2023

High Dimensional Data Analysis: variable screening and inference

lei fang
University of Kentucky, lfa246@uky.edu

Digital Object Identifier: https://doi.org/13023/etd.2023.379

Recommended Citation

fang, lei, "High Dimensional Data Analysis: variable screening and inference" (2023). *Theses and Dissertations--Statistics*. 70.
https://uknowledge.uky.edu/statistics_etds/70

lei fang, Student

Dr. Chenglong Ye, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

High Dimensional Data Analysis: Variable Screening and Inference

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Lei Fang
Lexington, Kentucky

Director: Dr.Chenglong Ye, Assistant Professor of Statistics
Lexington, Kentucky
2023

https://orcid.org/0009-0007-0990-3811

ABSTRACT OF DISSERTATION

High Dimensional Data Analysis: Variable Screening and Inference

This dissertation focuses on the problem of high dimensional data analysis, which arises in many fields including genomics, finance, and social sciences. In such settings, the number of features or variables is much larger than the number of observations, posing significant challenges to traditional statistical methods.

To address these challenges, this dissertation proposes novel methods for variable screening and inference. The first part of the dissertation focuses on variable screening, which aims to identify a subset of important variables that are strongly associated with the response variable. Specifically, we propose a robust nonparametric screening method to effectively select the predictors that marginally independent but conditionally dependent on the response.

The second part of the dissertation focuses on an application of high dimensional inference problem in microbiome related disease study. The microbial community in the human gut is teeming with metabolic activity and plays a key role in host physiology and health. But the host-microbiome interactions are not well understood in terms of the molecular mechanism, while the microbial metabolites have been hypothesized to play a critical role. We developed a statistical framework that not only integrate the microbiome and metabolites but also integrate multi-view microbiome data, to inference the causal effect of metabolites for disease outcome. We borrow the idea of debiasing lasso to construct the inference procedures. In numerical study and a real data application, we demonstrate our method's superior performance.

KEYWORDS: High Dimensional inference, Variable Screening, Nonparametric Independence Measure, Causal Inference, Debiased Lasso

Lei Fang

July 19, 2023

High Dimensional Data Analysis: Variable Screening and Inference

By
Lei Fang

Dr.Chenglong Ye
Director of Dissertation

Dr. Katherine Thompson
Director of Graduate Studies

July 19, 2023
Date

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Dr.Ye, whose guidance, support, and patience have been invaluable throughout my PhD journey. He has challenged me to think deeply, to question assumptions, and to strive for excellence in everything I do. Without his guidance, this thesis would not have been possible.

I would also like to thank the members of my thesis committee, Dr.Derek Young, Dr.Arnold Stromberg, Dr.Katherine Thompson, Dr.Solomon W. Harrar, and Dr.Reuben Adatorwovor for their valuable feedback and insightful comments on my work. Their expertise and constructive criticism have made this thesis a better piece of scholarship.

I would also want to thank for my previous advisor Dr. Xiangrong Yin, who unfortunately passed away in 2020. He provided me with valuable guidance and support during the early stages of my research, and his expertise and mentorship were instrumental in shaping my intellectual interests and pursuits. I also want to give a special thank to my collaborators Dr.Yue Wang and Dr. Mai Zhou, for the valuable opportunity to work on interesting research projects.

Finally, I want to express my gratitude to to my family and friends, who have provided me with emotional support and encouragement throughout my PhD journey. Their unwavering belief in me has kept me going during the challenging times, and has made this achievement all the more meaningful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

With the increasing availability of data from various fields, such as genetics, finance, biomedical area, and social media, high dimensional data analysis has become a vital area of research in statistics. With high dimensional data, the number of variables, often much greater than the sample size, which presents challenges in traditional statistical methods, in terms of data processing, analysis, and interpretation. To analyze high dimensional data, typically we believe the sparsity assumption, that is only a few of the variables are truly related to the response variable. Based on this common and reasonable assumption, many high dimensional variable selection method has been proposed. The most famous and popular one is Lasso for its efficient variable selection performance in high dimensional linear model. However, there are limitations, as the consistency and the selection stability requires the assumption of dimensionality and sample size. Such assumption may not valid in real ultra-high dimensional data to guarantee its asymptotic property, e.g. in genetic studies, the number of SNPs can be millions while sample size can only be as large as hundreds. Other variable selection methods are also suffering the so-called "curse of dimensionality". To handle this issue, people often pre-processed the ultra-high dimensional data by variable screening. The goal of variable screening is to select a subset of the predictors, which has small cardinality, while catching all the true related predictors. This whole method depends on some independent measure, especially the nonparametric one to allow flexibility and much broader application in real data.

In this thesis, we aim to contribute to the field of high dimensional data analysis by first developing a novel variable screening method by a new nonparametric dependence measure : Martingale Difference Correlation. This is new measure is motivated to address the potential issue of missing the conditional only active predictors to the mean of response variable. In Chapter 2, we present this new measure and its unbiased estimator based on U-statistics. We also study its asymptotic properties, and most importantly, its sure screening property for variable screening. We demonstrate its robust and superior performance via simulation studies and real data application. All the proofs are included in the supplement A and additional discussion of the marginal measure is included appendix A.

The second part of our contribution to the high dimensional data analysis is a high dimensional inference problem arised in microbime-metabolome study. In general, we want to investigate how the microbime affect human body, indirectly through metabolites or directly associated with some traits or disease. In Chapter 3, we proposed a structural high dimensional linear model to include both microbes and metabolite and explore the causal effect of metabolite by letting some microbes as instrumental variable. We developed the inference for such causal effect following the idea of debiased Lasso. In addition, we have demonstrated the effectiveness of our method through a guided simulation study and a real data application of inflammatory bowl disease.

Chapter 2 Variable Screening via Conditional Martingale Difference Correlation

2.1 Introduction

Variable screening has been a research area that deals with ultrahigh-dimensional data, where high-dimensional methods may fail due to the curse of dimensionality, as [10] suggested. [9]'s seminal work suggests to screen the ultrahigh-dimensional data before conducting variable selection. They proposed a sure independent screening (SIS) method for linear models to screen out inactive variables based on Pearson correlation. After that, variable screening receives more attention since it only requires that the selected set of variables covers the set of active variables, which is referred to as the *sure screening property* [9]. Screening methods with this property suffer less from *instability* [44] that is seen in many variable selection methods.

Various model-based screening methods have been developed. For linear regression models, screening methods have been proposed based on different measures, including marginal Pearson correlation [9], forward regression [38], marginal empirical likelihood ratio [3], and Kendall's rank correlation [19]. For linear quantile regression, screening methods based on quantile partial correlation [23] and conditional quantile correlation [47] have been proposed to handle heterogeneous data. In the context of generalized linear models, screening methods based on the maximum marginal likelihood or its estimate [11], the sparsity-restricted maximum likelihood estimator [42], and Kolmogorov-Smirnov statistic [24] have also been proposed. Other screening methods include model settings such as linear regression models with interactions [15, 12, 17], Cox models [48], varying coefficient models [5, 32], and additive models [8].

An alternative approach is the model-free screening method, which recently has gained popularity due to its less stringent assumptions. [50] proposed a sure independence ranking and screening approach (SIRS) for index models. [20] proposed a screening method (DC-SIS) based on distance correlation, which can be applied to grouped variables. [6] proposed a model-free screening method (MV-SIS) based on empirical conditional distribution function for discriminant analysis. [31] proposed the use of martingale difference correlation (MDC), which can be applied to mean and quantile screening. [25] proposed the fused Kolmogorov filter that works with different types of response variables and high covariate correlation. [27] proposed a screening method based on covariate information number (CIN) motivated by Fisher information. [14] developed a screening framework from the perspective of loss functions and proposed a screening method based on conditional strictly convex losses. Based on *ball correlation*, [28] proposed a generic screening method for biomedical discovery.

As pointed out by [20, 31, 34], screening methods based on marginal measures (e.g., the marginal correlation between the response and each predictor) will possibly miss the marginally but not jointly independent predictors. Two types of approaches have been developed to handle this issue. One approach is *conditional screening*, a screening procedure based on a given set of variables. For example, [2] proposed conditional sure independence screening (CSIS) for assessing the conditional (on a given conditional set) contribution of a predictor to the response in generalized linear models. Based on conditional distance correlation, [40] proposed a method that adjusts for confounding variables. [35] proposed a conditional independence measure and its corresponding screening method (CIS) with false discovery rate (FDR) control, which also works for heavy-tailed predictors/responses. Another approach is *screening via iterative procedures.* For example, the aforementioned forward regression iteratively selects the variables. [42] considered a method based on sparse MLE, where the algorithm iteratively updates the coefficients in the link function. [49] proposed a model-free forward screening method that iteratively updates the conditional set and is robust to outliers. [34] proposed an iterative variable screening method based on random subspace ensembles (RaSE) with a theoretical guarantee for iterative screening procedures.

However, two challenges remain. Model-based iterative methods (e.g., iterative SIS) may rely on a specific variable selection method, which makes the procedure less stable. On the other hand, the conditional screening method requires prior knowledge of the conditional set and its performance becomes unstable if an unreasonable conditional set is selected. It motivates us to develop a stable model-free screening method that identifies both marginally and jointly dependent variables to the response. We propose a kernel-based measure that captures both conditional and marginal mean-independent relationships. In particular, via Bochner's theorem [41], we transform the problem of choosing weights, a key element in our independence measure, to the problem of choosing kernels and their bandwidths in reproducing kernel Hilbert space (RKHS). This flexible kernel-based fashion allows our method to perform well in various settings, as illustrated in the synthetic and real data analysis.

The advantages of our method are as follows. First, we propose a kernel-based independence measure ($\mathrm{CMD}_{\mathcal{H}}$) that is able to characterize both conditional and marginal mean independence. Thus, we propose a $\mathrm{CMD}_{\mathcal{H}}$-based screening method that can detect both marginally and jointly dependent/active variables. Second, the proposed model-free screening method is stable against outliers, data heterogeneity, and high covariate correlation. Third, we show the sure screening property holds for screening both marginally and jointly dependent variables under mild regularity conditions. We also suggest selecting a data-driven conditional set for conducting conditional screening when no prior information is available.

The rest of the article is organized as follows. Section 2.2 introduces the proposed independence measure and its theoretical properties. In Section 2.3, we propose a model-free variable screening procedure and present its sure screening property. The simulation results and two real data examples are reported in Section 2.4, followed by the conclusion in Section 2.5. Additional theorems are presented in the appendix. Auxiliary simulation results and technical proofs are included in the supplementary material.

2.2 General methodology

Notations. Throughout the article, we use upper case (e.g., V) to denote a random variable and use bold font to denote a random vector (e.g., $\boldsymbol{U}$). We use $(\boldsymbol{U}_1', \boldsymbol{U}_2', V')$ and $(\boldsymbol{U}_1'', \boldsymbol{U}_2'', V'')$ to denote *i.i.d* copies of $(\boldsymbol{U}_1, \boldsymbol{U}_2, V)$. For a complex function $f(\boldsymbol{s}) : \mathbb{R}^q \to \mathbb{C}^p$, denote its RKHS norm as $||f(\boldsymbol{s})||^2_{\mathcal{H}_k} = \int_{\mathbb{R}^q} |f(\boldsymbol{s})|^2 w(\boldsymbol{t}) d\boldsymbol{t}$, where $w(\boldsymbol{t})$ corresponds to the kernel function k in RKHS.

For a sequence $\{t_{ij}\}$ with double indices $i, j = 1, ..., n$, we define

$$t^*_{ij} = t_{ij} - \bar{t}_{i.} - \bar{t}_{.j} + \bar{t}_{..}, \tag{2.1}$$

where $\bar{t}_{.j} = \frac{1}{(n-2)} \sum_{i=1}^n t_{ij}$, $\bar{t}_{i.} = \frac{1}{(n-2)} \sum_{j=1}^n t_{ij}$, and $\bar{t}_{..} = \frac{1}{(n-1)(n-2)} \sum_{i=1}^n \sum_{j=1}^n t_{ij}$. Denote $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ as the inner product of any two vectors $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension.

2.2.1 Conditional Martingale Difference Divergence

A motivating example. Assume two random variables X_1 and X_2 are independent with $E(X_2) = 0$, and let $Y = X_1 X_2$. We have $E(Y|X_1) - E(Y) = E(X_2) \cdot (X_1 - E(X_1)) = 0$ as long as $E(X_2) = 0$. The condition "$E(X_2) = 0$" is mild as we can standardize predictors in the dataset to have mean 0 in practice. This example indicates that mean independence measures based on the relationship $E(Y|X_1) - E(Y)$ will misleadingly suggest that X_1 is independent from mean of Y. It was pointed out by [20] and [31] that the marginal measures/methods such as DC and MDC will possibly miss the variables that only jointly contribute to the response variable. It motivates us to consider the following:

$$E(Y|X_1, X_2) - E(Y|X_2),$$

which equals to $X_2 \cdot (X_1 - E(X_1))$ in this example. More generally, we consider the equality $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$, i.e., the response variable V and covariate vector $\boldsymbol{U}_2$ given the covariate vector $\boldsymbol{U}_1$. If in addition, $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are independent, we have $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ if and only if $E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle}|\boldsymbol{U}_2) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})$ for any $\boldsymbol{t}_1$ (the proof is presented in supplementary material S2 (b)). This motivates us to propose the following independence measure in Definition 1, which can be treated as either a conditional or a marginal (see Remark 2) mean independence measure.

Definition 1. *Given a random vector $\boldsymbol{U}_1 \in \mathbb{R}^p$, the **c**onditional **m**artingale difference **d**ivergence of a random variable V and a random vector $\boldsymbol{U}_2 \in \mathbb{R}^q$ is defined as*

$$\begin{aligned} &\mathrm{CMD}_{\mathcal{H}}{}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \\ &= \iint |E(Ve^{i(\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle + \langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle)}) - E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})|^2 w_1(\boldsymbol{t}_1) w_2(\boldsymbol{t}_2) d\boldsymbol{t}_1 d\boldsymbol{t}_2, \end{aligned}$$

where $w_1(\boldsymbol{t}_1)$ and $w_2(\boldsymbol{t}_2)$ are weight functions.

The measure $\mathrm{CMD}_{\mathcal{H}}$ depends on two ingredients: a mean independence measure of a random vector and a random variable, and an adjusting method of the effect of a third vector. It provides valuable information on the conditional contribution of $\boldsymbol{U}_2$ to the mean of V given $\boldsymbol{U}_1$.

Remark 1. *We choose the weight function $w_1(\boldsymbol{t}_1)$ and $w_2(\boldsymbol{t}_2)$ to be integrable, relaxing the strong assumption of the boundedness of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ as in the literature. The choice of an integrable weight function makes the proposed independence measure more flexible. In particular, we can rewrite $\mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ as a functional of kernel functions in RKHS. See more details in Theorem 1(a) and Remark 3.*

We now define a scale-invariant version of the proposed measure.

Definition 2. *Let k_1 and k_2 be the two kernel functions that correspond to the weight function $w_1(\boldsymbol{t}_1)$ and $w_2(\boldsymbol{t}_2)$, as illustrated in Theorem 1(a). We define the **c**onditional **m**artingale difference **c**orrelation*

$$\mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \begin{cases} \dfrac{\mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)}{\sqrt{v(k_2, \boldsymbol{U}_2)v(k_{1_V}, \boldsymbol{U}_1)}} & \textit{if } v(k_2, \boldsymbol{U}_2)v(k_{1_V}, \boldsymbol{U}_1) > 0 \\ 0 & \textit{otherwise,} \end{cases}$$

where $v(k, \boldsymbol{U}) := E[k^2(\boldsymbol{U}, \boldsymbol{U}')] + E^2[k(\boldsymbol{U}, \boldsymbol{U}')] - 2E[k(\boldsymbol{U}, \boldsymbol{U}^{'}) \cdot k(\boldsymbol{U}, \boldsymbol{U}^{''})]$ and $k_V(\boldsymbol{U}, \boldsymbol{U}') := VV^{'}k(\boldsymbol{U}, \boldsymbol{U}')$ for any kernel function k.

Remark 2. *When $\boldsymbol{U}_1$ contains no useful information ($\boldsymbol{U}_1 = \emptyset$, $\boldsymbol{U}_1 \equiv \mathbf{c}$, or $\boldsymbol{U}_1$ is independent from $(V, \boldsymbol{U}_2)$), the definition of $\mathrm{CMD}_{\mathcal{H}}$ reduces to a marginal mean independence measure. That is,*

$$\mathrm{MD}_{\mathcal{H}}{}^2(V, \boldsymbol{U}_2) := \int |E(Ve^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}) - E(V)E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})|^2 w(\boldsymbol{t}_2)d\boldsymbol{t}_2.$$

Note that $\mathrm{MD}_{\mathcal{H}}(V, \boldsymbol{U}_2)$ is a generalized version of MDD *[31] by kernerlizing the L_2 distance of $\boldsymbol{U}_2$ and its i.i.d. copy $\boldsymbol{U}_2'$. The standardized version $\mathrm{MC}_{\mathcal{H}}$ is defined similarly as $\mathrm{CMC}_{\mathcal{H}}$.*

As will be seen in Theorem 1, the definition of the $\mathrm{CMC}_{\mathcal{H}}$ is more convenient for variable screening purpose since it takes values in $[0, 1]$. More detailed discussion of $\mathrm{CMC}_{\mathcal{H}}$ is included in Section 2.3. Now we show the theoretical properties of the proposed conditional independence measure.

Theorem 1. *Assume $E(V^2) < \infty$, we have the following properties:*

(a). We can rewrite $\mathrm{CMD}_{\mathcal{H}}{}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ as

$$\begin{aligned} &\mathrm{CMD}_{\mathcal{H}}{}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \\ &= E(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) + E(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'}))E(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) \\ &- 2E(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{''})), \end{aligned}$$

where k_1 and k_2 are RHKS kernel functions determined by $w_1(\boldsymbol{t}_1)$ and $w_2(\boldsymbol{t}_2)$ defined in Definition 1, respectively.

(b). $0 \leq \mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \leq 1$, and $\mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = 0 \Leftrightarrow E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s. if $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$.

(c). *Given two constants* $d \in \mathbb{R}$ *and* $e \in \mathbb{R}$, $\mathrm{CMC}_{\mathcal{H}}(a + bV, \boldsymbol{c} + d\boldsymbol{U}_2|e\boldsymbol{U}_1) = \mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ *for any scalars* $a, b \in \mathbb{R}$ *and* $\boldsymbol{c} \in \mathbb{R}^q$. *If the kernels* k_1 *and* k_2 *in (a) are scale-invariant, the above equality holds for any scalars* d *and* e *as well.*

(d). *If the random variables* $U, V \in \mathbb{R}$ *are independent, then*

$$\mathrm{MC}_{\mathcal{H}}{}^2(VU, U) = \frac{E^2(V)}{Var(V) + E^2(V) + E^2(U)\frac{Var(V)}{Var(U)}}\mathrm{MC}_{\mathcal{H}}{}^2(U, U).$$

Furthermore, if $E(V) = 0$, *then* $\mathrm{MC}_{\mathcal{H}}(VU, U) = 0$.

Remark 3. *If we take non-integrable weight functions* $w_1(\boldsymbol{t}_1)$ *and* $w_2(\boldsymbol{t}_2)$ *in Definition 1, then* k_1 *and* k_2 *in Theorem 1(a) may not be translation-invariant kernels in RKHS (e.g., the Euclidean distance function). See dCov [33] for an example that adopts a non-integrable weight in its definition.*

Remark 4. *Property (b) shows the equivalence between the conditional mean independence and* $\mathrm{CMC}_{\mathcal{H}}$ *being 0, which suggests* $\mathrm{CMC}_{\mathcal{H}}$ *is a suitable tool for conducting variable screening. Note that the independence of* $\boldsymbol{U}_1$ *and* $\boldsymbol{U}_2$ *in Property (b) is to ease the proof. Indeed, if we define a new independence measure* $\mathrm{CMD}_{\mathcal{H},new}^{\;2}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \iint |E(Ve^{i(\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle + \langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle)}|\boldsymbol{U}_1) - E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle}|\boldsymbol{U}_1)E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}|\boldsymbol{U}_1)|^2 w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2$, *then*

$$\mathrm{CMD}_{\mathcal{H},new}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = 0 \ a.s. \Leftrightarrow E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1) \ a.s..$$

This removes the independence condition of $\boldsymbol{U}_1$ *and* $\boldsymbol{U}_2$. *Then we need to replace our* U*-statistics estimator with the conditional* U*-statistics to estimate the new independence measure. Note that this new measure* $\mathrm{CMD}_{\mathcal{H},new}$ *is a function of the random vector* $\boldsymbol{U}_1$. *Such a conditional measure and its associated screening method is left as future work of interest. In this article, we stick to our original proposed measure* $\mathrm{CMD}_{\mathcal{H}}$. *In the simulations, as we see, even* $\boldsymbol{U}_1$ *and* $\boldsymbol{U}_2$ *are not independent (e.g., high variable correlations* $\rho = 0.5, 0.8, 0.9$ *in Example 1, and nonlinearly associated predictors in Example 4), our variable screening method still performs well, or even outperforms other methods in almost all the simulation settings.*

Remark 5. *Property (c) shows that the proposed* $CMC_{\mathcal{H}}$ *is scale-invariant. Property (d) directly shows the deficiency of marginal-type mean independence measure* ($\mathrm{MC}_{\mathcal{H}}$) *in interaction screening. Thus we propose the variable screening procedure based on* $\mathrm{CMC}_{\mathcal{H}}$.

2.2.2 Empirical Estimators and Asymptotic Properties

Based on property (a) in Theorem 1, we construct the U-statistic to estimate $\mathrm{CMC}_{\mathcal{H}}$.

Definition 3. *Let* $(\boldsymbol{U}_{1i}, \boldsymbol{U}_{2i}, V_i)_{i=1}^n$ *be i.i.d. observations of* $(\boldsymbol{U}_1, \boldsymbol{U}_2, V)$. *Denote* $a_{ij} = V_i V_j k_1(\boldsymbol{U}_{1i}, \boldsymbol{U}_{1j})$ *and* $b_{ij} = k_2(\boldsymbol{U}_{2i}, \boldsymbol{U}_{2j})$ *for* $i, j = 1, ..., n$. *Define the corresponding*

a_{ij}^* and b_{ij}^* as in Equation (2.1). The U-statistic estimator of $\mathrm{CMD}_{\mathcal{H}}$ is

$$\widehat{\mathrm{CMD}}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \frac{1}{n(n-3)} \sum_{1 \le i \ne j \le n} a_{ij}^* b_{ij}^*,$$

and the corresponding estimator of $\mathrm{CMC}_{\mathcal{H}}$ is:

$$\widehat{\mathrm{CMC}}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \frac{\sum_{1 \le i \ne j \le n} a_{ij}^* b_{ij}^*}{\sqrt{\sum_{1 \le i \ne j \le n} a_{ij}^{*2} \sum_{1 \le i \ne j \le n} b_{ij}^{*2}}}.$$

Remark 6. *Compared to the adoption of V-statistic estimator, we choose the U-statistic because it is unbiased and less computationally expensive.*

We now show the strong consistency of the proposed estimators.

