Econometric Society*, pages 1429–1445, 1978.

R. E. Lucas Jr and E. C. Prescott. Investment under uncertainty. *Econometrica: Journal of the Econometric Society*, pages 659–681, 1971.

J. Luo, D. J. Duggan, Y. Chen, J. Sauvageot, C. M. Ewing, M. L. Bittner, J. M. Trent, and W. B. Isaacs. Human prostate cancer and benign prostatic hyperplasia. *Cancer research*, 61(12):4683–4688, 2001.

J.-H. Luo, Y. P. Yu, K. Cieply, F. Lin, P. Deflavia, R. Dhir, S. Finkelstein, G. Michalopoulos, and M. Becich. Gene expression analysis of prostate cancers. *Molecular carcinogenesis*, 33(1):25–35, 2002.

J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, pages 3498–3528, 2009.

J. A. Magee, T. Araki, S. Patil, T. Ehrig, L. True, P. A. Humphrey, W. J. Catalona, M. A. Watson, and J. Milbrandt. Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer research*, 61(15):5692–5696, 2001.

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

A. A. Markov. Theory of algorithms, trudy math. *Inst. VA Steklova*, 42, 1954.

R. Mazumder, J. H. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495): 1125–1138, 2011.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

J. Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77, 1977.

M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM Journal on Applied Mathematics*, 61(2):633–658, 2000.

R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

E. Paparoditis and D. N. Politis. The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, 52(1):139–159, 2000.

J. L. Peixoto. Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41(4):311–313, 1987.

J. L. Peixoto. A property of well-formulated polynomial regression models. *The American Statistician*, 44(1):26–30, 1990.

N. Pochet, F. De Smet, J. A. Suykens, and B. L. De Moor. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, 20(17):3185–3195, 2004.

F. Pukelsheim. The three sigma rule. *The American Statistician*, 48(2):88–91, 1994.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, and M. J. Daly. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

M. Rosenblatt and D. Slepian. Nth order markov chains with every N variables independent. *Journal of the Society for Industrial and Applied Mathematics*, 10 (3):537–549, 1962.

A. Schick. On U-statistics with random kernels. *Statistics and probability letters*, 34 (3):275–283, 1997.

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

T. M. Sellke and S. H. Sellke. Chebyshev inequalities for unimodal distributions. *The American Statistician*, 51(1):34–40, 1997.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

L. Su and A. Ullah. Testing conditional uncorrelatedness. *Journal of Business and Economic Statistics*, 27(1):18–29, 2009.

L. Su and H. White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864, 2008.

L. Su and H. White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27–44, 2014.

L. Su and H. L. White. Conditional independence specification testing for dependent processes with local polynomial quantile regression. In *Essays in Honor of Jerry Hausman*, pages 355–434. Emerald Group Publishing Limited, 2012.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

G.-A. Thanei, N. Meinshausen, and R. D. Shah. The xyz algorithm for fast interaction search in high-dimensional data. *arXiv preprint arXiv:1610.05108*, 2016.

H. Theil. Economic forecasts and policy. *Amsterdam: North-Holland*, 1961.

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

X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang, and W. Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010a.

X. Wan, C. Yang, Q. Yang, H. Xue, N. L. Tang, and W. Yu. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, 26(1):30–37, 2010b.

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

X. Wang and C. Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

G. Y. Weintraub, C. L. Benkard, and B. Van Roy. Markov perfect industry dynamics with many firms. *Econometrica*, 76(6):1375–1411, 2008.

E. Wigner. On the interaction of electrons in metals. *Physical Review*, 46(11):1002–1011, 1934.

J. Wu, B. Devlin, S. Ringquist, M. Trucco, and K. Roeder. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, 34(3):275–285, 2010.

C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu. Snpharvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, 25(4):504–511, 2009.

C. Yang, X. Wan, Q. Yang, H. Xue, and W. Yu. Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. *BMC bioinformatics*, 11(1):S18, 2010.

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

X. Zhang, F. Zou, and W. Wang. Fastanova: an efficient algorithm for genome-wide association study. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 821–829. ACM, 2008.

X. Zhang, F. Zou, and W. Wang. Fastchi: an efficient algorithm for analyzing gene-gene interactions. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 528. NIH Public Access, 2009.

X. Zhang, S. Huang, F. Zou, and W. Wang. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26(12):i217–i227, 2010a.

X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang. Coe: a general approach for efficient genome-wide two-locus epistasis test in disease association study. *Journal of Computational Biology*, 17(3):401–415, 2010b.

Y. Zhang and J. S. Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

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

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

CURRICULUM VITAE

Academic qualifications of the thesis author, Mr. ZHOU Min:

- Received the degree of Bachelor of Science (Mathematics and Applied Mathematics) from Soochow University, July 2006.

- Received the degree of Master of Science (Probability and Mathematical Statistics) from Soochow University, July 2009.

August 2017