Theorem 2. *(Consistency) If $E\left(V^2\right) < \infty$, then*

$$lim_{n \to \infty} \widehat{\mathrm{CMD}}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \text{ a.s.}, \tag{2.2}$$

and

$$lim_{n \to \infty} \widehat{\mathrm{CMC}}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \text{ a.s..} \tag{2.3}$$

In the next theorem, we derive the asymptotic distribution for $\mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$. Denote the following functions: $g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) := E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$, $g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1) := E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})$, and $F(\boldsymbol{t}_1, \boldsymbol{t}_2) := E(V^2 e^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle} e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$. Define the covariance function $\mathrm{cov}_\Gamma((\boldsymbol{t}_1, \boldsymbol{t}_2), (\boldsymbol{t}_1', \boldsymbol{t}_2')) := F(\boldsymbol{t}_1 - \boldsymbol{t}_1', \boldsymbol{t}_2 - \boldsymbol{t}_2') + (F(\boldsymbol{t}_1 - \boldsymbol{t}_1', 0) + g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\overline{g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1')})\{g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}_2')} - g_{\boldsymbol{U}_2}(\boldsymbol{t}_2 - \boldsymbol{t}_2')\} - F(\boldsymbol{t}_1 - \boldsymbol{t}_1', \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}_2')} - F(\boldsymbol{t}_1 - \boldsymbol{t}_1', -\boldsymbol{t}_2')g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)$.

Theorem 3. *Assume $E(V^2) < \infty$, we have the following:*

a. *If $\mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = 0$, then*

$$n\widehat{\mathrm{CMD}}_{\mathcal{H}}^{\,2}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \xrightarrow{d} ||\Gamma(s)||_{\mathcal{H}_k}^2 \tag{2.4}$$

as $n \to \infty$, where $\Gamma(\cdot)$ is a complex-valued zero-mean Gaussian random process with covariance function $\mathrm{cov}_\Gamma((\boldsymbol{t}_1, \boldsymbol{t}_2), (\boldsymbol{t}_1', \boldsymbol{t}_2'))$.

b. *If $\mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = 0$ and $E(V^2|\boldsymbol{U}_2) = E(V^2)$, then*

$$n\widehat{\mathrm{CMD}}_{\mathcal{H}}^{\,2}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)/S_n \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j$$

as $n \to \infty$, where $S_n = (\frac{1}{n}\sum_i V_i^2 - \frac{1}{n(n-1)}\sum_{i \ne j} a_{ij})(1 - \frac{1}{n(n-1)}\sum_{i \ne j} b_{ij})$, $Z_j \overset{i.i.d.}{\sim} \chi_1^2$, and $\{\lambda_j\}_{j=1}^{\infty}$ are nonnegative constants such that $E(\sum_{j=1}^{\infty} \lambda_j Z_j) = 1$.

c. *If $\mathrm{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) > 0$, then $n \cdot \widehat{\mathrm{CMD}}_{\mathcal{H}}^{\,2}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)/S_n \xrightarrow{p} \infty$ as $n \to \infty$.*

The properties stated in the theorems of this section motivate us to propose variable screening algorithm based on $\mathrm{CMC}_{\mathcal{H}}$ and its estimate $\widehat{\mathrm{CMC}}_{\mathcal{H}}$.

2.3 $\mathrm{CMC}_{\mathcal{H}}$-based Variable Screening

In this section, we show the sure screening property of $\mathrm{CMC}_{\mathcal{H}}$ in Section 2.3.1. In Section 2.3.2, we introduce a variable screening algorithm named S-$\mathrm{CMC}_{\mathcal{H}}$ to accommodate the dependence among predictors. The algortihm works reasonably well when the data suffer from outliers, high correlation, and heterogeneity as seen in the numerical studies.

2.3.1 Sure Screening Property

Without loss of generality, let Y be a univariate continuous response variable and $\boldsymbol{X} = (X_1, ..., X_p)^T$ be the predictor vector. Denote the sample as $(X_{1k}, ..., X_{pk}, Y_k)_{k=1}^n$, where $p \gg n$. For any index set $S \subseteq \{1, ..., p\}$, denote $\boldsymbol{X}_S := \{\boldsymbol{X}_j : j \in S\}$. Given a conditional set $\boldsymbol{X}_S$ with cardinality d_1, we define

$$\mathcal{D}_S = \{j : \mathbb{E}(Y|(\boldsymbol{X}_S, X_j)) \text{ depends on } X_j\}$$

as the index set of dependent/active predictors conditional on $\boldsymbol{X}_S$, and

$$\mathcal{I}_S = \{j : \mathbb{E}(Y|(\boldsymbol{X}_S, X_j)) \text{ is independent from } X_j\}$$

as the index set of independent/inactive predictors conditional on $\boldsymbol{X}_S$. Note that $\mathcal{D}_S$ is a subset of $\mathcal{D} := \{j : \mathbb{E}(Y|\boldsymbol{X}) \text{ depends on } X_j\}$, the set of all dependent predictors. Suppressing S, we denote $\omega_j = \mathrm{CMC}^2_{\mathcal{H}}(Y, X_j|\boldsymbol{X}_S)$ as the dependence score of X_j given $\boldsymbol{X}_S$. Let $\hat{\omega}_j = \widehat{\mathrm{CMC}_{\mathcal{H}}}^2(Y, X_j|\boldsymbol{X}_S)$ be the estimator of ω_j and

$$\hat{\mathcal{D}}_S = \{j : \hat{\omega}_j \geq cn^{-\kappa}, \text{ for } j \in S^c\}$$

be the set of selected variables after screening. Before stating the sure screening property, we assume the following conditions.

(A1) There exists a constant $s_0 > 0$ such that $E(\exp(sY^2)) < \infty$ for all $0 < s \leq 2s_0$.

(A2) For any given $\boldsymbol{X}_S$, $\min_{j \in \mathcal{D}_S} \omega_j \geq 2cn^{-\kappa}$ for some constant $c > 0$ and $0 \leq \kappa < 1/2$.

Condition ($A1$) puts constraint on the tail distribution of the response variable and Condition (A2) requires that the conditional active/dependent variables and inactive/independent variables are well separated.

Theorem 4. *Under Condition* (*A*1)*, for any* $0 < \gamma < 1/2 - \kappa$*, there exist positive constants* c_1 *and* c_2 *such that*

$$P\left\{\max_{1 \leq j \leq p-d_1} |\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}\right\} \leq O((p - d_1)[\exp(-c_1 n^{1-2(\kappa+\gamma)}) + n\exp(-c_2 n^{\gamma})]). \tag{2.5}$$

If Conditions (*A*2) *also holds, we have*

$$P(\mathcal{D}_S \subset \hat{\mathcal{D}}_S) \geq 1 - O(s_n[\exp(-c_1 n^{1-2(\kappa+\gamma)}) + n\exp(-c_2 n^{\gamma})]) \tag{2.6}$$

for any conditional set $\boldsymbol{X}_S$*, where* s_n *is the cardinality of* $\mathcal{D}_S$*. In particular, let* $\delta = \min_{j\in\mathcal{D}_S}\omega_j - \max_{j\in\mathcal{I}_S}\omega_j$*, we have the ranking consistency:*

$$P\left\{\max_{j\in\mathcal{I}_S}\widehat{\omega}_j < \min_{j\in\mathcal{D}_S}\widehat{\omega}_j\right\} \geq 1 - 2O((p-d_1)[\exp(-c_1'\delta^2 n^{1-2\gamma}) + n\exp(-c_2 n^{\gamma})]) \tag{2.7}$$

for a positive constant c_1'*.*

Remark 7. *The above theorem shows the sure screening property holds for any given conditional set* $\boldsymbol{X}_S$*. Define* $\mathcal{M} := \{j : \mathbb{E}(Y|X_j)$ *depends on* $X_j\}$ *as the set of marginally dependent/active predictors. For the special case where the conditional set* $\boldsymbol{X}_S = \boldsymbol{X}$*, we have* $\mathcal{D}_S = \mathcal{D}$*, which is the common sure screening property in the literature. In Appendix A.3, we also show that, similarly to Theorem 4, the sure screening property holds when the conditional set is empty. In that case,* $\mathrm{CMC}_{\mathcal{H}}$ *reduces to* $\mathrm{MC}_{\mathcal{H}}$*, and the sure screening property holds for selecting the set* $\mathcal{D}$ *as well as* $\mathcal{M}$*. We discuss more details of selecting the conditional set* $\boldsymbol{X}_S$ *in Section 2.3.2.*

Remark 8. *The error terms* $\exp(-c_1 n^{1-2(\kappa+\gamma)})$ *and* $n\exp(-c_2 n^{\gamma})$ *in (3.5) comes from estimating the three terms in CMD*$_{\mathcal{H}}$ *as in Theorem 1(a). In the proof, take the first term* $E(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) := E(h)$ *for example, we decompose it into a bounded term* $E[hI(h < M)]$ *plus an unbounded term* $E[hI(h > M)]$ *for some large enough* $M > 0$*. Similar decomposition is done to the other two terms. Setting* $M = n^{\gamma}$ *for some* $0 < \gamma < 1/2 - \kappa$*, we obtain the two error terms* $\exp(-c_1 n^{1-2(\kappa+\gamma)})$ *and* $n\exp(-c_2 n^{\gamma})$ *for estimating the sum of the bounded terms and that of the unbounded terms, respectively. The role of the parameter* γ *is a trade-off of estimating the bounded and unbounded terms. By setting* $\gamma = \frac{1-2\kappa}{3}$*, we achieve a balance and obtain the optimal convergence rate. As mentioned in [31], their bound (3.5) can be further improved by assuming a stronger moment condition on* Y*, i.e.,* $E(\exp(sY^4)) < \infty$ *for all* $s \in (0, 2s_0]$*. Their improved bound is the same as our bound. It is also worth mentioning that we do not impose moment conditions on the variable X as in [31]. The reason why our method enjoys a better rate under a weaker condition is that our proposed measure CMC*$_{\mathcal{H}}$ *only computes the RKHS kernel functions of X (as shown in Theorem 1 (a)). Such kernels are bounded, which frees us from assuming additional moment conditions on X. In contrast, the Martingale Difference Correlation requires to calculate the Euclidean distance (Theorem 1 (1) in [31]), which is unbounded. Thus, they require the stronger assumptions.*

2.3.2 A Variable Screening Algorithm S-CMC$_{\mathcal{H}}$

The sure screening property holds for any conditional set. In practice, if the conditional set is not given, we use the top d_1 predictors suggested by $\mathrm{MC}_{\mathcal{H}}$, which also enjoys the sure screening property, as shown in Theorems 10 and 11 in the appendix.

We propose to use the following variable screening algorithm S-CMC$_{\mathcal{H}}$ as stated in Algorithm 1.

Algorithm 1 The procedure of the S-CMC$_{\mathcal{H}}$ for variable screening.

```
Input: The conditional set $\boldsymbol{X}_S$ (optional) and its cardinality $d_1$ (optional), the number of variables $d_2$ to select (optional), and the data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$.

  1. If $\boldsymbol{X}_S$ is not given, calculate $\widehat{\mathrm{MC}_{\mathcal{H}}}(X_i) := \widehat{\mathrm{MC}_{\mathcal{H}}}(Y, X_i)$ for each $i = 1, ..., p$. Let the conditional set $\boldsymbol{X}_S$ be the set of the top $d_1$(if not given, $d_1 = \left\lfloor \sqrt{n/\log n} \right\rfloor$ where $\lfloor\cdot\rfloor$ is the floor function) predictors with the largest $\widehat{\mathrm{MC}_{\mathcal{H}}}(X_i)$.

  2. For each $i \in \{1, ..., p - d_1\}$, calculate $\widehat{\mathrm{CMC}_{\mathcal{H}}}(X_{c_i}) := \widehat{\mathrm{CMC}_{\mathcal{H}}}(Y, X_{c_i}^{\perp} | \boldsymbol{X}_S)$, where each $c_i$ is from $S^c$ and $X_{c_i}^{\perp} = X_{c_i} - P_{\boldsymbol{X}_S} X_{c_i}$ with $P_{\boldsymbol{X}_S}$ being the projection matrix onto the column space of $\boldsymbol{X}_S$.

  3. Calculate the score $A_i := \max\left(\frac{\mathrm{MC}_{\mathcal{H}}(X_{c_i})}{\max\limits_{1 \le i \le p-d_1} \mathrm{MC}_{\mathcal{H}}(X_{c_i})}, \frac{\mathrm{CMC}_{\mathcal{H}}(X_{c_i})}{\max\limits_{1 \le i \le p-d_1} \mathrm{CMC}_{\mathcal{H}}(X_{c_i})}\right)$ for each $i = 1, ..., p - d_1$. Keep the top $d_2 - d_1$ predictors with the largest scores. If $d_2$ is not given, $d_2 = \lfloor n/\log n \rfloor$.

Output: The index set of the $d_2$ selected variables $\{i_1, ..., i_{d_2}\} \subseteq \{1, ..., p\}$ (the $d_1$ variables selected in Step 1 plus the $d_2 - d_1$ variables selected in Step 3).
```

The R code of S-CMC$_{\mathcal{H}}$ is available in the supplement file.

Remark 9. *The parameters d_1 and d_2 are predefined values. In general, larger d_1 will lead to worse performance if the set $\boldsymbol{X}_S$ contains larger proportion of inactive variables. It will not affect the theoretical performance of* CMC$_{\mathcal{H}}$. *However, computationally, the estimation of the expected kernel function of long vectors inside* CMC$_{\mathcal{H}}$ *becomes less reliable if the sample size is limited. Thus we recommend to choose small d_1(e.g., $d_1 = \left\lfloor \sqrt{n/\log n} \right\rfloor$) in practice.*

Remark 10. *The adoption of $X_{c_i}^{\perp}$ is to handle the scenario when the independence assumption of $\boldsymbol{X}_S$ and X_i is violated. Note that this only removes the linear dependence. See more discussion in Remark 4.*

Remark 11. *In the last step, an alternative way is to use a linear combination $w_1 \frac{\mathrm{MC}_{\mathcal{H}}(X_{c_i})}{\max\limits_{1 \le i \le p-d_1} \mathrm{MC}_{\mathcal{H}}(X_{c_i})} + w_2 \frac{\mathrm{CMC}_{\mathcal{H}}(X_{c_i})}{\max\limits_{1 \le i \le p-d_1} \mathrm{CMC}_{\mathcal{H}}(X_{c_i})}$ to rank all the predictors. The weights can be adaptively chosen driven by the data, which we leave as a potential future work.*

2.3.3 Extension to Quantile Screening

In this section, we extend our method to the quantile screening setting. For a univariate random response Y, denote $Y_\tau = \tau - \mathbf{1}(Y \le q_\tau)$ as its binary version with $\tau \in (0, 1)$, where q_τ is the τ-th quantile of the distribution of Y. Given *i.i.d.* observations $\{y_k\}_{k=1}^n$ of Y, denote $\hat{Y}_\tau = \tau - \mathbf{1}(Y \le \hat{q}_\tau)$ as the estimate of Y_τ, where $\hat{q}_\tau$ is the sample τ-th quantile. So for each observation y_k, we denote $y_{k_\tau} = \tau - \mathbf{1}(y_k \le \hat{q}_\tau)$.

Let $\omega_j(Y_\tau) = \mathrm{CMC}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$ and $\widehat{\omega}_j(\hat{Y}_\tau) = \widehat{\mathrm{CMC}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$. Similarly, we denote $\mathcal{D}_{q_\tau} = \{j : \mathbb{E}(Y_\tau|(\boldsymbol{X}_S, X_j))$ depends on $X_j\}$ as the quantile active predictors conditional on $\boldsymbol{X}_S$, and denote $\widehat{\mathcal{D}}_{q_\tau} = \{j : \widehat{\omega}_j(\hat{Y}_\tau) \geq cn^{-\kappa}$, for $j \in S^c\}$ as the selected variables.

Next we show the sure screening property for the quantile version of $\mathrm{CMC}_{\mathcal{H}}$.

Theorem 5. *Under condition (C1) in the appendix, for any $0 < \gamma < 1/2 - \kappa$ and $\kappa \in (0, 1/2)$, there exist positive constants c_1, c_2 such that for any $c > 0$,*

$$P\left\{\max_{1\leq j\leq p-d_1} |\widehat{\omega}_j(\hat{Y}_\tau) - \omega_j(Y_\tau)| \geq cn^{-\kappa}\right\} \leq O((p-d_1)[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]). \tag{2.8}$$

If Condition (C2) in the appendix holds in addition, we have

$$P(\mathcal{D}_{q_\tau} \subseteq \hat{\mathcal{D}}_{q_\tau}) \geq 1 - O(\widetilde{s}_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]), \tag{2.9}$$

where $\widetilde{s}_n$ is the cardinality of $\mathcal{D}_{q_\tau}$.

2.4 Numerical Studies

In this section, we evaluate the finite-sample performance of the proposed method $\mathrm{CMC}_{\mathcal{H}}$.

The choice of kernel functions Denote the translation-invariant Gaussian kernel as

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) := \exp(-\frac{1}{h}(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)), \tag{2.10}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^t, t \in \mathbb{N}$, and h is the bandwidth. For the proposed $\mathrm{CMC}_{\mathcal{H}}$, we use $K(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for both k_1 and k_2. In our simulations, the performance of $\mathrm{CMC}_{\mathcal{H}}$ for variable screening is robust against the bandwidths of k_1 and k_2. So we set $h = 2$ for both k_1 and k_2. For $\mathrm{MC}_{\mathcal{H}}$, we adopt Gaussian kernel and conduct a sensitivity analysis of the bandwidth, the results of which are presented in the supplementary material S1.1. The performance of $\mathrm{MC}_{\mathcal{H}}$ is sensitive to the bandwidth h. In particular, $\mathrm{MC}_{\mathcal{H}}$ with smaller h performs better for selecting covariates that are linearly related to the response variable, while larger h is more suitable for selecting nonlinearly related covariates. To select the conditional set $\boldsymbol{X}_S$ and avoid cherry-picking, we first calculate the values of $\mathrm{MC}_{\mathcal{H}}$ using two bandwidths $h = 2\hat{\sigma}^2_{X_i}$ and $h = 6\hat{\sigma}^2_{X_i}$ for each predictor X_i, where $\hat{\sigma}^2_{X_i}$ is sample variance of X_i and $i = 1, ..., p$. Our experience is that $h = 2\hat{\sigma}^2_{X_i}$ and $h = 6\hat{\sigma}^2_{X_i}$ generally perform well for the simulations. Then we take the maximum of the two $\mathrm{MC}_{\mathcal{H}}$ values for each predictor. One can also use Laplacian kernel and Cauchy kernel in practice. However, in our examples, they yield similar performance to that of the Gaussian kernel.

Criteria of evaluating variable screening performance Following [20], we consider three criteria for evaluating the variable screening performance: 1) $\mathcal{S}_q$: the $(100q)$-th quantile of the minimal model size required to contain all the active predictors. 2) $\mathcal{P}_i$: the proportion that the predictor X_i is selected, and 3) $\mathcal{P}_{\text{all}}$: the proportion of all active predictors being selected. Essentially, an $\mathcal{S}_q$ closer to the total number of active predictors is preferred. The three criteria are connected in a way that a smaller minimal model size suggests a larger $\mathcal{P}_{all}$ and a lager $\mathcal{P}_i$ for each active variable.

Screening thresholds We compare S-CMC$_{\mathcal{H}}$ with six variable screening methods, including two marginal screening method (DCSIS2 in [17] and MDC), three conditional screening methods (CSIS in [2], CDCSIS in [40] and CIS in [35]), and one iterative methods (RaSE$_1$-eBIC in [34]. For each screening method, we keep the top $d_2 = \lfloor n/log(n) \rfloor$ variables, where n is the sample size. We report the $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ values based on 100 repetitions for each example. Since conditional screening methods require a pre-selected conditional set $\boldsymbol{X}_S$, we either artificially set up the conditional set based on the variables in the true model or select the top $d_1 = \lfloor \sqrt{n/log(n)} \rfloor$ variables suggested by MC$_{\mathcal{H}}$. We also include a sensitivity analysis of using different methods (e.g. SIS, LASSO and forward regression) to choose the conditional set in the supplementary material S1.2.

2.4.1 Simulation

Example 1 (Marginally inactive but jointly active predictors). Following the idea of [9], we generate samples $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ from the linear regression model

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 - cX_6 + \epsilon,$$

where the coefficient c is designed so that $\text{cov}(X_6, Y) = 0$. That is, the predictor X_6 is marginally independent from the response Y. The predictor vector $\boldsymbol{X} = (X_1, ..., X_p) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has (i,j)-th entry $\sigma_{ij} = \rho^{I\{i \neq j\}}$. The error term $\epsilon \sim N(0,1)$ and is independent from $\boldsymbol{X}$. We set the sample size $n = 200$ and the dimension $p = 3000$. We consider three cases: $(c, \rho) \in \{(2.5, 0.5), (4, 0.8), (4.5, 0.9)\}$. Note that X_6 is dependent with Y if given one or more predictors from $\{X_1, ..., X_5\}$.

The simulation results are reported in Table 2.1. The marginal screening method MDC fails to detect the marginally independent predictor X_6 in all cases, while other methods identify X_6 as an active predictor. As the correlation increases from $\rho = 0.5$ to $\rho = 0.9$, the selection proportion $\mathcal{P}_i$ decreases for $i = 1, 2, ..., 5$. Consequently, $\mathcal{P}_{all}$ decreases as the correlation increases. Note that if the conditional set $\boldsymbol{X}_{S_1} = \{X_1\}$, the proportion of selecting X_1 is set to 1. We set $d_2 = \lfloor n/log(n) \rfloor = 37$ in this example. Note that the minimal model size $\mathcal{S}_{0.5}$ is small only in the first case where $\rho = 0.5$, which explains the low values of $\mathcal{P}_{all}$ for all the methods for the high correlation case ($\rho = 0.8$ or 0.9). The three conditional methods: CIS, CSIS and CDC-SIS, fail to detect the active variables X_2, X_3, X_4 and X_5 when the correlation increases to $\rho = 0.9$.

Table 2.1: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 1.

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_6$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
$c = 2.5$, $\rho = 0.5$								
MDC	0.89	0.94	0.93	0.91	0.90	0.00	0.00	3000.0
CSIS ($\boldsymbol{X}_{S_1}$)	1.00	0.97	0.98	0.98	0.97	0.87	0.78	14.0
CSIS ($\boldsymbol{X}_{S_2}$)	0.71	0.75	0.75	0.72	0.72	1.00	0.17	1580.5
CDC-SIS ($\boldsymbol{X}_{S_1}$)	1.00	0.95	0.95	0.95	0.93	0.16	0.14	708.5
CDC-SIS ($\boldsymbol{X}_{S_2}$)	0.73	0.76	0.74	0.72	0.72	1.00	0.16	1541.5
CIS($\boldsymbol{X}_{S_1}$)	1.00	0.83	0.88	0.87	0.87	0.71	0.34	58.0
CIS($\boldsymbol{X}_{S_2}$)	0.97	0.91	0.94	0.94	0.97	0.01	0.01	2997.5
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_1}$)	1.00	0.94	0.94	0.92	0.92	0.53	0.39	90.5
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_2}$)	0.91	0.95	0.94	0.93	0.92	1.00	0.72	17.5
RaSE$_1$-eBIC	0.99	1.00	0.99	0.99	1.00	1.00	0.97	6.0
$c = 4$, $\rho = 0.8$								
MDC	0.61	0.63	0.64	0.63	0.62	0.00	0.00	3000.0
CSIS ($\boldsymbol{X}_{S_1}$)	1.00	0.48	0.44	0.40	0.44	1.00	0.10	320.0
CSIS ($\boldsymbol{X}_{S_2}$)	0.39	0.34	0.36	0.41	0.33	1.00	0.00	2988.5
CDC-SIS ($\boldsymbol{X}_{S_1}$)	1.00	0.56	0.53	0.54	0.54	0.93	0.06	264.0
CDC-SIS ($\boldsymbol{X}_{S_2}$)	0.39	0.34	0.36	0.41	0.34	1.00	0.00	2623.5
CIS($\boldsymbol{X}_{S_1}$)	1.00	0.39	0.31	0.34	0.42	1.00	0.03	546.0
CIS($\boldsymbol{X}_{S_2}$)	0.70	0.59	0.67	0.64	0.69	1.00	0.09	268.0
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_1}$)	1.00	0.73	0.71	0.66	0.67	1.00	0.26	127.5
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_2}$)	0.61	0.73	0.72	0.66	0.67	1.00	0.16	156.5
RaSE$_1$-eBIC	0.80	0.92	0.76	0.79	0.87	1.00	0.35	2312.5
$c = 4.5$, $\rho = 0.9$								
MDC	0.48	0.40	0.39	0.48	0.40	0.00	0.00	3000.0
CSIS ($\boldsymbol{X}_{S_1}$)	1.00	0.14	0.09	0.13	0.08	1.00	0.00	2339.0
CSIS ($\boldsymbol{X}_{S_2}$)	0.28	0.24	0.23	0.24	0.20	1.00	0.00	2992.5
CDC-SIS ($\boldsymbol{X}_{S_1}$)	1.00	0.29	0.28	0.29	0.28	0.98	0.01	875.0
CDC-SIS ($\boldsymbol{X}_{S_2}$)	0.28	0.24	0.23	0.24	0.21	1.00	0.00	2875.0
CIS($\boldsymbol{X}_{S_1}$)	1.00	0.11	0.08	0.07	0.16	1.00	0.00	2037.5
CIS($\boldsymbol{X}_{S_2}$)	0.43	0.44	0.47	0.46	0.41	1.00	0.00	897.0
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_1}$)	1.00	0.48	0.47	0.51	0.45	1.00	0.07	269.5
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_2}$)	0.53	0.48	0.47	0.51	0.47	1.00	0.03	338.0
RaSE$_1$-eBIC	0.62	0.89	0.73	0.65	0.68	1.00	0.19	2316.0

The conditional set is either $\boldsymbol{X}_{S_1} = \{X_1\}$ or $\boldsymbol{X}_{S_2} = \{$the first $d_1 = 6$ predictors selected by MDC$_{\mathcal{H}}\}$.

When the conditional set changes from an oracle set $\boldsymbol{X}_{S_1}$ to a data-dependent set $\boldsymbol{X}_{S_2}$, the performances of the three conditional screening methods (CIS, CSIS and CDC-SIS) deteriorate. Our proposed S-CMC$_{\mathcal{H}}$ instead shows robustness against the conditional set. The performance of our method decreases due to the extra cost of selecting X_1 when the conditional set becomes data-dependent.

In this example, from the perspective of the minimal model size $\mathcal{S}_{0.5}$, the minimal model size for RaSE$_1$-eBIC changes from 6 to 2316 when ρ increases from 0.5 to 0.9. In contrast, the stable performance of S-CMC$_{\mathcal{H}}$ in $\mathcal{S}_{0.5}$ indicates that our method is more robust against the correlation ρ. In terms of $\mathcal{P}_{all}$ and $\mathcal{P}_i$, RaSE$_1$-eBIC and S-CMC$_{\mathcal{H}}$ are better than any other method. RaSE$_1$-eBIC is better than S-CMC$_{\mathcal{H}}$, which may be due to the good performance of RaSE$_1$-eBIC in the linear case.

Example 2 (Interaction terms). We consider the following model with interaction terms:

$$Y = X_1 + X_5 + X_{10} + X_1X_{15} + 1.5X_5X_{20} + 2X_{10}X_{25} + \epsilon.$$

The predictor vector $\boldsymbol{X} = (X_1, ..., X_p) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has (i, j)-th entry $\sigma_{ij} = \rho^{|i-j|}$. We consider two cases: $\rho \in \{0, 0.9\}$. The error term $\epsilon \sim N(0, 1)$ and is

Table 2.2: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2.

	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
$\rho = 0$								
MDC	0.94	0.96	0.90	0.01	0.00	0.03	0.00	2662.5
CSIS ($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.04	0.03	0.07	0.00	2452.0
CSIS($\boldsymbol{X}_{S_2}$)	0.97	0.93	0.90	0.03	0.02	0.04	0.00	2147.0
CDC-SIS ($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.01	0.03	0.26	0.00	1639.0
CDC-SIS ($\boldsymbol{X}_{S_2}$)	0.84	0.86	0.84	0.07	0.17	0.43	0.00	1529.0
DCSIS2	0.40	0.70	0.99	0.11	0.35	0.85	0.00	1372.5
CIS($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.02	0.12	0.45	0.00	566.0
CIS($\boldsymbol{X}_{S_2}$)	0.99	0.98	0.96	0.01	0.01	0.02	0.00	2278.0
S-CMC$_{\mathcal{H}}$ ($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.18	0.62	0.99	0.09	196.5
S-CMC$_{\mathcal{H}}$($\boldsymbol{X}_{S_2}$)	0.94	0.96	0.91	0.11	0.37	0.77	0.01	851.5
RaSE$_1$-eBIC	0.87	0.82	0.79	0.01	0.01	0.04	0.00	2295.5
$\rho = 0.9$								
MDC	1.00	1.00	1.00	0.97	0.52	0.14	0.13	350.5
CSIS ($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.08	0.15	0.13	0.02	1887.0
CSIS ($\boldsymbol{X}_{S_2}$)	0.78	0.97	0.74	0.08	0.12	0.13	0.02	1870.0
CDCSIS ($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.03	0.04	0.06	0.00	2024.5
CDCSIS($\boldsymbol{X}_{S_2}$)	0.52	0.96	0.80	0.86	1.00	1.00	0.40	66.0
DCSIS2	0.93	1.00	1.00	0.87	0.91	0.81	0.62	30.5
CIS($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.28	0.93	1.00	0.27	62.0
CIS($\boldsymbol{X}_{S_2}$)	0.78	0.99	0.83	0.43	0.92	0.85	0.2	88.5
S-CMC$_{\mathcal{H}}$($\boldsymbol{X}_{S_1}$)	1.00	1.00	1.00	0.90	1.00	1.00	0.90	25.0
S-CMC$_{\mathcal{H}}$($\boldsymbol{X}_{S_2}$)	1.00	1.00	1.00	0.90	1.00	0.99	0.89	26.0
RaSE$_1$-eBIC	0.41	0.42	0.37	0.02	0.02	0.01	0.01	2403.0

We set the conditional set to be $\boldsymbol{X}_{S_1}$=$\{X_1, X_5, X_{10}\}$ or $\boldsymbol{X}_{S_2}$ =$\{$the first $d_1 = 6$ predictors selected by MC$_{\mathcal{H}}\}$.

independent from $\boldsymbol{X}$. We set the sample size $n = 200$ and the dimension $p = 3000$.

The results are reported in Table 2.2. The three variables X_{15}, X_{20}, X_{25} all jointly contribute to the mean of Y, but are marginally independent of the mean of Y. It is difficult for the marginal screening methods to detect these three terms. Note that X_{25} has a larger coefficient in its interaction term than that of X_{20} or X_{15}. As the signal/coefficient of the interaction term increases, its effect is easier to be detected ($\mathcal{P}_{25} > \mathcal{P}_{20} > \mathcal{P}_{15}$) for all the methods. RaSE$_1$-eBIC fails in detecting those three variables in this example. This is possibly due to the fact that RaSE$_1$-eBIC targets additive models. In comparison to the interaction screening method DCSIS2, the proposed method S-CMC$_{\mathcal{H}}$ has a comparable performance in selecting X_{15}, X_{20} and X_{25} for both $\rho = 0$ and $\rho = 0.9$. But S-CMC$_{\mathcal{H}}$ performs better than DCSIS2 in selecting the marginally active variables X_1, X_5 and X_{10}. What's more, S-CMC$_{\mathcal{H}}$ has the smallest $\mathcal{S}_{0.5}$ among all the methods. Similar to Example 1, the performance of S-CMC$_{\mathcal{H}}$ is stable against the conditional set. We also did a sensitivity analysis against the conditional set. We consider three more cases: the conditional set is selected by Lasso, SIS, and forward regression. The results are reported in the supplementary material S1.2, demonstrating a better and more stable performance of S-CMC$_{\mathcal{H}}$ compared to other conditional methods (CSIS, CDC-SIS). Finally, we include a block structure correlation setting where the correlation among active predictors are 0.2 and 0.1 otherwise. We report the result in the supplementary material S1.3, which clearly shows the advantage of S-CMC$_{\mathcal{H}}$.

Example 3 (Heteroscedasticity & Quantile screening). In this example, we demonstrate that our screening method can help in heteroskedastic model specification. We consider the following model:

$$Y = X_1 + X_5 + X_1 X_{10} + 1.5 X_5 X_{15} + \epsilon \cdot \exp(X_{35} + X_{40}).$$

The predictor vector $\boldsymbol{X} = (X_1, ..., X_p) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has (i,j)-th entry $\sigma_{ij} = \rho^{|i-j|}$. We consider two cases: $\rho \in \{0, 0.9\}$. The error term $\epsilon \sim N(0,1)$ is independent from $\boldsymbol{X}$. We set the sample size $n = 400$ and the dimension $p = 3000$. For the purpose of quantile screening, we change the continuous response Y to a binary response $Y_\tau = \tau - \mathbf{1}(Y \leq \hat{q}_\tau)$, where $\hat{q}_\tau$ is the τ-th sample quantile of the response. Then we apply S-CMC$_{\mathcal{H}}$ and MDC on the data $(Y_\tau, \boldsymbol{X})$. Note that in fixed design, the population quantile q_τ does not depend on (X_{20}, X_{25}) if and only if $\tau = 0.5$. We consider two choices of the quantile: $\tau = 0.5$ and $\tau = 0.75$.

The results are presented in Table 2.3 with conditional set selected by MC$_{\mathcal{H}}$. In this example, we also evaluate the method QaSIS [16], a quantile-adaptive model-free variable screening method for heterogeneous data. Overall, our proposed method S-CMC$_{\mathcal{H}}$ performs the best across all four combinations of (ρ, τ). In particular, S-CMC$_{\mathcal{H}}$ is dominantly better in $\mathcal{S}_{0.5}$ than any other method even under high correlation setting. When there is no correlation among predictors ($\rho = 0$), from the perspective of quantile screening, MDC, CDC-SIS, QaSIS, and S-CMC$_{\mathcal{H}}$ correctly differentiate the different roles of X_{35} and X_{40} under two values of τ. However, these three methods (MDC, CDC-SIS, QaSIS) fail to identify X_{15} (compared to X_{10}) when $\tau = 0.5$. In contrast, S-CMC$_{\mathcal{H}}$ has a better performance in separating the active variables from the inactive variables. When high correlation exists among predictors, all methods receive improved performance and S-CMC$_{\mathcal{H}}$ still remains competitive. This is because the marginal relationship between X_{10} (X_{15}) and the response Y is strengthened by the high correlation among predictors. We also include the scenario when conditional set is $\{X_1, X_5, X_{10}\}$ in the supplementary material S1.4.

Example 4 (Nonlinear case). We consider the following nonlinear case.

$$Y = 3I(X_1 > 0.5)X_2 + 3sin^2(2\pi X_1)X_3 + 3(X_1^2 - 1)X_4 + \exp(X_1)X_5 + \epsilon,$$

where we first generate U_1, $U_2 \overset{i.i.d}{\sim}$ Unif$[0,1]$ and then let $X_1 = (U_1 + U_2)/2$ and $X_k = (Z_k + 2U_1)/4$ for $k = 2, 3, ..., p$. We consider two scenarios, symmetric distribution $Z_k \overset{i.i.d}{\sim} N(0,1)$ and asymmetric distribution $Z_k \overset{i.i.d}{\sim} \chi^2_{(1)}$. The error term ϵ is independently drawn from standard normal distribution. We set the sample size $n = 200$ and the dimension $p = 3000$. The results are reported in Table 2.4.

In this example, the active predictors are nonlinearly associated with each other. Each element in this model is an interaction effect of two variables. The proposed S-CMC$_{\mathcal{H}}$ demonstrates its competitive performance for selecting all the active predictors

Table 2.3: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 3 with $\boldsymbol{X}_S$ selected by MC$_\mathcal{H}$.

Method	τ	$\rho = 0$								$\rho = 0.9$							
		$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{35}$	$\mathcal{P}_{40}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{35}$	$\mathcal{P}_{40}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
MDC	0.5	1.00	1.00	0.06	0.03	0.04	0.02	0.00	1910.0	1.00	1.00	0.97	0.49	0.06	0.05	0.49	71.0
	0.75	0.99	1.00	0.24	0.38	0.52	0.55	0.03	986.5	1.00	1.00	1.00	1.00	0.91	0.92	0.86	29.5
CSIS	0.5	1.00	0.99	0.01	0.03	0.20	0.16	0.00	2025.5	1.00	1.00	0.19	0.07	0.31	0.24	0.00	1854.5
	0.75	0.97	0.95	0.06	0.14	0.39	0.42	0.00	2207.0	0.86	0.96	0.39	0.11	0.31	0.30	0.00	2138.5
CDC-SIS	0.5	1.00	0.99	0.13	0.28	0.23	0.30	0.04	757.0	1.00	1.00	0.28	0.94	1.00	1.00	0.27	172.5
	0.75	0.96	0.97	0.13	0.51	0.42	0.47	0.04	1159.5	0.88	0.97	0.41	0.87	0.95	0.92	0.25	214.5
QaSIS	0.5	1.00	1.00	0.16	0.16	0.28	0.27	0.03	1085.0	1.00	1.00	1.00	0.90	0.69	0.55	0.90	20.5
	0.75	0.92	0.99	0.14	0.38	0.69	0.72	0.03	600.0	0.99	1.00	1.00	0.97	0.99	0.97	0.92	35.5
DCSIS2	0.5	0.18	0.41	0.04	0.08	0.93	0.90	0.00	1443.5	0.07	0.15	0.06	0.05	1.00	1.00	0.00	918.0
	0.75	0.18	0.41	0.04	0.08	0.93	0.90	0.00	1443.5	0.07	0.15	0.06	0.05	1.00	1.00	0.00	918.0
CIS	0.5	1.00	0.99	0.07	0.04	0.01	0.02	0.01	1982.5	1.00	1.00	0.31	0.89	0.95	0.98	0.29	121.5
	0.75	1.00	0.99	0.06	0.14	0.21	0.3	0.00	2537.5	1.00	0.99	0.47	0.76	0.9	0.91	0.33	129.0
S-CMC$_\mathcal{H}$	0.5	1.00	1.00	0.18	0.80	0.05	0.04	0.13	240.0	1.00	1.00	0.97	0.93	0.34	0.38	0.90	16.0
	0.75	0.99	1.00	0.33	0.67	0.49	0.58	0.05	574.5	1.00	1.00	1.00	0.95	0.97	0.93	0.86	28.0
RaSE$_1$-eBIC	0.5	0.49	0.31	0.00	0.00	0.03	0.01	0.00	1816.0	0.25	0.07	0.00	0.00	0.04	0.03	0.00	2425.5
	0.75	0.49	0.31	0.00	0.00	0.00	0.00	0.00	2427.0	0.25	0.07	0.00	0.00	0.01	0.02	0.00	2427.0

Table 2.4: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 4 with $\boldsymbol{X}_S$ selected by MC$_\mathcal{H}$.

	$Z_k \overset{i.i.d}{\sim} N(0,1)$							$Z_k \overset{i.i.d}{\sim} \chi^2_{(1)}$						
	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
MDC	1.00	0.79	0.84	0.00	0.85	0.00	2974.5	1.00	0.64	0.80	0.00	0.85	0.00	2725.0
CSIS	0.99	0.56	0.55	1.00	0.69	0.17	554.5	1.00	0.61	0.70	1.00	0.81	0.26	100.5
CDC-SIS	0.99	0.62	0.53	0.34	0.68	0.11	1101.5	1.00	0.45	0.60	0.65	0.60	0.11	962.5
DCSIS2	1.00	0.84	0.57	0.00	0.52	0.00	2235.0	1.00	0.90	0.71	0.05	0.70	0.01	1284.5
CIS	1.00	0.95	0.90	0.31	0.94	0.24	153.0	1.00	0.76	0.78	0.42	0.78	0.17	155.0
S-CMC$_\mathcal{H}$	1.00	0.79	0.84	0.95	0.85	0.52	33.5	1.00	0.63	0.78	0.90	0.82	0.35	77.0
RaSE$_1$-eBIC	0.28	0.60	0.60	0.58	0.74	0.06	2250.0	0.98	0.47	0.63	0.95	0.73	0.18	2253.5

with relatively high $\mathcal{P}_i$ and $\mathcal{P}_{all}$, and smallest $\mathcal{S}_{0.5}$, in both scenarios. The methods DCSIS2 and MDC fail to detect X_4. It is worth pointing out that the contribution of variable X_4 is also underestimated by the methods CDC-SIS and CIS. This may be because that X_4's conditional contribution to Y is diffused by the exponential term $\exp(X_1)X_5$, as indicated by the much larger $\mathcal{P}_5$ in the four methods mentioned above.

2.4.2 Data Applications

2.4.2.1 Single Cell Malt Tumor CITE-seq Dataset

The Malt Tumor **C**ellular **I**ndexing of **T**ranscriptomes and **E**pitopes by sequencing (CITE-seq) dataset (*http://www.10xgenomics.com*) contains single-cell level sequencing RNA data and as well as the surface protein expression count. We are interested in identifying the genes that affect the surface protein level. The dataset contains 33555 genes and proteins from 8412 single cells. Following the data pre-processing procedure in [35] (filtering out cells with more than 90% zero entries and genes that has zero variance), we obtain a sample of $n = 207$ single cells and $p = 18702$ genes. And we set protein CD8 as the response variable and the protein CD3 as conditional variable. Interested readers are referred to [35] for detailed scientific explanations of using CD3 as the conditional variable. To evaluate the prediction performance of each screening method, we randomly split the observations into a training set of size 176 and a test set of size 31. We select $d_2 = 38$ variables by the following four conditional

Table 2.5: The prediction accuracy in MALT data.

Method	MSE	paired t-test p-value
CSIS	1.364	$< 2.2 \times 10^{-16}$
CDC-SIS	1.162	$< 2.2 \times 10^{-16}$
CIS	1.268	$< 2.2 \times 10^{-16}$
S-CMC$_{\mathcal{H}}$	0.748	–

Table 2.6: The prediction accuracy in Riboflavin production data.

Method	MSE	paired t-test p-value
MDC	0.301	0.017
CSIS	0.357	2.909×10^{-15}
CDC-SIS	0.323	5.62×10^{-5}
CIS	0.323	5.62×10^{-5}
RaSE$_1$-eBIC	0.293	0.401
S-CMC$_{\mathcal{H}}$	0.292	–

The conditional set $\boldsymbol{X}_S$ contains the top $d_1 = 4$ variables suggested by MC$_{\mathcal{H}}$.

methods: CSIS, CDCSIS, CIS and SCMC$_{\mathcal{H}}$ including CD3. Then a random forest model is fitted with the selected variables and the response, with log transformations on the response and the variables. The mean squared errors (MSE) of the methods on the test set for 100 repetitions are reported in Table 2.5. In each of the 100 repetitions, we calculate the difference between the MSE of S-CMC$_{\mathcal{H}}$ and that of each competing method. Then we conducted a one-sided paired two-sample t-test with the alternative hypothesis that our method S-CMC$_{\mathcal{H}}$ has a smaller average mean squared error (MSE). Our method has the smallest MSE and the p-values of the paired t-test indicates S-CMC$_{\mathcal{H}}$ outperforms each method in prediction accuracy.

2.4.2.2 Riboflavin Production Dataset

The dataset [18] contains information about riboflavin (vitamin B2) production by $n = 71$ bacillus subtiliswith, where $p = 4088$ gene expression levels are recorded. The dataset is provided by Royal DSM (Switzerland) and is available in the R package `hdi`.

Our goal is to find which genes are most related in predicting the riboflavin production rate. We randomly split the sample into a training set of size 60 and a test set of size 11. To evaluate the prediction performance of each screening method, we select $d_2 = 16$ variables for each screening method and train a random forest model on the training set. Then we calculate the mean squared error (MSE) of each method on the test set. The average MSE's based on 100 data splittings into training/test sets are reported in Table 2.6. Similar to Section 2.4.2.1, we conducted the same one-sided paired two-sample t-test with the alternative hypothesis that our method S-CMC$_{\mathcal{H}}$ has a smaller averge mean squared error (MSE).Our method has the smallest MSE and the p-values suggest that it significantly improves the MSE.

2.5 Conclusion

In this article, we propose the **c**onditional **m**artingale difference **d**ivergence ($\mathrm{CMD}_{\mathcal{H}}$) to measure the dependence between a response variable and a predictor vector given a third vector. It is primarily designed to overcome the limitation of marginal independence measures. Based on $\mathrm{CMD}_{\mathcal{H}}$, we develop a new screening procedure called S-$\mathrm{CMC}_{\mathcal{H}}$ by combining the merits of the $\mathrm{CMC}_{\mathcal{H}}$ and $\mathrm{MC}_{\mathcal{H}}$ for selecting both marginal and jointly active variables. The proposed framework can be easily extended to quantile screening. The simulations and real data applications demonstrate that S-$\mathrm{CMC}_{\mathcal{H}}$ has a competitive and stable performance under variety model settings for mean or quantile screening. We also provide a data-driven method for selecting the conditional set $\boldsymbol{X}_S$. The limitation of this method is that we do need to predetermine a proper number of variables in conditional set to get a satisfactory performance. Using $\lfloor \sqrt{n/\log n} \rfloor$, as done in our numerical study, may not suffice for the cases when the true underlying model consists jointly only active variables that depends on a large conditional set. Designing a variable screening method that is free of tuning the cardinality of the conditional set is a challenging future research topic.

Chapter 3 High dimensional inference in integration of multiview microbiome data for causal discoveries of complex trait-metabolite associations

3.1 Introduction

The human microbiome refers to the collection of microorganisms, including bacteria, viruses, fungi, and other microbes, that reside within and on our bodies. These microbes play a crucial role in maintaining our health and well-being by aiding in digestion, immune system regulation, and synthesizing essential nutrients.

However, an imbalance or disruption in the microbiome can lead to various microbiome-related diseases. These conditions arise when the composition and diversity of the microbial community are altered, resulting in negative health effects. Examples of such disease including inflammatory bowel disease (IBD), obesity and metabolic disorders, allergies and asthma, mental health disorders, and clostridium difficile infection. Understanding the complex interactions between the microbiome and various diseases is critical. It would provide promising directions for developing targeted interventions, such as probiotics, prebiotics, and fecal microbiota transplantation, to restore a healthy microbiome and improve patient outcomes. The primary mode that microbiome interact with host is through the microbiome metabolism, producing the corresponding metabolites. It is naturally to suspect that microbiome metabolism plays an important role during the pathogenesis of those diseases.

We demonstrate such hypothesis is reasonable with a real data example: iHMP-IBDMDB-2019 (IHMP) dataset in [22], which comes from a longitudinal IBD study consists of 79 patients and 26 controls. The microbiome abundance measurement is from shotgun and metabolites is measured from the untargeted and four complimentary LC-MS methods, see [22] for a detailed description of data collection. We only consider subjects' baseline measurement obtained at the time of their enrollment. The detailed data processing are described in Section3.4.

Our response is the IBD biomarker: C-Reactive Protein (CRP) level in mg/L, which is a protein that human liver makes. Elevated levels of CRP are observed in the bloodstream during inflammation, which naturally makes it a good indicator for IBD. See [37] for a detailed illustration and justification. Indeed, the exclusive availability of CRP level measurement is the reason we choose this iHMP dataset. By a simple linear regression on each metabolite, we obtain their marginal coefficient and p-value and make a volcano plot in Figure 3.1. Despite the limited sample size of 54, we have discovered 43 significant metabolites (annotated in blue dots).

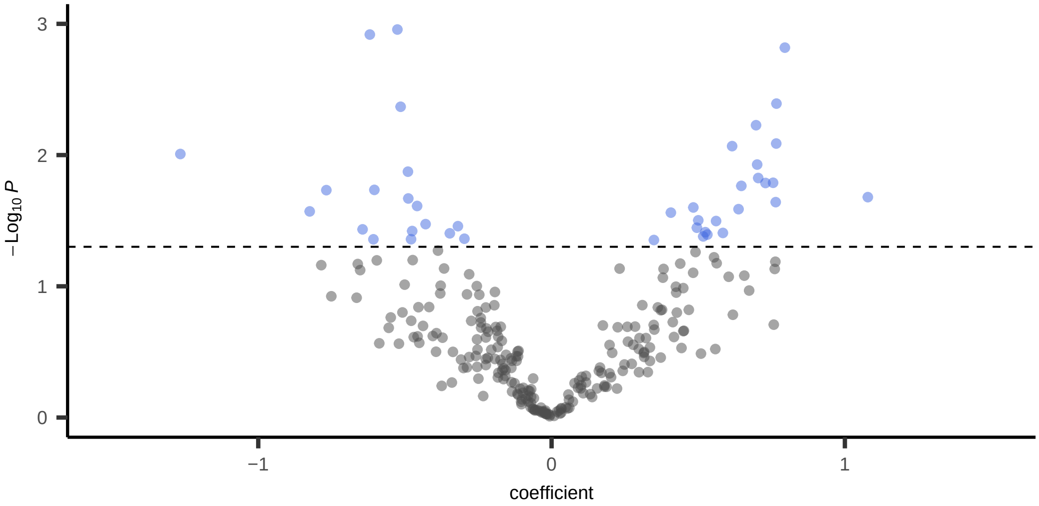

Figure 3.1: volcano plot of each metabolites by univariate linear model

We then want to test whether these 43 significant metabolites are actually associated with microbiome. From Microbiome Regression-Based Kernel Association Tests (MiRKAT), which is a kernel based global test in [4], 28 metabolites shows significance.

Now, we have presented a reasonable hypothesized pathogenesis for IBD: microbiome → metabolites → IBD(CRP). Our ultimate goal here is to understand how does microbiome affect such disease. To be more specific, we want to test the hypothesis that microbiome has an **causal** affect on IBD through the microbiome metabolism (or metabolites). To this end, it is not sufficient to draw conclusions from marginal analysis on metabolites, as such discovery could be false positive. It is necessary to include both metabolites and microbiome abundance in the analysis.

Fortunately, we have such data available called multi-view microbiome data. Each observation contains the high dimensional metabolites and microbiome abundance data. Indeed, the aforementioned IHMP dataset is one such data. However, there are two challenges. The first one is that such datasets usually have small sample size as metabolites measurement is expensive. Small sample size is problematic because of weak power for the test, especially in a high dimensional setting. The second difficulty is the availability a particular microbiome generated metabolites of interest or phenotype response. For example, some metabolites are not measured in IHMP dataset but it is found in another IBD dataset from [13]. Another situation is that, while large microbiome-wide association data are available, they lack metabolites measurement. Can we also integrate it with multi-view mcirobiome data and potentially improve the power?

To address these issues, we develop a general framework to integrate microbiome datasets, as well as metabolites and microbiome. Specifically, with a structural model in section 3.2.1 , we are able to test the causal effect of the metabolite. Our proposed method is very natural: we can impute the unobserved metabolite from another informative multi-view microbiome dataset. And we show that the resulting inference remains reliable both theoretically and numerically. One particular appealing advantage of our method is that the type I error is well controlled regardless the choice of external multi-view microbiome dataset.

The rest of the article is organized as follows: in Section 3.2.1, we introduce our proposed structural model and describe the inference procedure in detail. In Section 3.3, we demonstrate the performance of our method with simulation study and a real data application. The concluding discussion and future research area in Section 3.5.

3.2 Methodology

3.2.1 Model

We consider the following linear structral equation model for our target *i.i.d.* data $\{m_i, X_{i,\cdot}, Y_i\}$:

$$Y_i = M_i\theta^* + \boldsymbol{\beta}^{**\top} X_{i,\cdot} + \varepsilon_i \tag{3.1}$$

and

$$M_i = \boldsymbol{\gamma}^\top X_{i,\cdot} + \delta_i, \tag{3.2}$$

where $Y_i \in \mathbb{R}$ is the univariate phenotype, and $X_{i,\cdot} \in \mathbb{R}^p$ denotes the CLR-transformed microbiome abundance from p microbes, $M_i \in \mathbb{R}$ represents the unobserved metabolite of interest. ε_i refer to the error term but independent from microbiome abundance. Without loss of generality, we assume δ_i and ε_i are from normal distribution with mean 0, but we allow $\mathrm{cov}(\delta_i, \varepsilon_i) \neq 0$. In other words, ε is correlated with M. The θ^* hence can be viewed as the causal effect of metabolite.

To see this, we rewrite $\varepsilon_i = \phi^\top H_{i,\cdot} + \xi_i$, where term $H_{i,\cdot} \in \mathbb{R}^q$ represents the true hidden confounding variables that correlated with M and $\xi_i \in \mathbb{R}$ represents the random error that independent from M. Model 3.1 is now transformed as a common hidden confounding model. Because we account for the hidden confounding variable in our full model (3.1). Our goal is to estimate and test the causal effect θ^*.

When $\mathrm{cov}(\delta_i, \varepsilon_i) \neq 0$, even if every M_i is observed, the estimation of θ^* under OLS setting will be biased as the independence assumption is violated. In high dimensional setting, this is also problematic as even the consistency estimation of θ^* is generally not attainable, see [21] for detailed discussion.

Since the interested metabolite M is unobserved, a natural strategy is to impute it by a predicted $\hat{M}$ from an external multi-view microbiome dataset. The validity of conducting inference based on predicted predictor has been discussed in [39]. Therefore, our working model (3.3) is the following:

$$Y_i = \hat{M}_i\theta + \boldsymbol{\beta}^{*\top} X_{i,\cdot} + \epsilon_i, \tag{3.3}$$

where the $\hat{M}_i = \hat{\boldsymbol{\gamma}}^\top X_{i,\cdot}$. And $\hat{\boldsymbol{\gamma}}$ is predicted from a informative external dataset with the following model:

$$M'_j = \boldsymbol{\gamma}^\top X'_j + \delta'_j, \tag{3.4}$$

where each δ'_j is i.i.d from normal distribution with mean 0. As $\hat{M}_i$ is predicted from external dataset, the error term ϵ_i is independent from $\hat{M}_i$. The aforementioned bias issue is potentially bypassed, as well as the unavaibility of M, see Thorem 6 for details. However, our working model (3.3) may not be identifiable as the $\hat{M}_i$ is potentially collineared with $X_{i,\cdot}$.

Indeed, a necessary and sufficient identifibility condition in our working model is that there exists at least one instrumental microbe. To see this, we let $\mathcal{A} = \{1, ..., p\}$ index all p microbes, and categorize the microbes into the following four scenarios based on their coefficients in $\boldsymbol{\beta^{**}}$ and $\boldsymbol{\gamma}$:

(1) $\mathcal{G}_1 = \{l \in \mathcal{A} \mid \beta^*_l \neq 0 \text{ and } \gamma_l \neq 0\}$, microbes in $\mathcal{G}_1$ are related to both outcome and the metabolites. These microbes are confounders when the interest is to infer metabolome-outcome associations.

(2) $\mathcal{G}_2 = \{l \in \mathcal{A} \mid \beta^*_l = 0 \text{ and } \gamma_l \neq 0\}$, microbes in $\mathcal{G}_2$ are only related to the metabolites but not directly associated with outcome. These microbes may be served as instrumental variables for deciphering the causal relationship between the metabolites and the outcome.

(3) $\mathcal{G}_3 = \{l \in \mathcal{A} \mid \beta^*_l \neq 0 \text{ and } \gamma_l = 0\}$, microbes in $\mathcal{G}_3$ directly associated with the outcome without affecting the synthesis of the metabolites.

(4) $\mathcal{G}_4 = \{l \in \mathcal{A} \mid \beta^*_l = 0 \text{ and } \gamma_l = 0\}$, microbes in $\mathcal{G}_4$ are irrelavent.

See Figure 3.2 shows below for a clear illustration of their relationship.

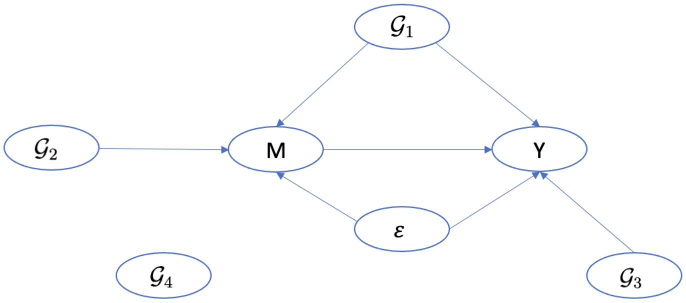

Figure 3.2: confouder $\mathcal{G}_1$ and instrumental variable $\mathcal{G}_2$

Throughout the rest of the paper, we assume $\mathcal{G}_2 \neq \emptyset$ and we leave the gap between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ to the variable selection consistency property of the variable selection methods.

3.2.2 Two Stage Parameter Estimation and Inference

From our working model (3.3), the estimation of θ^* consists of two stages. The first stage is to estimate $\boldsymbol{\gamma}$ from an external dataset and the second stage is plugging in the predicted $\hat{\boldsymbol{M}}$ to estimate θ^*. In the first stage, we adopt the popular variable selection method Lasso and the estimator $\hat{\boldsymbol{\gamma}}$ is obtained from the following optimization:

$$\hat{\boldsymbol{\gamma}} = \underset{\gamma}{\operatorname{argmin}} \left\{ \frac{\|\boldsymbol{M'} - \boldsymbol{X'}\boldsymbol{\gamma}\|_2^2}{2m} + \lambda_\gamma \sum_{i=1}^{p} |\gamma_i| \right\}, \tag{3.5}$$

where λ_γ is a tuning parameter. The $\hat{\boldsymbol{M}}$ is estimated as $\boldsymbol{X}\hat{\boldsymbol{\gamma}}$. In the second stage, we also use Lasso to the estimate θ^* . A notable difference here is that since we want to keep microbiom realated metabolite $\hat{\boldsymbol{M}}$ during the variable selection, we do not put penalty on $\hat{\boldsymbol{M}}$ as seen in (3.6). This is similar to weighted lasso where the weight on penalty for $\hat{\boldsymbol{M}}$ is 0.

$$\hat{\theta}, \hat{\boldsymbol{\beta}}^{\boldsymbol{*}} = \underset{\theta, \boldsymbol{\beta}^{\boldsymbol{*}}}{\operatorname{argmin}} \left\{ \frac{\|Y - \hat{\boldsymbol{M}}\theta - \boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{*}}\|_2^2}{2n} + \lambda \sum_{i=1}^{p} |\beta_i| \right\} \tag{3.6}$$

Proposition 1. *(consistency of Lasso estimator) denote $S_0 =: \{\beta_i \neq 0\}$, $s_{0_\gamma} = \|\gamma\|_0$ and $s_{0_\beta} = \|\beta\|_0$. With the compatibility condition (see supplement), there exist constant ϕ^2, such that $\|\theta^* - \hat{\theta}, \beta^* - \hat{\beta}\|_1 \leq 4\lambda(s_{0_\beta}+1)/\phi^2$ for a $\lambda \asymp (\lambda_\gamma s_{0_\gamma}^2 \vee \sqrt{\log p/n})$. If we choose $\lambda_\gamma \asymp \sqrt{\log p/m}$ for $\hat{\gamma}$, if $s_{0_\beta} = o(\sqrt{m/(s_{0_\gamma}^4 \log p)} \wedge \sqrt{n/\log p})$, then $\|\theta^* - \hat{\theta}, \beta^* - \hat{\beta}\|_1 \to 0$.*

However, the consistency estimator $\hat{\theta}$ from Lasso is not unbiased nor it has a distribution to allow for hypothesis test. Fortunately, such problem has been solved by the development of debiased lasso (see [36] or [46]). We hence follow the idea in [46] to construct an unbiased estimator of θ^*.

Let $\boldsymbol{z}$ denote a score function of $\hat{\boldsymbol{M}}$ defined as follows:

$$\widehat{\boldsymbol{z}} = \hat{\boldsymbol{M}} - \boldsymbol{X}\hat{\boldsymbol{b}}, \hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \left\{ \frac{\|\hat{\boldsymbol{M}} - \boldsymbol{X}\boldsymbol{b}\|_2^2}{2n} + \lambda_z \sum_{k=1}^{p} w_k |b_k| \right\}. \tag{3.7}$$

Our debiased Lasso estimator of $\tilde{\theta}$ is defined as $\hat{\theta} + \dfrac{\widehat{\boldsymbol{z}}^\top(\boldsymbol{Y} - \hat{\boldsymbol{M}}\hat{\theta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\boldsymbol{*}})}{\widehat{\boldsymbol{z}}^\top \hat{\boldsymbol{M}}}$, where $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{*}}$ are the initial lasso estimator of θ and $\boldsymbol{\beta}^{\boldsymbol{*}}$ in (3.6).

Remark 12. *$\tilde{\theta}$ can also be decomposed into of three components: θ^*, noise and bias, that is*

$$\tilde{\theta} - \theta^* = \frac{\boldsymbol{z}^\top \boldsymbol{\varepsilon}}{\boldsymbol{z}^\top \hat{\boldsymbol{M}}} + \frac{\boldsymbol{z}^\top \boldsymbol{\delta}\theta}{\boldsymbol{z}^\top \hat{\boldsymbol{M}}} + \underbrace{\sum_{i=1}^{p} \frac{\boldsymbol{z}^\top \boldsymbol{x}_i(\beta_i - \hat{\beta}_i)}{\boldsymbol{z}^\top \hat{\boldsymbol{M}}} + \sum_{i=1}^{p} \frac{\boldsymbol{z}^\top \boldsymbol{x}_i(\gamma_i - \hat{\gamma}_i)\theta^*}{\boldsymbol{z}^\top \hat{\boldsymbol{M}}}}_{\text{bias}}. \tag{3.8}$$

Theorem 6. *Suppose the regularity conditions in [46] are satisfied and we choose the λ_z from their algorithm, $\tilde{\theta}$ is an unbiased estimator of θ^*. Furthermore, under the null hypothesis when $\theta = 0$, we have*

$$\frac{\boldsymbol{z}^\top \hat{\boldsymbol{M}}}{\|\boldsymbol{z}\|_2 \hat{\sigma}_\varepsilon} \tilde{\theta} \to N(0, 1) \tag{3.9}$$

if $\lambda_\gamma \asymp \sqrt{\log p/m}$ and $\lambda \asymp (\lambda_\gamma s_{0_\gamma}^2 \vee \sqrt{\log p/n})$ and $s_0 = o(\sqrt{m}/(s_{0_\gamma}^2 \log p) \wedge \sqrt{n}/\log p)$.

Under the null hypothesis $\theta^* = 0$, the bias part $\sum_{i=1}^{p} \dfrac{\boldsymbol{z}^\top \boldsymbol{x}_i(\gamma_i - \hat{\gamma}_i)\theta}{\boldsymbol{z}^\top \hat{\boldsymbol{M}}}$ in $\tilde{\theta}$ from (3.8) is vanished, which implies the impact from the poor estimation of $\boldsymbol{\gamma}$ (hence the prediction of $\boldsymbol{M}$) is negligible. Thus the test statistics of θ^* in Theorem 1 is robust against external dataset and so does the type I error. This finding has a very important implication in practice, because it provides additional guard for choosing a suitable external dataset, if we are concerned more about type I error.

3.2.3 Partially-Informative External Dataset

In reality, it is hard to guarantee that external dataset share exactly the same relationship between the metabolites and the microbiome abundance as the target dataset. When the $\boldsymbol{\gamma}$ in model (3.4) is different from the model (3.2), a natural question arises: is our framework still valid? We do need some additional assumptions for the new $\boldsymbol{\gamma}'$ in external dataset. See proposition (2) for details.

Proposition 2. *(variation from external dataset) If the external dataset has a slight different relationship between $\boldsymbol{M}$ and $\boldsymbol{X}$, say $\boldsymbol{M}' = \boldsymbol{X}'\boldsymbol{\gamma}' + \boldsymbol{\delta}$, then Proposition 1 and Theorem 6 also holds if $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_1 \lesssim s_{0_{\gamma'}}\sqrt{\log p/m}$ and $s_{0_{\gamma'}} \lesssim s_{0_{\gamma}}$.*

Unfortunately, in practice, we can not verify if our external dataset is informative or not as we have no information of $\boldsymbol{\gamma}$ in the target dataset. But based on Theorem (6), this will not affect the type I error under the null hypothesis $\theta^* = 0$.

3.3 Numerical Study

3.3.1 Simulation Study

We conduct guided simulation studies to demonstrate the performance of proposed test. Compare to general simulation studies, in guided simulation studies, the design matrix is generated from the real data, which is the microbiome abundance data in our case. The rest terms are simulated, e.g. the error terms $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$, our response $\boldsymbol{y}$, and both metabolite $\boldsymbol{M}$ and $\boldsymbol{M}'$. The data was obtained from a collection curated data from 14 gut microbiome-metabolomic studies in [26]. Specifically, we let the shotgun measured microbiome abundance from [13] to be our design matrix in target dataset, and the one from [43] as our external dataset. We pick these two for their relative large sample size of 347 and 220, respectively. We first filter out the microbes that have missing values more than 10% from each sample. Then we select their shared microbes for our design matrix, resulting the dimension p of 393. We finally perform the centered log ratio (CLR) transformation on the compositional microbiome abundance, which is a typical approach for handling such compositional data.

In terms of model setting, the coefficient $\boldsymbol{\gamma}$ for microbiome-metabolite association consist of -0.5 for the first 20 microbes and 0.5 for the last 20 microbes, with 0 for the rest. The coefficient $\boldsymbol{\beta}^*$ consist of 0.1 for the first 10 microbes followed by -0.1 for 30 microbes and 0 for the remaining microbiome abundance. Such small scale of coefficient is from the practical consideration that the associations between microbiome and phenotype are generally weak. This parameter setting actually lead to 20

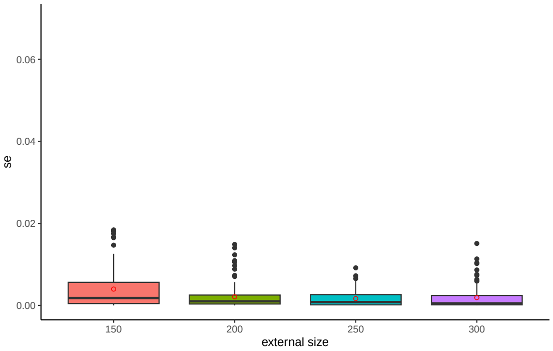

Figure 3.3: Boxplot of se for each external size

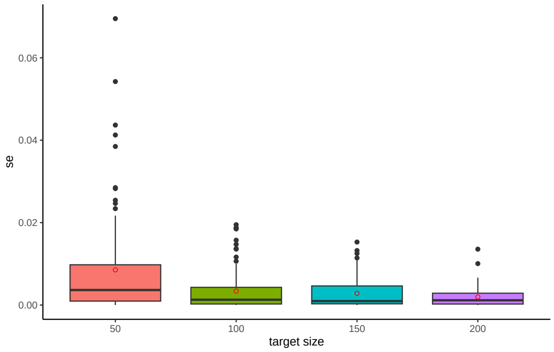

Figure 3.4: Boxplot of se for each target size

confounding microbes belong to $\mathcal{G}_1$ and 20 causal microbes belong to $\mathcal{G}_2$ and 20 microbes belong to $\mathcal{G}_3$. The error terms in (3.1) and (3.2) are generated from a bivariate normal distribution, whose covaraince matrix consist of 0.5 for off-diagonal entry and 1 for diagonal entry. All samples are randomly draw from the aforementioned two dataset. All simulation result are based on 100 times repetition.

We first fix $\theta^* = 0.2$, and investigate how the sample size of external and target dataset would affect the accuracy of our estimation of θ^*. We plot their squared error in the boxplot as shown in Figure 3.3 and Figure 3.4.

It is not surprising to see the estimation of θ^* becomes more accurate as the increase of the external sample size in Figure 3.3. Figure 3.4 reveals the similar pattern for target sample size. A interesting pattern here is that Figure 3.3 suggests a moderate external sample size (e.g. 200) is sufficient for a good estimation of θ.

We further explore the the impact of external and target sample size to the Type I error and power. And we plot the rejection rate of H_0 against the θ^* in Figure 3.5 and Figure 3.6. Figure 3.5 shows the Type I error is well controlled regardless the external sample size, and there is not much power gain from increasing it from 200 to 300. In Figure 3.6, we do notice the small target sample size (e.g. 50) would yield a slightly Type I error inflation and require a strong signal to achieve a high power. And it seems the power is more sensitive to the target sample size.

3.3.2 Variation Between Target and External Dataset

In this section, we demonstrate that how will the Type I error power change when we have a partially-informative external dataset. That is the $\boldsymbol{\gamma}$ would be different from external and target datasets. Theorem 6 implies that at least the Type I error will not be affected. Since there are many different ways how the $\boldsymbol{\gamma}$ can vary, it is not reasonable to explore them all. Here, we only consider one simple case where each of the nonzero γ_i in $\boldsymbol{\gamma}$ will now independently fluctuate with a random error from normal distribution with mean 0 and standard deviation of s, where s varies from 0 to 0.1. s here actually represent the magnitude of the fluctuation. The rest parameters will be set the same as in the previous guided simulation, but we use the full target and external dataset with $\theta^* = 0.2$. The Figure 3.7 shows the type I error

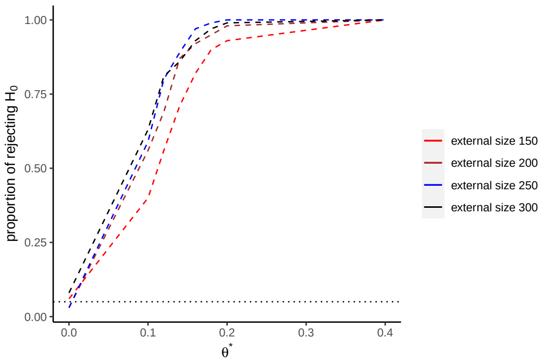

Figure 3.5: rejecting H_0 rate for each external size

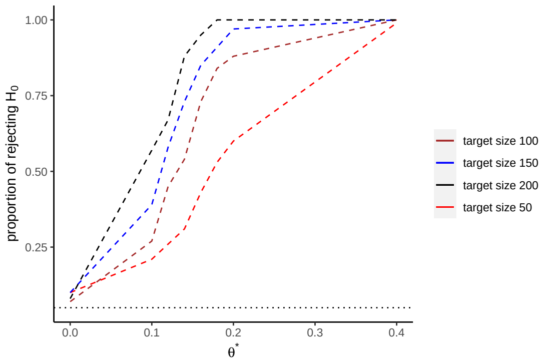

Figure 3.6: rejecting H_0 rate for each target size

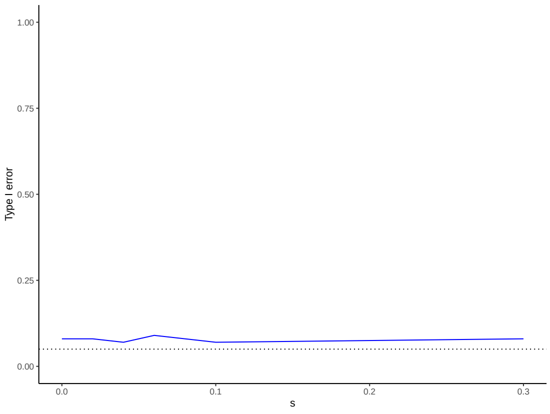

Figure 3.7: Type I error for different s

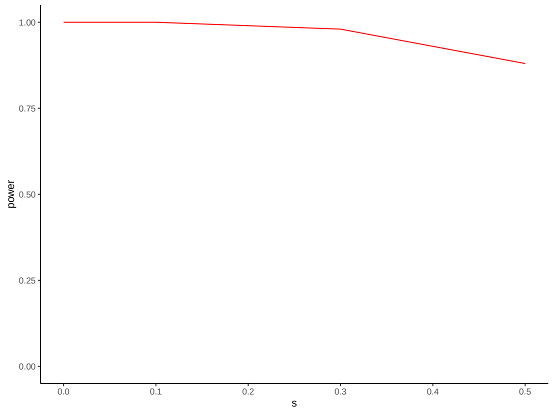

Figure 3.8: power plot for different s

is stable under different s, which is consistent with Theorem 6. Figure 3.8 shows the power would slowly decrease if the fluctuate of the variation is excessive.

3.4 A Real Data Application

In this section, we use the IHMP dataset and FRAN dataset from 3.1 to demonstrate the new findings with our method. We report the significant causal metabolites in the IBD pathogenesis.

3.4.1 Data Description

We use their curated version from [26] to be consistent. Our target dataset is the IHMP dataset mentioned in the section 3.1. which comes from a longitudinal IBD study consists of 79 patients and 26 controls. The microbiome abundance measurement is from shotgun and metabolites is measured from the untargeted and four complimentary LC-MS methods, see [22] for a detailed description of data collection. We only consider subjects' baseline measurement obtained at the time of their enrollment. Our response is the IBD biomarker: C-Reactive Protein (CRP) level in mg/L, which is a protein that human liver makes. Elevated levels of CRP are observed in the bloodstream during inflammation, which naturally makes it a good indicator for IBD. See [37] for a detailed illustration and justification. After removing the missing CRP, the final sample size is reduced to 54. We also take log transformation on it to deal with the heterogeneity.

For the preprocess of the microbiome abundance data. We first remove the microbes with more than 30% missing values. Since microbiome abundance data are compositional, we take the central log ratio (CLR) transformation as suggested by [1]. Because of the missing value, we add a small value 10^{-8} to each microbiome abundance before taking CLR. In terms of the metabolites measurement, we first remove the metabolites measurements with missing values and then take log transformation as they are count data. Since we only consider microbiome related metabolites, we apply MiRKAT to screen them based on the microbiome abundance. We identified total 157 significant metabolites.

We choose FRANZOSA-IBD-2019 (FRAN) to be the external dataset, which is from another IBD study in [13]. FRAN dataset has total 220 sample size including 56 controls. FRAN and IHMP datasets share the same method for microbiome abundance measurement and metabolites measurement. However, FRAN dataset does not have the response CRP measurement. We apply the same procedure to process the microbiome adundance data and metabolites measurement. We have identified 113 metabolites in FRAN dataset. And we use the common microbes ($p = 448$) from both datasets for the design matrix X.

In the MiRKAT screened metabolites from the FRAN dataset, 70 of them are both found in IHMP. 43 of them are exclusively from FRAN dataset, what makes it more interesting is that 14 of them are not even measured in IHMP, and the rest are measured but not selected by MiRKAT. The takeaway message here is that using external dataset can also help picking up additional microbiome correlated metabolites.

3.4.2 Analysis Results

Our focus here is to test the causal effect of the selected 43 metabolites in FRAN dataset. Our method has found found 10 new significant metabolites (annotated in red dots) as shown in Figure 3.9 below. Among the 10 metabolites, some of them have a valid clinical meaning, e.g., ADMA may play a role in suppressing inflammatory processes, according to [7].

For completeness, we also test the causal effect of rest 70 metabolites, see figure 3.10 below. Among the 12 significant metabolites(annotated in red dot), arachidonic acid pathway is a central regulator of inflammatory response as stated in [29]. And pantothenate, also known as vitamin B5, may have antioxidant effect that reduces low-grade inflammation. We leave our significant findings to be confirmed to the future experimental valididations.

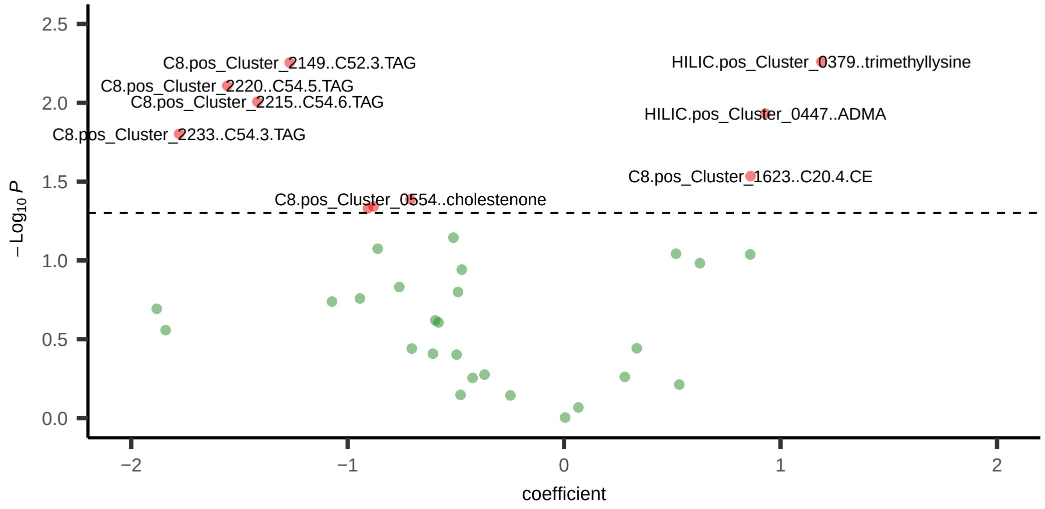

Figure 3.9: volcano plot for the 43 metabolites

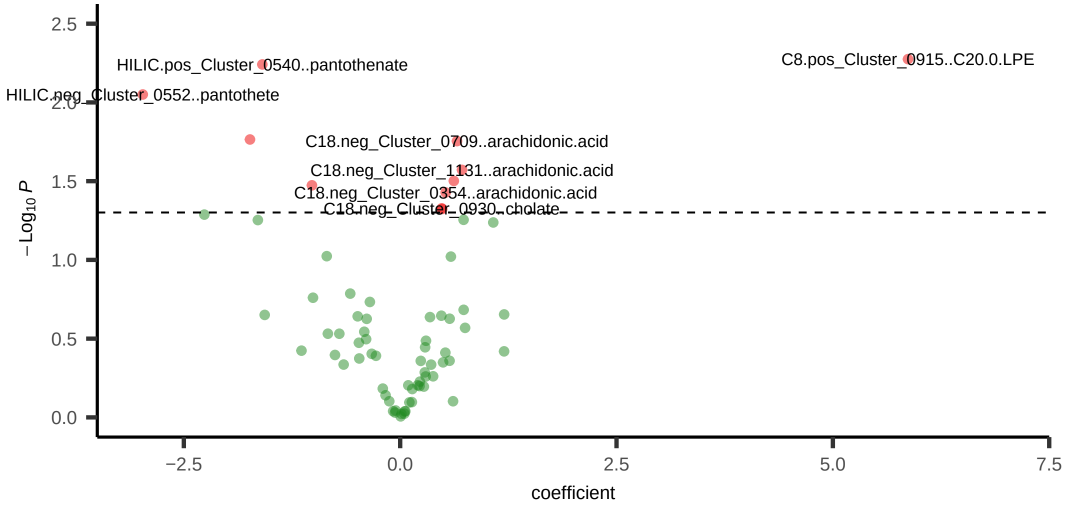

Figure 3.10: volcano plot for the rest 70 metabolites

3.5 Discussion

In this article, we proposed a new framework to study the metabolites human trait association with the microbiome by integrating multiview microbiome data. Our method can efficiently estimate the causal effect of an metabolite that unfortunately not observed in the target dataset. We also demonstrate that our proposed method has a good control of type I error as well as satisfying power when conducting the hypothesis test. We also explore the impact of variation of the microbiome and metabolites association across the training and testing sample. Our model only involves single metabolite for analysis, it would be ideal but challenging to include all existing metabolites. Specifically, it would require all the metabolites are not collineared with each other, and at least one instrumental microbe for each metabolites. Can we address identifiability issue with a more relaxed assumption? This would be a challenging but important question left for a future research topic. Essentially, developing complicated but efficient models that contain both information of microbiome and metabolites is the key to study their relationship and help us really understand how the microbiome affect human body.

Appendices

Appendix A: Proofs and Supplemental Materials of Chapter 2

A.1 Bochner's Theorem

Lemma 1. *[41, Theorem 6.6] A continuous function $k : \mathbb{R}^{p+q} \to \mathbb{R}$ is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure $W(\xi)d\xi$ on $\mathbb{R}_{p+q}$, that is,*

$$k(z) = \int_{\mathbb{R}^{(p+q)}} e^{-iz^T\xi} W(\xi)d\xi, \ \forall z \in \mathbb{R}^{p+q}.$$

A.2 Properties of $\mathrm{MC}_{\mathcal{H}}$

Some important properties of $\mathrm{MD}_{\mathcal{H}}$ and $\mathrm{MC}_{\mathcal{H}}$ are presented as follows.

Definition 4. *Given i.i.d. observations $(\boldsymbol{U}_i, V_i)_{i=1}^n$ from the distribution of $(\boldsymbol{U}, V)$. Let $a_{ij} = V_iV_j$ and $b_{ij} = k(\boldsymbol{U}_i - \boldsymbol{U}_j)$, where k is a kernel in RKHS. The unbiased sample RKHS type martingale difference divergence $\widehat{\mathrm{MD}_{\mathcal{H}}}^2(V, \boldsymbol{U})$ is defined as*

$$\widehat{\mathrm{MD}_{\mathcal{H}}}^2(V, \boldsymbol{U}) = \frac{1}{n(n-3)} \sum_{i \neq j}^{n} a^*_{ij} b^*_{ij} \tag{3.10}$$

and the unbiased sample RKHS type martingale difference correlation $\widehat{\mathrm{MC}_{\mathcal{H}}}^2(V, \boldsymbol{U})$ is defined by

$$\widehat{\mathrm{MC}_{\mathcal{H}}}^2(V, \boldsymbol{U}) = \begin{cases} \dfrac{\widehat{\mathrm{MD}_{\mathcal{H}}}^2(V, \boldsymbol{U})}{\mathrm{var}_n(V)\mathrm{var}_{n\mathcal{H}}(\boldsymbol{U})} & \text{if } \mathrm{var}_n(V)\mathrm{var}_{n\mathcal{H}}(\boldsymbol{U}) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{3.11}$$

where $var_n(V) = (\frac{1}{n(n-3)} \sum_{i \neq j}^n |a^*_{ij}|^2)^{1/2}$, and $var_{n\mathcal{H}}(\boldsymbol{U}) = (\frac{1}{n(n-3)} \sum_{i \neq j}^n |b^*_{ij}|^2)^{1/2}$.

Theorem 7. *The following properties hold if $E(V^2) < \infty$:*

a. $\mathrm{MD}^2_{\mathcal{H}}(V, \boldsymbol{U}) = E[(V - E(V))(V' - E(V'))k(\boldsymbol{U} - \boldsymbol{U}')]$.

b. $0 \leq \mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U}) \leq 1$, *and* $\mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U}) = 0 \Leftrightarrow E(V|\boldsymbol{U}) = E(V)$ *almost surely.*

c. $\mathrm{MC}_{\mathcal{H}}(a + bV, c + \boldsymbol{U}) = \mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U})$ *for any scalars a, $b \in \mathbb{R}$ and $c \in \mathbb{R}^q$.*

Theorem 8. *If $E\ (V^2) < \infty$, then*

$$lim_{n\to\infty}\widehat{\mathrm{MD}_{\mathcal{H}}}(V, \boldsymbol{U}) = \mathrm{MD}_{\mathcal{H}}(V, \boldsymbol{U}) \ a.s., \tag{3.12}$$

and

$$lim_{n\to\infty}\widehat{\mathrm{MC}_{\mathcal{H}}}(V, \boldsymbol{U}) = \mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U}) \ a.s.. \tag{3.13}$$

Theorem 9. *Assume $E(V^2) < \infty$, we have the following:*

a. If $\mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U}) = 0$, then

$$n\widehat{\mathrm{MD}_{\mathcal{H}}}^2(V, \boldsymbol{U}) \xrightarrow[n\to\infty]{D} ||\Gamma(s)||^2_{\mathcal{H}_k}, \tag{3.14}$$

where $\Gamma(\cdot)$ denotes a complex-valued zero-mean Gaussian random process with covariance function

$$\begin{aligned}\mathrm{cov}_\Gamma(s, s_0) = F(s - s_0) - g_{\boldsymbol{U}}(s - s_0)E^2(V) + \{E(V^2) + E^2(V)\}g_{\boldsymbol{U}}(s)\overline{g_{\boldsymbol{U}}(s_0)} \\ - F(s)\overline{g_{\boldsymbol{U}}(s_0)} - g_{\boldsymbol{U}}(s)\overline{F(s_0)}\end{aligned}$$

with $s, s_0 \in \mathbb{R}^q$ and $g_{\boldsymbol{U}}(s) = E(e^{i\langle s, \boldsymbol{U}\rangle})$, $F(s) = E[V^2 \exp(i\langle \boldsymbol{U}, s\rangle)]$.

b. If $\mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U}) = 0$ and $E(V^2|\boldsymbol{U}) = E(V^2)$, then

$$n\widehat{\mathrm{MD}_{\mathcal{H}}}^2(V, \boldsymbol{U})/S_n \xrightarrow[n\to\infty]{D} Q,$$

where $S_n = (1 - \frac{1}{n(n-1)}\sum_{k\neq l} k(\boldsymbol{U}_k - \boldsymbol{U}_l))(\frac{1}{n}\sum_k (V_k - \bar{V}_n)^2)$, and Q is a nonnegative quadratic form $Q = \sum_{i=1}^{\infty} \lambda_i Z_i^2$, where Z_i are independent standard normal random variables. $\{\lambda_i\}$ are nonnegative constants that depend on the distribution of (U, V) and $E(Q) = 1$.

c. If $\mathrm{MC}_{\mathcal{H}}(V, \boldsymbol{U}) \neq 0$, then $n\widehat{\mathrm{MD}_{\mathcal{H}}}^2(V, \boldsymbol{U})/S_n \xrightarrow[n\to\infty]{P} \infty$.

A.3 Sure Screening Property of $\mathrm{MC}_{\mathcal{H}}$

Let $\psi_i = \mathrm{MC}_{\mathcal{H}}(Y, X_i)$ for predictor X_i and $\widehat{\psi}_i = \widehat{\mathrm{MC}_{\mathcal{H}}}(Y, X_i)$. Denote $\widehat{\mathcal{M}} = \{j : \widehat{\psi}_j \geq cn^{-\kappa}$, for $1 \leq j \leq p\}$. Similar to $\mathrm{CMC}_{\mathcal{H}}$, we need the following two assumptions.

(B1) There exists a positive constant s_0 such that for all $0 < s \leq 2s_0$, then $E\{\exp(sY^2)\} < \infty$.

(B2) The minimum $\mathrm{MC}_{\mathcal{H}}$ value of active predictors is greather than $2cn^{-\kappa}$, for some constant $c > 0$ and $0 \leq \kappa < \dfrac{1}{2}$.

Theorem 10. *Under Assumption $(B1)$, for any $0 < \gamma < 1/2 - \kappa$, there exist postive constants c_1 and c_2 such that*

$$P\left\{\max_{1\leq j\leq p} |\widehat{\psi}_j - \psi_j| \geq cn^{-\kappa}\right\} \leq O(p[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]). \tag{3.15}$$

Under conditions $(B1)$ and $(B2)$, we have that

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}) \geq 1 - O(s_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]), \tag{3.16}$$

where s_n is the cardinality of $\mathcal{M}$.

For the sure screening property of quantile screening by $\mathrm{MC}_{\mathcal{H}}$, we require the condition (C1) in the following Section A.4. Denote $\mathcal{M}_{q_\tau} := \{j : \mathbb{E}(Y_\tau|X_j)$ depends on $X_j\}$ and $\widehat{\mathcal{M}}_{q_\tau} := \{j : \widehat{\psi}_j(\hat{Y}_\tau) \geq cn^{-\kappa}$, for $1 \leq j \leq p\}$.

Theorem 11. *Under (C1), for any $0 < \gamma < 1/2 - \kappa$ and $\kappa \in (0, 1/2)$, there exists positive constants c_1, c_2 such that for any $c > 0$,*

$$P\left\{\max_{1\leq j\leq p}|\widehat{\psi}_j(\hat{Y}_\tau) - \psi_j(Y_\tau)| \geq cn^{-\kappa}\right\} \leq O(p[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]). \tag{3.17}$$

If the minimum $MC_{\mathcal{H}}$ value of active predictors satisfies $\min_{j\in\mathcal{M}_{q_\tau}} \psi_j(Y_\tau) \geq 2cn^{-\kappa}$ for some constant $c > 0$ and $0 \leq \kappa < 1/2$, we can show that

$$P(\mathcal{M}_{q_\tau} \subseteq \widehat{\mathcal{M}}_{q_\tau}) \geq 1 - O(\widetilde{s}_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]), \tag{3.18}$$

where $\widetilde{s_n}$ is the cardinality of $\mathcal{M}_{q_\tau}$.

A.4 Sure Screening Property of Quantile Screening using $\mathrm{CMC}_{\mathcal{H}}$

We require the two assumptions below:

(C1) The CDF of Y (F_Y) is continuously differentiable in a small neighborhood of $q_\tau = q_\tau(Y)$, say $[q_\tau - \delta_0, q_\tau + \delta_0]$ for $\delta > 0$. Let $G_1(\delta_0) = \inf_{y\in[q_\tau-\delta_0, q_\tau+\delta_0]} f_Y(y)$, and $G_2(\delta_0) = \sup_{y\in[q_\tau-\delta_0, q_\tau+\delta_0]} f_Y(y)$ where f_Y is the density function of Y. Assume that $0 < G_1(\delta_0) \leq G_2(\delta_0) < \infty$.

(C2) The minimum $\mathrm{CMC}_{\mathcal{H}}$ value of active predictors satisfies $\min_{j\in\mathcal{D}_{q_\tau}} \omega_j(Y_\tau) \geq 2cn^{-\kappa}$ for some constant $c > 0$ and $0 \leq \kappa < 1/2$.

The following proposition from [31] is necessary for proving the sure screening property.

Proposition 3. *Under condition (C1), there exists $\epsilon_0 > 0$ and $c_1 > 0$, such that for any $\epsilon \in (0, \epsilon_0)$,*

$$P\left(\frac{1}{n}\sum_{l=1}^{n}|\hat{y}_{l_\tau} - y_{l_\tau}| > \epsilon\right) \leq 3\exp(-2nc_1\epsilon^2) \tag{3.19}$$

A.5 Sensitivity Analysis of Using Different Bandwidths in $\mathrm{MC}_{\mathcal{H}}$

We consider the following example from [31]. Let $g_1(x) = x, g_2(x) = (2x-1)^2, g_3(x) = sin(2\pi x)/(2 - sin(2\pi x))$, and $g_4(x) = 0.1sin(2\pi x) + 0.2cos(2\pi x) + 0.3sin^2(2\pi x) + 0.4cos^3(2\pi x) + 0.5sin^3(2\pi x)$.

Example A. $Y = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + 1.5g_1(X_5) + 1.5g_2(X_6) + 1.5g_3(X_7) + 1.5g_4(X_8) + 2g_1(X_9) + 2g_2(X_{10}) + 2g_3(X_{11}) + 2g_4(X_{12}) + \sqrt{0.5184}\epsilon$, where the predictors $X_j, j = 1, ..., p$ are *i.i.d.* from Unif(0,1), and ϵ is independent from the predictors and follows the standard normal distribution. We set $n = 400$ and $p = 1000$, and select $\lfloor n/log(n) \rfloor = 66$ variables.

Table 7: Sensitivity on Example A based on 500 replications.

bandwidth	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_6$	$\mathcal{P}_7$	$\mathcal{P}_8$	$\mathcal{P}_9$	$\mathcal{P}_{10}$	$\mathcal{P}_{11}$	$\mathcal{P}_{12}$	$\mathcal{P}_{all}$
0.001	0.21	0.21	0.58	0.76	0.49	0.55	0.97	1.00	0.83	0.88	1.00	1.00	0.00
0.50	0.68	0.46	0.95	0.99	0.96	0.90	1.00	1.000	1.00	1.00	1.00	1.00	0.22
2	0.74	0.24	0.91	0.90	0.97	0.63	1.00	1.00	1.00	0.95	1.00	1.00	0.07
4	0.75	0.11	0.89	0.79	0.97	0.30	1.00	0.99	1.00	0.68	1.00	1.00	0.01
8	0.75	0.07	0.88	0.74	0.97	0.12	1.00	0.98	1.00	0.23	1.00	1.00	0.00
100	0.78	0.05	0.86	0.58	0.98	0.04	0.99	0.88	1.00	0.08	1.00	1.00	0.00

We report the results in Table 7. For X_1 which is linearly related to Y, $\mathcal{P}_1$ increases as the bandwidth increases. But for X_2, X_3 and X_4, the corresponding selection proportions decrease as the bandwidth increases except when bandwidth is 0.001. Similar pattern can be observed with other predictors.

A.6 Sensitivity Analysis against the Conditional Set

Table 8 indicates our proposed method has a stable performance against the choice of the conditional set.

A.7 Performance under the block correlation structure

The correlations among active predictors are 0.2 and 0.1 otherwise. In addition, the variance of predictor is set as 1. Table 9 shows the superior performance of S-CMC$_{\mathcal{H}}$ with the block correlation structure among predictors.

A.8 Additional Simulation Results of Example 2

Per reviewer's comment, we evaluate the selection probability for all the interaction terms in Example 2 with a modified dimension $p_1 = 100$. In particular, the input data now has $p_1 + p_1(p_1 - 1)/2$ predictors. The results are presented in Table 10. We can see that if we consider all two-way interaction terms as predictors, we are able to select all the six terms in the true model with most of the screening methods except RASE$_1$-eBIC. It is worth pointing out that the main effects X_{15}, X_{20}, X_{25} have low selection probability. In this new example, we are indeed only screening the marginally active variables among the $p_1 + p_1(p_1 - 1)/2$ predictors. Thus, by including all the interaction effects as predictors, we are unable to demonstrate the ability of the conditional variable screening methods on screening conditionally active predictors.

Table 8: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2, with conditional set selected by different methods.

Example 2 with $\rho = 0, n = 200, p = 3000$								
	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
Conditonal set selected by Lasso								
CSIS	0.97	0.96	0.92	0.02	0.02	0.06	0.00	2241.0
CDC-SIS	0.67	0.27	0.35	0.02	0.06	0.15	0.00	2178.0
CIS	0.92	0.69	0.70	0.00	0.00	0.00	0.00	2269.0
S-CMC$_{\mathcal{H}}$	0.95	0.94	0.82	0.04	0.02	0.04	0.00	2263.0
Conditonal set selected by SIS								
CSIS	0.97	0.93	0.91	0.03	0.00	0.04	0.00	2179.5
CDC-SIS	0.87	0.88	0.86	0.06	0.16	0.41	0.00	1285.5
CIS	0.99	0.97	0.95	0.02	0.02	0.04	0.00	2368.5
S-CMC$_{\mathcal{H}}$	0.94	0.96	0.94	0.12	0.36	0.75	0.02	1051.0
Conditonal set selected by Forward Regression								
CSIS	0.94	0.91	0.90	0.04	0.03	0.04	0.00	2347.0
CDC-SIS	0.89	0.85	0.87	0.04	0.20	0.48	0.01	1396.0
CIS	0.99	0.96	0.96	0.02	0.01	0.06	0.00	2496.0
S-CMC$_{\mathcal{H}}$	0.95	0.95	0.93	0.15	0.43	0.81	0.07	801.5

Table 9: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2, with block design correlation structure.

Example 2 with $n = 200, p = 3000$								
	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
ρ=0								
MDC	0.98	1.00	0.98	0.22	0.27	0.27	0.05	536.0
CSIS(X_{S_1})	1.00	1.00	1.00	0.03	0.05	0.07	0.00	2199.5
CSIS(X_{S_2})	0.96	0.96	0.93	0.08	0.07	0.16	0.00	2574.5
CDC-SIS(X_{S_1})	1.00	1.00	1.00	0.05	0.10	0.22	0.00	1542.0
CDC-SIS(X_{S_2})	1.00	0.99	0.99	0.1	0.07	0.17	0.01	2190.5
DCSIS2	0.59	0.83	0.97	0.29	0.39	0.89	0.06	434.0
CIS(X_{S_1})	1.00	1.00	1.00	0.10	0.21	0.54	0.00	529.0
CIS(X_{S_2})	0.53	0.20	0.27	0.07	0.00	0.13	0.00	1901.5
S-CMC$_{\mathcal{H}}$(X_{S_1})	1.00	1.00	1.00	0.30	0.61	1.00	0.19	121.0
S-CMC$_{\mathcal{H}}$(X_{S_2})	0.98	0.99	0.97	0.25	0.42	0.83	0.10	273.0
RaSE.ebic1	0.84	0.78	0.78	0.01	0.00	0.01	0.00	2245.0

Table 10: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2 consider all two-way interaction terms.

	Example 2 with $n = 200, p = 100$										
	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{1,15}$	$\mathcal{P}_{5,20}$	$\mathcal{P}_{10,25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
				ρ=0							
MDC	0.92	0.89	0.92	0.01	0.00	0.04	0.86	1.00	1.00	0.63	23.0
CSIS(X_{S_2})	1.00	1.00	1.00	0.00	0.00	0.02	1.00	1.00	1.00	1.00	7.0
CDC-SIS(X_{S_2})	0.77	0.74	0.76	0.11	0.01	0.00	0.72	0.99	1.00	0.32	176.5
S-CMC$_{\mathcal{H}}$(X_{S_2})	0.99	0.97	0.99	0.13	0.43	0.62	0.91	1.00	1.00	0.86	13.0
RASE$_1$-eBIC	0.49	0.11	0.41	0.07	0.07	0.16	0.36	0.77	0.98	0.01	4238.5
CIS	1	1	0.99	0.02	0.04	0.06	0.61	0.91	0.96	0.51	35.5

Table 11: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 3, with $\boldsymbol{X}_S = \{X_1, X_5\}$.

Method	τ	$\rho = 0$								$\rho = 0.9$							
		$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
MDC	0.5	1.00	1.00	0.06	0.03	0.04	0.02	0.00	1910.0	1.00	1.00	0.97	0.49	0.06	0.05	0.49	71.0
	0.75	0.99	1.00	0.24	0.38	0.52	0.55	0.03	986.5	1.00	1.00	1.00	1.00	0.91	0.92	0.86	29.5
CSIS	0.5	1.00	1.00	0.00	0.04	0.24	0.17	0.00	2071.0	1.00	1.00	0.02	0.05	0.31	0.24	0.00	1957.0
	0.75	1.00	1.00	0.00	0.04	0.24	0.17	0.00	2345.5	1.00	1.00	0.02	0.05	0.31	0.24	0.00	2389.5
CDC-SIS	0.5	1.00	1.00	0.86	1.00	0.95	0.92	0.86	8.0	1.00	1.00	0.71	0.98	1.00	1.00	0.70	30.5
	0.75	1.00	1.00	0.86	1.00	0.95	0.92	0.76	18.0	1.00	1.00	0.71	0.98	1.00	1.00	0.70	30.5
QaSIS	0.5	1.00	1.00	0.16	0.16	0.28	0.27	0.03	1085.0	1.00	1.00	1.00	0.90	0.69	0.55	0.90	20.5
	0.75	0.92	0.99	0.14	0.38	0.69	0.72	0.03	600.0	0.99	1.00	1.00	0.97	0.99	0.97	0.92	35.5
DCSIS2	0.5	0.18	0.41	0.04	0.08	0.93	0.90	0.00	1443.5	0.07	0.15	0.06	0.05	1.00	1.00	0.00	918.0
	0.75	0.18	0.41	0.04	0.08	0.93	0.90	0.00	1443.5	0.07	0.15	0.06	0.05	1.00	1.00	0.00	918.0
CIS	0.5	1.00	1.00	0.92	1.00	0.96	1.00	0.92	15.0	1.00	1.00	0.98	1.00	1.00	1.00	0.98	23.0
	0.75	1.00	1.00	0.92	1.00	0.96	1.00	0.88	20.0	1.00	1.00	0.98	1.00	1.00	1.00	0.98	23.0
S-CMC$_{\mathcal{H}}$	0.5	1.00	1.00	0.64	1.00	0.06	0.01	0.64	37.5	1.00	1.00	0.93	0.99	0.29	0.35	0.93	13.0
	0.75	1.00	1.00	0.69	0.96	0.54	0.61	0.27	183.5	1.00	1.00	1.00	1.00	0.99	0.98	0.97	26.0
RaSE$_1$-eBIC	0.5	0.49	0.31	0.00	0.00	0.03	0.01	0.00	1816.0	0.25	0.07	0.00	0.00	0.04	0.03	0.00	2425.5
	0.75	0.49	0.31	0.00	0.00	0.00	0.00	0.00	2427.0	0.25	0.07	0.00	0.00	0.01	0.02	0.00	2427.0

A.9 Additional Simulation Results of Example 3

Table 11 shows the performance all screening methods when conditional set is selected as $\{X_1, X_5\}$.

A.10 Proof of Theorem 1

(a) Let $g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle} e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}), g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1,) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})$, and $g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) = E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$. If we expand the CMD$_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ in the representation of characteristic functions, it is

$$\int |g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) - g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1) g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2 w_1(\boldsymbol{t}_1) w_2(\boldsymbol{t}_2) d\boldsymbol{t}_1 d\boldsymbol{t}_2$$

$$= \int (|g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2)|^2 - g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) \bar{g}_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1) \bar{g}_{\boldsymbol{U}_2}(\boldsymbol{t}_2) -$$

$$\bar{g}_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) g_{\boldsymbol{V},\boldsymbol{U_1}} g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) + |g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2 |g_{\boldsymbol{U_2}}(\boldsymbol{t}_2)|^2) w_1(\boldsymbol{t}_1) w_2(\boldsymbol{t}_2) d\boldsymbol{t}_1 d\boldsymbol{t}_2,$$

where

$$
\begin{aligned}
&|g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}^{'}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}^{'}_2\rangle}],\\
&g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)\bar{g}_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\bar{g}_{\boldsymbol{U}_2}(\boldsymbol{t}_2) = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}^{'}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}^{''}_2\rangle}],\\
&\bar{g}_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}^{'}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}^{''}_2\rangle}],\\
&|g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2|g_{\boldsymbol{U_2}}(\boldsymbol{t}_2)|^2 = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}^{'}_1\rangle}]\cdot E[e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}^{''}_2\rangle}].
\end{aligned}
$$

The weight function $w_1(\boldsymbol{t}_1)$ and $w_2(\boldsymbol{t}_2)$ are integrable. With *Bochner's theorem*, for a translation-invariant positive-definite kernel $k(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x}-\boldsymbol{x}')$, we can immediately get

$$
\begin{aligned}
&\mathrm{CMD}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1)\\
&= E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) + E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1))\cdot E(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2))\\
&- 2E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)k_2(\boldsymbol{U}_2,\boldsymbol{U}^{''}_2))
\end{aligned}
$$

(b) To show $\mathrm{CMC}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) \le 1$, we can rewrite $\mathrm{CMD}^2_{\mathcal{H}}$ as

$$
\begin{aligned}
&\mathrm{CMD}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1)\\
&= E[(E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)) + VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1) - E_{V,\boldsymbol{U}_1}(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1))-\\
&E_{V',\boldsymbol{U}^{'}_1}(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1))) \times (E(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) + k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)-\\
&E_{\boldsymbol{U}_2}(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) - E_{\boldsymbol{U}^{'}_2}(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2))].
\end{aligned}
$$

We have $v(k_2,\boldsymbol{U}_2)$

$$
\begin{aligned}
&= E(k^2_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) + E^2(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) - 2E[k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)\cdot k_2(\boldsymbol{U}_2,\boldsymbol{U}^{''}_2)]\\
&=E(E(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) + k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2) - E_{\boldsymbol{U}_2}(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) - E_{\boldsymbol{U}^{'}_2}(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)))^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&v(k_{1_V},\boldsymbol{U}_1)\\
&= E(V^2(V')^2k^2_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)) + E^2(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)) - 2E[V^2V'V''k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)\cdot k_1(\boldsymbol{U}_1,\boldsymbol{U}^{''}_1)]\\
&= E(E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)) + k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1) - E_{\boldsymbol{U}_1}(k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)) - E_{\boldsymbol{U}^{'}_1}(k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)))^2,
\end{aligned}
$$

where $\mathrm{CMC}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) \le 1$ follows from an application of the *Cauchy-Schwarz* inequality. Furthermore, it is trivial to see that $\mathrm{CMC}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) \ge 0$.

Recall that $\mathrm{CMC}^2_{\mathcal{H}}$ is a standardized version of $\mathrm{CMD}^2_{\mathcal{H}}$:

$$
\begin{aligned}
&\mathrm{CMD}_{\mathcal{H}}{}^2(V,\boldsymbol{U}_2|\boldsymbol{U}_1)\\
&= \iint |E(Ve^{i(\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle+\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle)}) - E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle})|^2 w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2.
\end{aligned}
$$

Thus, $\mathrm{CMC}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) = 0$ is equivalent to $E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle})$ for any $\boldsymbol{t}_1 \in \mathbb{R}^p$ and $\boldsymbol{t}_2 \in \mathbb{R}^q$. To prove "Given $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$,

$E(V|\boldsymbol{U}_1,\boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s. if and only if $\text{CMC}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) = 0$", it suffices to prove "Given $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$, $E(V|\boldsymbol{U}_1,\boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s. if and only if $E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle})$ for any $\boldsymbol{t}_1 \in \mathbb{R}^p$ and $\boldsymbol{t}_2 \in \mathbb{R}^q$."

We first prove the " $\Rightarrow$ " direction. By definition, with probability being one,

$$\int V\frac{f(V,\boldsymbol{U}_1,\boldsymbol{U}_2)}{f(\boldsymbol{U}_1,\boldsymbol{U}_2)}dV =: E(V|\boldsymbol{U}_1,\boldsymbol{U}_2) = E(V|\boldsymbol{U}_1) := \int V\frac{f(V,\boldsymbol{U}_1)}{f(\boldsymbol{U}_1)}dV,$$

where by a slight abuse of notation, we denote $f(\cdot)$ as the corresponding probability density functions. Plugging in $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$ (i.e., $f(\boldsymbol{U}_1,\boldsymbol{U}_2) = f(\boldsymbol{U}_1)f(\boldsymbol{U}_2)$), we have almost surely

$$\int V\frac{f(V,\boldsymbol{U}_1,\boldsymbol{U}_2)}{f(\boldsymbol{U}_2)}dV = \int Vf(V,\boldsymbol{U}_1)dV. \tag{3.20}$$

Denote $g_1(\boldsymbol{U}_1) := \int V\frac{f(V,\boldsymbol{U}_1,\boldsymbol{U}_2)}{f(\boldsymbol{U}_2)}dV$ and $g_2(\boldsymbol{U}_1) := \int Vf(V,\boldsymbol{U}_1)dV$. Recall a function $f(x)$ can be transformed to its Fourier transform $\hat{f}(\xi) := \int_{-\infty}^{\infty} f(x)e^{-i2\pi\xi x}dx$ when the Dirichlet's condition holds, i.e., the integral of $f(x)$ is finite over every finite measure of support of x. Since $g_1(\boldsymbol{U}_1) = g_2(\boldsymbol{U}_1)$ and they are Fourier transformable, taking the Fourier transformation on $g_1(\boldsymbol{U}_1)$ yields

$$\hat{g}_1(-\boldsymbol{t}_1/2\pi) = \int_{\mathbb{R}^p} g_1(\boldsymbol{U}_1)e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}d\boldsymbol{U}_1 = \int_{\mathbb{R}^p} g_2(\boldsymbol{U}_1)e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}d\boldsymbol{U}_1, \tag{3.21}$$

which is equivalent to $E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}|\boldsymbol{U}_2) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})$ a.s. for any $\boldsymbol{t}_1 \in \mathbb{R}^p$. Denote $g_3(\boldsymbol{U}_2) := f(\boldsymbol{U}_2)\cdot E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}|\boldsymbol{U}_2)$ and $g_4(\boldsymbol{U}_2) := f(\boldsymbol{U}_2)\cdot E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})$, and we have $g_3(\boldsymbol{U}_2) = g_4(\boldsymbol{U}_2)$. Again, taking the Fourier transformation on $g_3(\boldsymbol{U}_2)$ leads to

$$\hat{g}_3(-\boldsymbol{t}_2/2\pi) = \int_{\mathbb{R}^q} g_3(\boldsymbol{U}_2)e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = \int_{\mathbb{R}^q} g_4(\boldsymbol{U}_2)e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}d\boldsymbol{U}_2, \tag{3.22}$$

which is equivalent to $E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle})$ for any $\boldsymbol{t}_1 \in \mathbb{R}^p$ and $\boldsymbol{t}_2 \in \mathbb{R}^q$. This completes the "only if" direction.

We next prove the " $\Leftarrow$ " direction. By the reverse Fourier transformation (or known as Fourier inversion theorem): the function $f(x)$ can be recovered by its Fourier transform $\hat{f}(\xi)$. That is, $f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)e^{i2\pi\xi x}dx$. Since $E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle})$ for any $t_1 \in \mathbb{R}^p$ and $t_2 \in \mathbb{R}^q$, we have

$$\int_{\mathbb{R}^q} g_3(\boldsymbol{U}_2)e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = \int_{\mathbb{R}^q} g_4(\boldsymbol{U}_2)e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}d\boldsymbol{U}_2, \tag{3.23}$$

which implies the equality of their Fourier transform functions, i.e.,

$$\hat{g}_3(-\boldsymbol{t}_2/2\pi) = \hat{g}_4(-\boldsymbol{t}_2/2\pi).$$

Thus,

$$g_3(\boldsymbol{U}_2) = \int_{\mathbb{R}^q} \hat{g}_3(-\boldsymbol{t}_2/2\pi)e^{-i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = \int_{\mathbb{R}^q} \hat{g}_4(-\boldsymbol{t}_2/2\pi)e^{-i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = g_4(\boldsymbol{U}_2).$$

By the definition of $g_3(\cdot)$ and $g_4(\cdot)$, dividing $f(\boldsymbol{U}_2)$ on both sides of $g_3(\boldsymbol{U}_2) = g_3(\boldsymbol{U}_2)$, we have

$$E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle}|\boldsymbol{U}_2) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})$$

a.s. for any $\boldsymbol{t}_1 \in \mathbb{R}^p$, which is equivalent to the equality (3.21). Similar arguments of applying the reverse Fourier transformation to (3.21) yields $g_1(\boldsymbol{U}_1) = g_2(\boldsymbol{U}_1)$. Dividing $f(\boldsymbol{U}_1)$ on both sides of $g_1(\boldsymbol{U}_1) = g_2(\boldsymbol{U}_1)$, together with the condition $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$, we have $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s.. This completes the "if" direction.
(c) The proof is straightforward by the definition of $\mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ and is omitted here.
(d) The proof is straightforward by the definition of $\mathrm{MC}_{\mathcal{H}}(V|\boldsymbol{U})$ and is omitted here.□

A.11 Proof for U-statistic of $\mathrm{CMC}_{\mathcal{H}}$

Recall the following notations: $a_{ij} = V_iV_jk_1(\boldsymbol{U}_{1i}, \boldsymbol{U}_{1j})$ and $b_{ij} = k_2(\boldsymbol{U}_{2i}, \boldsymbol{U}_{2j})$, $a^*_{ij} = a_{ij} - \dfrac{1}{n-2}\sum_{j=1}^n a_{ij} - \frac{1}{n-2}\sum_{i=1}^n a_{ij} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n a_{ij}$, and $b^*_{ij} = b_{ij} - \frac{1}{n-2}\sum_{j=1}^n b_{ij} - \frac{1}{n-2}\sum_{i=1}^n b_{ij} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n b_{ij}$. In addition, we let $a_{i.} = \sum_{j=1}^n a_{ij}$, $a_{.j} = \sum_{i=1}^n a_{ij}$, $a_{..} = \sum_{i,j=1}^n a_{ij}$, $b_{i.} = \sum_{j=1}^n b_{ij}$, $b_{.j} = \sum_{i=1}^n b_{ij}$, $b_{..} = \sum_{i,j=1}^n b_{ij}$, $\bar{a}_{i.} = \frac{1}{n-2}a_{i.}$, $\bar{a}_{.j} = \frac{1}{n-2}a_{.j}$, $\bar{a}_{..} = \frac{1}{(n-1)(n-2)}a_{..}$. $\bar{b}_{i.} = \frac{1}{n-2}b_{i.}$, $\bar{b}_{.j} = \frac{1}{n-2}b_{.j}$, and $\bar{b}_{..} = \frac{1}{(n-1)(n-2)}b_{..}$.
Then,

$$\begin{aligned}
\sum_{i\neq j} A_{i,j}B_{i,j} =& \sum_{i\neq j}(a_{ij}b_{ij} - a_{ij}\bar{b}_{i.} - a_{ij}\bar{b}_{.j} + a_{ij}\bar{b}_{..} - \bar{a}_{i.}b_{ij} + \bar{a}_{i.}\bar{b}_{i.} + \bar{a}_{i.}\bar{b}_{.j} - \bar{a}_{i.}\bar{b}_{..} \\
& - \bar{a}_{.j}b_{ij} + \bar{a}_{.j}\bar{b}_{i.} + \bar{a}_{.j}\bar{b}_{.j} - \bar{a}_{.j}\bar{b}_{..} + \bar{a}_{..}b_{ij} - \bar{a}_{..}\bar{b}_{i.} - \bar{a}_{..}\bar{b}_{.j} + \bar{a}_{..}\bar{b}_{..}) \\
=& \sum_{i\neq j} a_{ij}b_{ij} - \sum_i a_{i.}\bar{b}_{i.} - \sum_j a_{.j}\bar{b}_{.j} + a_{..}\bar{b}_{..} \\
& - \sum_i \bar{a}_{i.}b_{i.} + (n-1)\sum_i \bar{a}_{i.}\bar{b}_{i.} + \sum_{i\neq j}\bar{a}_{i.}\bar{b}_{.j} - (n-1)\sum_j \bar{a}_{.j}\bar{b}_{..} \\
& - \sum_j \bar{a}_{.j}b_{.j} + \sum_{i\neq j}\bar{a}_{i.}\bar{b}_{.j} + (n-1)\sum_i \bar{a}_{i.}\bar{b}_{i.} - (n-1)\sum_j \bar{a}_{.j}\bar{b}_{..} \\
& + \bar{a}_{..}b_{..} - (n-1)\sum_i \bar{a}_{..}\bar{b}_{i.} - (n-1)\sum_j \bar{a}_{..}\bar{b}_{.j} + n(n-1)\bar{a}_{..}\bar{b}_{..}.
\end{aligned}$$

Let $T_1 = \sum_{i \neq j} a_{ij} b_{ij}$, $T_2 = a_{..} b_{..}$, $T_3 = \sum_i a_{i.} b_{i.}$. Then,

$$\begin{aligned}\sum_{i \neq j} A_{i,j} B_{i,j} =& T_1 - \frac{1}{n-2} T_3 - \frac{1}{n-2} T_3 + - \frac{1}{(n-1)(n-2)} T_2 \\ & - \frac{1}{n-2} T_3 + \frac{n-1}{(n-2)^2} T_3 + \frac{1}{(n-2)^2}(T_2 - T_3) - \frac{1}{(n-2)^2} T_2 \\ & - \frac{1}{n-2} T_3 + \frac{1}{(n-2)^2}(T_2 - T_3) + \frac{n-1}{(n-2)^2} T_3 - \frac{1}{(n-2)^2} T_2 \\ & + \frac{1}{(n-1)(n-2)} T_2 - \frac{1}{(n-2)^2} T_2 - \frac{1}{(n-2)^2} T_2 + \frac{n}{(n-1)(n-2)^2} T_2 \\ & = T_1 - \frac{2}{n-2} T_3 + \frac{1}{(n-1)(n-2)} T_2.\end{aligned}$$

Let $(n)_k = n!/(n-k)!$ and I_k^n be the collections of k-tuples of indices (chosen from 1,2, ...,n) such that each index occurs exactly once. Then corresponding U-statistics estimator is

$$\begin{aligned}E(VV' k_1(\boldsymbol{U}_{1i}, \boldsymbol{U}'_{1j}) k_2(\boldsymbol{U}_2, \boldsymbol{U}'_2)) =& (n)_2^{-1} E(\sum_{(i,j) \in I_2^n}^{n} V_i V_j k_1(\boldsymbol{U}_{1i}, \boldsymbol{U}'_{1j}) k_2(\boldsymbol{U}_{1i}, \boldsymbol{U}'_{1j})) \\ =& (n)_2^{-1} E(T_1),\end{aligned}$$

$$\begin{aligned}E(VV' k_1(\boldsymbol{U}_1, \boldsymbol{U}'_1)) \cdot E(k_2(\boldsymbol{U}_2, \boldsymbol{U}'_2)) =& (n)_4^{-1} E(\sum_{(i,j,q,r) \in I_2^n}^{n} V_i V_j k_1(\boldsymbol{U}_{1i}, \boldsymbol{U}'_{1j}) k_2(\boldsymbol{U}_{2q}, \boldsymbol{U}'_{2r})) \\ =& (n)_4^{-1} E(T_2 - 4T_3 + 2T_1),\end{aligned}$$

and

$$E(VV' k_1(\boldsymbol{U}_1, \boldsymbol{U}'_1) k_2(\boldsymbol{U}_2, \boldsymbol{U}'_2)) = (n)_3^{-1} E(\sum_{(i,j,r) \in I_2^n}^{n} V_i V_j k_1(\boldsymbol{U}_{1i}, \boldsymbol{U}'_{1j}) k_2(\boldsymbol{U}_{2i}, \boldsymbol{U}'_{2r})) = (n)_3^{-1} E(T_2 - T_1).$$

Combine the expectations, we get $\text{CMD}_{\mathcal{H}} = E(T_1 - \frac{2}{n-2} T_3 + \frac{1}{(n-1)(n-2)} T_2)$. Then the unbiased estimator is $T_1 - \frac{2}{n-2} T_3 + \frac{1}{(n-1)(n-2)} T_2$. □

A.12 Proof of Theorem 2

If $\mathbb{E}(V^2) < \infty$, we need to prove almsot surely convergence for the following two expressions.

$$lim_{n \to \infty} \widehat{\text{CMD}}_{\mathcal{H}}(V, \boldsymbol{U}_2 | \boldsymbol{U}_1) = \text{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2 | \boldsymbol{U}_1), \tag{3.24}$$

and

$$lim_{n \to \infty} \widehat{\text{CMC}}_{\mathcal{H}}(V, \boldsymbol{U}_2 | \boldsymbol{U}_1) = \text{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2 | \boldsymbol{U}_1). \tag{3.25}$$

By the *Strong law of large numbers* for U-statistics, we can immediately get the result. □

A.13 Proof of Theorem 3

(a). Under the conditional independence, we have $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ and $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$. Define the process $\Gamma_n(\boldsymbol{t}_1, \boldsymbol{t}_2) = \sqrt{n}\xi_n(\boldsymbol{t}_1, \boldsymbol{t}_2) = \sqrt{n}(g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) - g^n_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)g^n_{\boldsymbol{U}_2}(\boldsymbol{t}_2))$, then $E(\Gamma_n(\boldsymbol{t}_1, \boldsymbol{t}_2)) = 0$. The weak convergence of $||\Gamma_n||^2_{\mathcal{H}}$ to $||\Gamma||^2_{\mathcal{H}}$ follows from multivariate *central limit theorem* and *continuous mapping theorem.* Let $F(\boldsymbol{t}_1, \boldsymbol{t}_2) = E(V^2 \exp(i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle)\text{exp}(i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle))$. We have

$$\begin{aligned}
&E[\Gamma_n(\boldsymbol{t}_1, \boldsymbol{t}_2)\overline{\Gamma_n(\boldsymbol{t}'_1, \boldsymbol{t}'_2)}] \\
&= E(n(g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) - g^n_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)g^n_{\boldsymbol{U}_2}(\boldsymbol{t}_2))(\overline{g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}'_1, \boldsymbol{t}'_2)} - \overline{g^n_{V,\boldsymbol{U}_1}(\boldsymbol{t}'_1)g^n_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)})).
\end{aligned}$$

A direct calculation yields to the following:

$$\begin{aligned}
E[\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)\overline{\Gamma_n(\boldsymbol{t}'_1, \boldsymbol{t}'_2)}] = {} & \frac{(n-1)^2}{n^2}F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, \boldsymbol{t}_2 - \boldsymbol{t}'_2) + \frac{n-1}{n}g_{\boldsymbol{U}_2}(\boldsymbol{t}_2 - \boldsymbol{t}'_2)[\frac{1}{n}F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, 0) - \\
& g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\overline{g_{V,\boldsymbol{U}_1}(\boldsymbol{t}'_1)}] + \frac{n-1}{n}[g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\overline{g_{V,\boldsymbol{U}_1}(\boldsymbol{t}'_1)} + \\
& \frac{n-2}{n}F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, 0)]g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)} - \frac{(n-1)^2}{n^2}(F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)} + \\
& g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, -\boldsymbol{t}'_2)).
\end{aligned}$$

In particular,

$$\begin{aligned}
E|\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 = {} & \frac{n-1}{n}E(V^2)(1 + \frac{n-2}{n}|g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) - \frac{n-1}{n}|g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2(1 - |g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) - \\
& \frac{(n-1)^2}{n^2}[F(0, \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)} + g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(0, -\boldsymbol{t}'_2)].
\end{aligned}$$

(b). According to the first assertion, we have

$$\begin{aligned}
E||\Gamma||^2_{\mathcal{H}} &= \int Cov_\Gamma((\boldsymbol{t}_1, \boldsymbol{t}_2), (\boldsymbol{t}_1, \boldsymbol{t}_2))dw \\
&= \int \{[E(V^2) - |g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2](1 - |g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) + 2E(V^2)|g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2 - F(0, \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)} - \\
&\quad g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(0, -\boldsymbol{t}_2)\}dw.
\end{aligned}$$

Under the assumption $E(V^2|\boldsymbol{U}_2) = E(V^2)$, which implies $F(0, \boldsymbol{t}_2) = E(V^2)g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)$, we have $E\|\Gamma\|^2_{\mathcal{H}} = E(V^2) - E(V^2)E(k_2(\boldsymbol{U}_2 - \boldsymbol{U}'_2)) - E(VV'k_1(\boldsymbol{U}_1 - \boldsymbol{U}'_1)) + E(VV'k_1(\boldsymbol{U}_1 - \boldsymbol{U}'_1)k_2(\boldsymbol{U}_2 - \boldsymbol{U}'_2))$.

Let $S_n = (\frac{1}{n}\sum_i V_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j} a_{ij})(1 - \frac{1}{n(n-1)}\sum_{i\neq j} b_{ij})$, then by the SLLN for U-statistics, $S_n \xrightarrow[n\to\infty]{a.s} E\|\Gamma\|^2_{\mathcal{H}}$. Therefore $n\widehat{\text{CMD}}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)/S_n \xrightarrow[n\to\infty]{D} Q$, where E(Q)=1 and Q is a nonnegative quadratic form of centered Gaussian random variable following the argument in the proof of Corollary 2 of Szekely er al.(2007).

(c). Suppose that $\text{CMD}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) > 0$, then Theorem 3 implies that $\widehat{\text{CMD}}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \xrightarrow[n\to\infty]{a.s.} \text{CMD}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) > 0$, therefore $n\widehat{\text{CMD}}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \xrightarrow[n\to\infty]{a.s.} \infty$. By the SLLN, S_n converges to a constant, and therefore $n\widehat{\text{CMD}}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)/S_n \xrightarrow[n\to\infty]{a.s.} \infty$. □

A.14 Proof of Theorem 4

For the notational convinence, we use Y, Z, X to represent the original notation $V, \boldsymbol{U}_1$ and $\boldsymbol{U}_2$. Let $S_1^j = E[YY'k_1(Z-Z')k_2(X_j-X_j')], S_2^j = E[YY'k_1(Z-Z')]E[k_2(X_j-X_j')]$ and $S_3^j = E[YY'k_1(Z-Z')k_2(X_j-X_j'')]$, where (X_j', Y', Z') and (X_j'', Y'', Z'') are *i.i.d.* copies of (X_j, Y, Z). $(n)_k = n!/(n-k)!$ and I_k^n be the collections of k-tuples of indices (chosen from $\{1, 2, ..., n\}$) such that each index occurs exactly once. Correspondingly, their unbiased sample counterparts are

$$S_{1n}^j = (n)_2^{-1} \sum_{(k,l)\in I_2^n} Y_kY_lk_1(Z_k - Z_l)k_2(X_{jk} - X_{jl}),$$

$$S_{2n}^j = (n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n} Y_kY_lk_1(Z_k - Z_l)k_2(X_{jh} - X_{jq}), \text{and}$$

$$S_{3n}^j = (n)_3^{-1} \sum_{(k,l,h)\in I_3^n} Y_kY_lk_1(Z_k - Z_l)k_2(X_{jk} - X_{jh}).$$

Since $\text{CMD}_{\mathcal{H}}^j$ and $\widehat{\text{CMD}}_{\mathcal{H}}^j$ can be expressed as $(\text{CMD}_{\mathcal{H}}^j)^2 = S_1^j + S_2^j - 2S_3^j$ and $(\widehat{\text{CMD}}_{\mathcal{H}}^j)^2 = S_{1n}^j + S_{2n}^j - 2S_{3n}^j$. We shall establish the consistency result for each part separately.

Consistency of S_{1n}^j. Since S_{1n}^j is a U-statistic with the kernel function $h_1(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k) = Y_kY_lk_1(Z_k - Z_l)k_2(X_{jk} - X_{jl})$. Rewrite $S_{1n}^j = \{n(n-1)\}^{-1}\sum_{k\neq l} h_1I\{|h_1| \leq M\} + \{n(n-1)\}^{-1}\sum_{k\neq l} h_1I\{|h_1| > M\} = S_{1n,1}^j + S_{1n,2}^j$. Correspondingly, its population counterpart can also be decomposed as $S_1^j = E[h_1I\{|h_1| \leq M\}] + E[h_1I\{|h_1| > M\}] = S_{1,1}^j + S_{1,2}^j$. Note that $S_{1n,1}^j + S_{1n,2}^j$ are unbiased estimators of $S_{1,1}^j + S_{1,2}^j$, respectively.

To show the consistency of $S_{1n,1}^j$, we note that all U-statistics can be expressed as an average of averages of iid random variables, see [30] (section 5.1.6). Denote $m = \lfloor n/2 \rfloor$, and define $\Omega(X_{j1}, Y_1, Z_1; ...; X_{jn}, Y_n, Z_n) = \frac{1}{m}\sum_{r=0}^{m-1} h_1^{(r)}I\{|h_1^{(r)}| \leq M\}$, where $h_1^{(r)} = h_1(X_{j\ 1+2r}, Y_{1+2r}, Z_{1+2r}; X_{j\ 2+2r}, Y_{2+2r}, Z_{2+2r})$. Then, we have $S_{1n,1}^j = (n!)^{-1}\sum_{n!}\Omega(X_{ji_1}, Y_{i_1}, Z_{i_1}; ...; X_{ji_n}, Y_{i_n}, Z_{i_n})$, where $\sum_{n!}$ denote summation over all $n!$ permutations $(i_1, ..., i_n)$ of $(1, ..., n)$. By *Jensen's inequality*, for $t > 0$, we have

$$\begin{aligned}
E[\exp(tS_{1n,1}^j)] &= E[\exp\{t(n!)^{-1}\sum_{n!}\Omega(X_{ji_1}, Y_{i_1}, Z_{i_1}; ...; X_{ji_n}, Y_{i_n}, Z_{i_n})\} \\
&\leq (n!)^{-1}\sum_{n!}E[\exp(t\sum_{r=0}^{m-1} h_1^{(r)}I\{|h_1^{(r)}| \leq M\}/m)] \\
&= E^m[\exp(th_1^{(r)}I\{|h_1^{(r)}| \leq M\}/m)],
\end{aligned}$$

which entails that

$$\begin{aligned}P(S^j_{1n,1} - S^j_{1,1} \geq \epsilon) &\leq \exp(-t\epsilon)\exp(-tS^j_{1,1})E[\exp(t\tilde{S}^j_{1n,1})]\\ &\leq \exp(-t\epsilon)\cdot E^m\{\exp[t(h^{(r)}_1 I\{|h^{(r)}_1| \leq M\} - S^j_{1,1})/m]\}\\ &\leq \exp(-t\epsilon)\cdot\exp\{t^2M^2/(2m)\},\end{aligned}$$

using *Markov's inequality* and *Hoeffding's inequality* (see lemma 1 of [20]) in the first and third inequality above, respectively. Set $t = \epsilon m/M^2$ and utilize the symmetry of U-statistics, we can then obtain $P(|S^j_{1n,1} - S^j_{1,1}| \geq \epsilon) \leq 2\exp\{-\epsilon^2 m/(2M^2)\}$.

The next part is for dealing with $S^j_{1n,2}$. By *Cauchy-Schwarz inequality* and *Markov's inequality*, $(S^j_{1,2})^2 = (E[h_1 I\{|h_1| > M\}])^2 \leq E[h^2_1]\cdot P\{|h_1| > M\} \leq E[h^2_1]E[|h_1|^q]\cdot M^{-q}$ for any $q \in \mathbb{N}$. From the inequality $|ab| \leq (a^2+b^2)/2, a, b \in \mathbb{R}$, we get $|h_1(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l)| \leq \frac{1}{2}(Y^2_k + Y^2_l)(k_1(Z_k - Z_l)k_2(X_{jk} - X_{jl})) \leq K^2Y^2_k$ as the kernel k_1 and k_2 are bounded by some constant K. Hence $E[|h_1|^q]$ is bounded basd on assumption (A1). Thus, if we let $M = n^\gamma$ for $0 < \gamma < 1/2 - \kappa$, then $S^j_{1,2} \leq \epsilon/2$ for sufficiently large n (in the sense we sepecify $\epsilon = cn^{-\kappa}$ and q can be any integer greater than $2\kappa/\gamma$). Hence, $P(|S^j_{1n,2} - S^j_{1,2}| \geq \epsilon) \leq P(|S^j_{1n,2}| \geq \epsilon/2)$. Since the event $\{|S^j_{1n,2}| \geq \epsilon/2\}$ implies the event $\{Y^2_k \geq M/K^2, \text{ for some } 1 \leq k \leq n\}$, we have that

$$\begin{aligned}P\{|S^j_{1n,2}| \geq \epsilon/2\} &\leq P(\cup^n_{k=1}\{Y^2_k \geq M/K^2\})\\ &\leq \sum^n_{k=1} P(\{Y^2_k \geq M/K^2\})\\ &\leq nP(\{Y^2_k \geq M/K^2\}).\end{aligned}$$

Invoking assumption (A1) and *Markov's inequality*, there must exist a constant C, such that $P(\{Y^2_k \geq M/K^2\} \leq C\exp(-sM/K^2)$ for any k and $s \in (0, 2s_0]$. Consequently, for sufficiently large n, $\max_{1\leq j\leq p} P(|S^j_{1n,2} - S^j_{1,2}| \geq \epsilon) \leq \max_{1\leq j\leq p} P(|S^j_{1n,2}| \geq \epsilon/2) \leq \max_{1\leq p\leq n} nP(\{Y^2_k \geq M/K^2\}) \leq nC\exp(-sM/K^2)$. In combination with the convergence result of $S^j_{1n,1}$, we get that for large enough n,

$$\begin{aligned}P(|S^j_{1n} - S^j_1| \geq 2\epsilon) &\leq P(|S^j_{1n,1} - S^j_{1,1}| \geq \epsilon) + P(|S^j_{1n,2} - S^j_{1,2}| \geq \epsilon)\\ &\leq 2\exp(-\epsilon^2 n^{1-2\gamma}/4) + Cn\exp(-sn^\gamma/K^2).\end{aligned}$$

Consistency of S^j_{2n}. We can rewrite S^j_{2n} as follows:

$$\begin{aligned}&S^j_{2n}\\ &= \frac{1}{(n)_4}\sum_{k<l<h<q} 4[Y_kY_lk_1(Z_k - Z_l)k_2(X_{jh} - X_{jq}) + Y_kY_hk_1(Z_k - Z_h)k_2(X_{jl} - X_{jq}) +\\ &Y_kY_qk_1(Z_k - Z_q)k_2(X_{jl} - X_{jh}) + Y_lY_hk_1(Z_l - Z_h)k_2(X_{jq} - X_{jk}) +\\ &Y_lY_qk_1(Z_l - Z_q)k(X_{jh} - X_{jk}) + Y_hY_qk_1(Z_h - Z_q)k_2(X_{jl} - X_{jk})]\\ &=24(n)^{-1}_4 \sum_{k<l<h<q} h_2(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l; X_{jh}, Y_h, Z_h; X_{jq}, Y_q, Z_q),\end{aligned}$$

where $h_2(X_{jk}, Y_k, Z_{jk}; X_{jl}, Y_l, Z_{jl}; X_{jh}, Y_h, Z_{jh}; X_{jq}, Y_q, Z_{jq})$ is the kernel function. Following the same argument in ⟨i⟩, we write S^j_{2n} as $S^j_{2n} = 24(n)_4^{-1} \sum_{k<l<h<q} h_2 I(|h_2| \leq M) + 24(n)_4^{-1} \sum_{k<l<h<q} h_2 I(|h_2| \geq M) = S^j_{2n,1} + S^j_{2n,2}$ and their population versions $S^j_2 = E[h_2 I\{|h_2| \leq M\}] + E[h_2 I\{|h_2| \geq M\}] = S^j_{2,1} + S^j_{2,2}$. Using the same argument as for $S^j_{1n,1}$, we can show that

$$P(|S^j_{2n,1} - S^j_{2,1}| \geq \epsilon) \leq 2\exp\{-\epsilon^2 m'/(2M^2)\},$$

where $m' = \lfloor n/4 \rfloor$, due to the fact that S^j_{2n} is a fourth-order U-statistics.

Now it remains to establish the uniform convergence of the other part $S^j_{2n,2}$. Note that $|h_2(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l; X_{jh}, Y_h, Z_h; X_{jq}, Y_q, Z_q)| \leq [\frac{1}{4}(Y_k^2 + Y_l^2 + Y_h^2 + Y_q^2)] * K^2$, so the event $\{|S^j_{2n}| \geq \epsilon/2\}$ implies the event $\{Y_k^2 \geq M/K^2\}$ for some $1 \leq k \leq n$. Therefore, following a similar argument as presented in Part I, we have

$$\begin{aligned} P(|S^j_{2n,1} - S^j_{2,1}| \geq \epsilon) &\leq P(|S^j_{2n,2}| \geq \epsilon/2) \\ &\leq P(\cup_{k=1}^n [Y_k^2 \geq M/K^2]) \\ &\leq Cn \exp(-sM/K^2), \end{aligned}$$

for any k and $s \in (0, 2s_0]$. Combining the two convergence results for $S^j_{3n,1}$ and $S^j_{3n,2}$ with $M = n^\gamma$ for some $0 < \gamma < 1/2 - \kappa$, it follows that

$$P(|S^j_{2n} - S^j_2| \geq 2\epsilon) \leq 2\exp(-\epsilon^2 n^{1-2\gamma}/8) + Cn\exp(-sn^\gamma/K^2).$$

Consistency of S^j_{3n}. We can rewrite S^j_{3n} as follows:

$$\begin{aligned} S^j_{3n} =& (n)_3{}^{-1} \sum_{k<l<h} [Y_k Y_l k_1(Z_k - Z_l) k_2(X_{jk} - X_{jh}) + \\ & Y_k Y_h k_1(Z_k - Z_h) k_2(X_{jk} - X_{jl}) + Y_l Y_k k_1(Z_l - Z_k) k_2(X_{jl} - X_{jh})] + \\ & Y_l Y_h k_1(Z_l - Z_h) k_2(X_{jl} - X_{jk}) + Y_h Y_k k_1(Z_h - Z_k) k_2(X_{jh} - X_{jl}) + \\ & Y_h Y_l k_1(Z_h - Z_l) k_2(X_{jh} - X_{jk}) \\ =& 6(n)_3{}^{-1} \sum_{k<l<h} h_3(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k; X_{jh}, Y_h, Z_h), \end{aligned}$$

where $h_3(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k; X_{jh}, Y_h, Z_h)$ is the kernel function. Again, we write S^j_{3n} as $S^j_{3n} = 6\{n(n-1)(n-2)\}^{-1} \sum_{k<l<h} h_3 I(|h_3| \leq M) + 6\{n(n-1)(n-2)\}^{-1} \sum_{k<l<h} h_3 I(|h_3| \geq M) = S^j_{3n,1} + S^j_{3n,2}$ and its population counterpart as $S^j_3 = E[h_3 I\{|h_3| \leq M\}] + E[h_3 I\{|h_3| \geq M\}] = S^j_{3,1} + S^j_{3,2}$. By using the same argument for $S^j_{1n,1}$, we can show that

$$P(|S^j_{3n,1} - S^j_{3,1}| \geq \epsilon) \leq 2\exp\{-\epsilon^2 m'/(2M^2)\},$$

where $m' = \lfloor n/3 \rfloor$, due to the fact that S^j_{3n} is a third-order U-statistics. Now it remains to establish the uniform convergence of the other part $S^j_{3n,2}$. Note that $|h_3(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k; X_{jh}, Y_h, Z_h)| \le [\frac{1}{3}(Y_k^2 + Y_l^2 + Y_h^2)] * K^2$, so the event $\{|S^j_{3n}| \ge \epsilon/2\}$ implies the event $\{Y_k^2 > M/K^2\}$ for some $1 \le k \le n$. Therefore, following a similar argument as presented in Part I, we have

$$\begin{aligned} P(|S^j_{3n,1} - S^j_{3,1}| \ge \epsilon) &\le P(|S^j_{3n,2}| \ge \epsilon/2) \\ &\le P(\cup_{k=1}^n [Y_k^2 \ge M/K^2]) \\ &\le Cn \exp(-sM/K^2), \end{aligned}$$

for any k and $s \in (0, 2s_0]$. Combining the two convergence results for $S^j_{3n,1}$ and $S^j_{3n,2}$ with $M = n^\gamma$ for some $0 < \gamma < 1/2 - \kappa$, it follows that

$$P(|S^j_{3n} - S^j_3| \ge 2\epsilon) \le 2\exp(-\epsilon^2 n^{1-2\gamma}/6) + Cn \exp(-sn^\gamma/K^2)).$$

This, together with the consistency in Part I, Part II and Part III, we have

$$\begin{aligned} &P\{|(2S^j_{3n} - S^j_{1n} - S^j_{2n}) - (2S^j_3 - S^j_1 - S^j_2)| \ge \epsilon\} \\ &\le P(|S^j_{3n} - S^j_3| \ge \frac{\epsilon}{4}) + P(|S^j_{2n} - S^j_2| \ge \frac{\epsilon}{4}) + P(|S^j_{1n} - S^j_1| \ge \frac{\epsilon}{4}) \\ &= O\{\exp(-c_1\epsilon^2 n^{1-2\gamma}) + n\exp(-c_2 n^\gamma)\}. \end{aligned}$$

for some positive constants c_1 and c_2 and the bound is uniform with respect to $j = 1, ..., p$. Analyzing the denominator of $\widehat{\omega}_j$ would have the same form of convergence rate, so we omit the details here. Let $\epsilon = cn^{-\kappa}$, where κ satisfies $0 < \kappa + \gamma < 1/2$, we then have

$$\begin{aligned} P\{\max_{1 \le j \le p-d_1} |\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}\} \le& (p - d_1) \max_{1 \le j \le p-d_1} P\{|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}\} \\ \le& O((p - d_1)[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^\gamma)]). \end{aligned}$$

If $\mathcal{D}_S \not\subseteq \widehat{\mathcal{D}}_S$, then there exist some $j \in \mathcal{D}_S$, such that $\widehat{\omega}_j < cn^{-\kappa}$. According to assumption (A2), for this particular j , we would have $|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}$, which implies that $A = \{\mathcal{D}_S \not\subseteq \widehat{\mathcal{D}}_S\} \subseteq \{|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}, \text{for some } j \in \mathcal{D}_S\} = B$ and hence $B^c \subseteq A^c$. Finally,

$$\begin{aligned} P(A^c) \ge P(B^c) = 1 - P(B) &= 1 - P(|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}, \text{for some } j \in \mathcal{D}_S) \\ &\ge 1 - s_n \max_{(j \in \mathcal{D}_S)} P(|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}) \\ &\ge 1 - O(s_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^\gamma)]). \end{aligned}$$

where the first inequality above is due to *Bonferroni's inequality.*

Ranking Consistence. Recall the assumption that $\delta = \min_{j\in\mathcal{D}_S}\omega_j - \max_{j\in\mathcal{D}_S}\omega_j$,

$$\begin{aligned}
&P\left\{\min_{j\in\mathcal{D}_S}\widehat{\omega}_j \le \max_{j\in\mathcal{I}_S}\widehat{\omega}_j\right\}\\
&= P\left\{\min_{j\in\mathcal{D}_S}\widehat{\omega}_j - \min_{j\in\mathcal{D}_S}\omega_j + \delta \le \max_{j\in\mathcal{I}_S}\widehat{\omega}_j - \max_{j\in\mathcal{I}_S}\omega_j\right\}\\
&\le P\left\{\max_{j\in\mathcal{D}_S}|\widehat{\omega}_j - \omega_j| \ge \delta/2\right\} + P\left\{\max_{j\in\mathcal{I}_S}|\widehat{\omega}_j - \omega_j| \ge \delta/2\right\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&P\left\{\min_{j\in\mathcal{I}_S}\widehat{\omega}_j < \max_{j\in\mathcal{D}_S}\widehat{\omega}_j\right\}\\
&\ge 1 - 2O((p-d_1)[\exp(-c_1'\delta^2 n^{1-2\gamma}) + n\exp(-c_2 n^\gamma)]).
\end{aligned}$$

□

A.15 Proof of proposition 1

Following the proof of Proposition 2 in [31], we rewrite $|y_{l_\tau} - \hat{y}_{l_\tau}|$ as $|I(y_l \le q_\tau) - I(y_l \le \widehat{q}_\tau)|$, which is $I(\widehat{q}_\tau < y_l \le q_\tau) + I(q_\tau < y_l \le \widehat{q}_\tau)$. Then $P(\frac{1}{n}\sum_{l=1}^n |\hat{y}_{l_\tau} - y_{l_\tau}| > \epsilon) \le P(\frac{1}{n}\sum_{l=1}^n |\hat{y}_{l_\tau} - y_{l_\tau}| > \epsilon, |\widehat{q}_\tau - q_\tau| \le \delta) + P(|\widehat{q}_\tau - q_\tau| \ge \delta| =: P_1 + P_2$. For P_2, we apply [30] Theorem 2.3.2 and get $P_2 = P(|\widehat{q}_\tau - q_\tau| \ge \delta) \le 2\exp(-2nL(\delta)^2)$, where $L(\delta) = \min\{F_Y(q_\tau + \delta) - \tau, \tau - F_Y(q_\tau + \delta)\}$. Under the Assumption (B1), we have $G_1(\delta_0)\delta \le L(\delta) \le G_2(\delta_0)\delta$. Let $\delta = \min(\epsilon/\{4G_2(\delta_0)\}, \delta_0)$ if $\epsilon < \epsilon_0 = 4G_2(\delta_0)\delta_0$. Then for $\epsilon \in (0, \epsilon_0)$, we have

$$P_2 \le 2\exp(-2nG_1(\delta_0)^2\delta^2) \le 2\exp(-2n\frac{G_1(\delta_0)^2}{16G_2(\delta_0)^2}\epsilon^2).$$

Setting $P_{q_\tau} = P(|y_l - q_\tau| \le \delta)$, we can find a bound for P_1:

$$P_1 \le P(\frac{1}{n}\sum_{l=1}^n I(|y_l - q_\tau| \le \delta) > \epsilon) = P(\frac{1}{n}\sum_{l=1}^n I(|y_l - q_\tau| \le \delta) - P_{q_\tau} > \epsilon - P_{q_\tau}).$$

By Hoeffding's inequality, $P_1 \le \exp(-2(\epsilon - P_{q_\tau})^2 n)$. Since $P_{q_\tau} \le 2\delta G_2(\delta_0) \le \epsilon/2$ when $\epsilon \in (0, \epsilon_0)$, then $I_1 \le \exp(-2n\epsilon^2/4)$. Together with the bound for P_2, we have $P\left(\dfrac{1}{n}\sum_{l=1}^n |\widehat{y_{l_\tau}} - y_{l_\tau}| > \epsilon\right) \le 3\text{exp}(-2nc_1\epsilon^2)$. □

A.16 Proof of theorem 5

We shall show the uniform consistency of $\widehat{\omega}_j(\hat{Y}_\tau) = \text{CMC}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$ under the assumptions (B1) and (B2). Due to the similarity of its numerator and denominator, we only present the numerator part, which is the consistency of $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S))$. First we show the consistency of $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$, and then show the $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$ is a consistent estimator of $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$. Following the similar procedures in the proof of section 5, for any $\gamma \in (0, 1/2-\kappa)$, there exist positive constants c_1 and c_2 such that

$$P(|\mathrm{CMD}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S) - \widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \geq \epsilon) \leq C[\exp\{-c_1\epsilon^2 n^{1-2\gamma}\} + n\exp(-c_2 n^\gamma)]. \tag{3.26}$$

for a sufficiently small ϵ (i.e. $\epsilon = cn^{-\kappa}$, which will be defined later). Next we focus on the difference between $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$ and $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$. Denote $\widehat{T}^j_{1n} = (n)_2^{-1} \sum_{(k,l)\in I_2^n} \hat{y}_{k_\tau}\hat{y}_{l_\tau} k_1(Z_k - Z_l)k_2(X_{jk} - X_{jl}), \widehat{T}^j_{2n} = (n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n} \hat{y}_{k_\tau}\hat{y}_{l_\tau} k_1(Z_k - Z_l)k_2(X_{jh} - X_{jq})$, and $\widehat{T}^j_{3n} = (n)_3^{-1} \sum_{(k,l,h)\in I_3^n} \hat{y}_{k_\tau}\hat{y}_{h_\tau} k_1(Z_k - Z_h)k_2(X_{jk} - X_{jl})$. Similarly, T^j_{1n}, T^j_{2n} and T^j_{3n} are defined as $\{\hat{y}_{k_\tau}\}_{k=1}^n$ replaced with $\{W_k\}_{k=1}^n$. Let $C_0 = \tau + 1$. By using the *triangle inequality* and the boundedness of y_{k_τ} and $\hat{y}_{k_\tau}$, we can derive that

$$
\begin{aligned}
&|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S) - \widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \leq |\widehat{T}^j_{1n} - T^j_{1n}| + |\widehat{T}^j_{2n} - T^j_{2n}| + 2|\widehat{T}^j_{3n} - T^j_{3n}| \\
&= |(n)_2^{-1} \sum_{(k,l)\in I_2^n} [\hat{y}_{k_\tau}\hat{y}_{l_\tau} - y_{k_\tau}y_{l_\tau}] k_1(Z_k - Z_l)k_2(X_{jk} - X_{jl})| + \\
&|(n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n}^{n} [\hat{y}_{k_\tau}\hat{y}_{l_\tau} - y_{k_\tau}y_{l_\tau}] k_1(Z_k - Z_l)k_2(X_{jh} - X_{jq})| + \\
&2|(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [\hat{y}_{k_\tau}\hat{y}_{h_\tau} - y_{k_\tau}y_{h_\tau}] k_1(Z_k - Z_h)k_2(X_{jk} - X_{jl})| \\
&\leq (n)_2^{-1} \sum_{(k,l)\in I_2^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k - Z_l)k_2(X_{jk} - X_{jl}) + \\
&(n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k - Z_l)k_2(X_{jh} - X_{jq}) + \\
&2(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [|\hat{y}_{k_\tau}(\hat{y}_{h_\tau} - y_{h_\tau})| + |y_{h_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k - Z_l)k_2(X_{jk} - X_{jl}).
\end{aligned}
$$

For the first part:

$$
\begin{aligned}
&(n)_2^{-1} \sum_{(k,l)\in I_2^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k - Z_l)k_2(X_{jk} - X_{jl}) \\
&\leq K\{(n)_2^{-1} \sum_{(k,l)\in I_2^n} |\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + (n)_2^{-1} \sum_{(k,l)\in I_2^n} |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|\} \\
&\leq K\{C_0(n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{l_\tau} - y_{l_\tau})| + C_0(n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{k_\tau} - y_{k_\tau})|\} \\
&= 2KC_0(n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{l_\tau} - y_{l_\tau})|.
\end{aligned}
$$

For the second part:

$$
\begin{aligned}
&(n)_4^{-1}\sum_{(k,l,h,q)\in I_4^n}[|\hat{y}_{k_\tau}(\hat{y}_{l_\tau}-y_{l_\tau})|+|y_{l_\tau}(\hat{y}_{k_\tau}-y_{k_\tau})|]k_1(Z_k-Z_l)k_2(X_{jh}-X_{jq})\\
&\le 2KC_0(n)_4^{-1}\sum_{(k,h,q)\in I_3^n}|\hat{y}_{k_\tau}-y_{k_\tau}|.
\end{aligned}
$$

For the third part:

$$
\begin{aligned}
&2(n)_3^{-1}\sum_{(k,l,h)\in I_3^n}[|\hat{y}_{k_\tau}(\hat{y}_{h_\tau}-y_{h_\tau})|+|y_{h_\tau}(\hat{y}_{k_\tau}-y_{k_\tau})|]k_1(Z_k-Z_l)k_2(X_{jk}-X_{jl})\\
&\le 2K\{(n)_3^{-1}\sum_{(k,l,h)\in I_3^n}[|\hat{y}_{k_\tau}(\hat{y}_{h_\tau}-y_{h_\tau})|+2(n)_3^{-1}\sum_{(k,l,h)\in I_3^n}[|\hat{y}_{h_\tau}(\hat{y}_{k_\tau}-y_{k_\tau})|]\}\\
&=2KC_0(n)_3^{-1}(\sum_{(k,h,l)\in I_3^n}|\hat{y}_{h_\tau}-y_{h_\tau}|+\sum_{(k,l)\in I_2^n}^{n}|(\hat{y}_{k_\tau}-y_{k_\tau})|).
\end{aligned}
$$

If we combine the above three parts together, we get

$$
\begin{aligned}
&|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau,X_j|\boldsymbol{X}_S)-\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau,X_j|\boldsymbol{X}_S)|\\
&\le 4KC_0(n)_2^{-1}\sum_{(k,l)\in I_2^n}|(\hat{y}_{l_\tau}-y_{l_\tau})|+4KC_0(n)_3^{-1}\sum_{(k,h,l)\in I_3^n}^{n}|\hat{y}_{h_\tau}-y_{h_\tau}|\\
&=\frac{8KC_0}{n}\sum_{l=1}^{n}|(\hat{y}_{l_\tau}-y_{l_\tau})|.
\end{aligned}
$$

By Proposition 1, we have:

$$
P(\frac{8C_0}{n}\sum_{k=1}^{n}|(\hat{y}_{l_\tau}-y_{l_\tau})|*Z\ge\epsilon)=P(\frac{1}{n}\sum_{k=1}^{n}|(\hat{y}_{l_\tau}-y_{l_\tau})|\ge\frac{\epsilon}{8KC_0})\le 3\exp(-2nc_1\epsilon^2).
$$

Consequently, in view of (3.26), we have that

$$
\begin{aligned}
&P(|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau,X_j|\boldsymbol{X}_S)-\mathrm{CMD}^2_{\mathcal{H}}(Y_\tau,X_j|\boldsymbol{X}_S)|\ge 2\epsilon)\\
\le&P(|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau,X_j|\boldsymbol{X}_S)-\mathrm{CMD}^2_{\mathcal{H}}(Y_\tau,X_j|\boldsymbol{X}_S)|\ge\epsilon)+\\
&P(|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau,X_j|\boldsymbol{X}_S)-\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau,X_j|\boldsymbol{X}_S)|\ge\epsilon)\\
\le&C[\exp\{-c_1\epsilon^2n^{1-2\gamma}\}+n\exp(-c_2n^\gamma)],
\end{aligned}
$$

for a sufficiently small $\epsilon>0$ and some positive constant c_1 and c_2. The analysis of the denominator of $\widehat{\mathrm{CMC}}^2_{\mathcal{H}}(\hat{Y}_\tau,X_j|\boldsymbol{X}_S)$ will generate a similar form of the convergence rate. Therefore, if we set $\epsilon=cn^{-\kappa}$, where κ satisfies $0<\kappa+\gamma<1/2$, we have

$$
\begin{aligned}
&P\{\max_{1\le j\le p-d_1}|\widehat{\omega}_j(\hat{Y}_\tau)-\omega_j(\hat{Y}_\tau)|\ge cn^{-\kappa}\}\\
&\le p\max_{1\le j\le p-d_1}P\{|\widehat{\omega}_j(\hat{Y}_\tau)-\omega_j(\hat{Y}_\tau)|\ge cn^{-\kappa}\}\\
&\le O((p-d_1)[\exp\{-c_1n^{1-2(\kappa+\gamma)}\}+n\exp(-c_2n^\gamma)]).
\end{aligned}
$$

□

The proofs of Theorems 6-10 in the appendix follow exactly the similar lines of argument as in their corresponding CMD$_{\mathcal{H}}$ or CMC$_{\mathcal{H}}$ versions, we thus omit the details here.

Appendix B: Proofs of Chapter 3

B.1 Proof of Proposition 1

We use $G = \boldsymbol{X\gamma}$ for the ease of the presentation. In the working model, we are using Lasso predicted $\hat{M}$ from internal dataset, we start with the basic inequalities in Lasso:

$$\begin{aligned}&\|Y - \hat{M}\hat{\theta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \|Y - \hat{M}\theta^* - \boldsymbol{X}\boldsymbol{\beta^*}\|_2^2/n + \lambda\|\boldsymbol{\beta^*}\|_1\\ &\|G\theta^* + \boldsymbol{X}\boldsymbol{\beta^*} + \delta\theta^* + \varepsilon - \hat{M}\hat{\theta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \|(G - \hat{M})\theta^* + \delta\theta^*\|_2^2/n + \lambda\|\boldsymbol{\beta^*}\|_1\end{aligned}$$

$$\begin{aligned}&\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}) + (G - \hat{M})\theta^* + \delta\theta^* + \epsilon\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}\|_1\\ &\leqslant \|(G - \hat{M})\theta^* + \delta\theta^* + \varepsilon\|_2^2/n + \lambda\|\boldsymbol{\beta^*}\|_1\end{aligned}$$

$$\begin{aligned}&\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n + 2(\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}))^\top((G - \hat{M})\theta^* + \delta\theta^* + \varepsilon)/n\\ &\leqslant \lambda\|\boldsymbol{\beta^*}\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1.\end{aligned}$$

$$\begin{aligned}&\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n \qquad (27)\\ &\leqslant 2(\hat{M}(\hat{\theta} - \theta^*) + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}))^\top((G - \hat{M})\theta^* + \delta\theta^* + \varepsilon)/n + \lambda\|\boldsymbol{\beta^*}\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1. \qquad (28)\end{aligned}$$

We start looking at the right hand side of the (28). Since ε is not independent of $\hat{M}$,

$$\begin{aligned}\hat{M}^\top(\varepsilon + \delta\theta^*)/n &= (\boldsymbol{X}\hat{\boldsymbol{\gamma}})^\top(\varepsilon + \delta\theta^*)/n \qquad (29)\\ &= (\boldsymbol{X}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}))^\top(\varepsilon + \delta\theta^*)/n + (\boldsymbol{X\gamma})^\top(\varepsilon + \delta\theta^*)/n \qquad (30)\end{aligned}$$

Now, $(\boldsymbol{X}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}))^\top(\varepsilon + \delta\theta^*)/n \le |\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}|_1|\boldsymbol{X}^\top(\varepsilon + \delta\theta^*)/n|_\infty$

On set $S := \{\max_{1<i<p} 2|\boldsymbol{X}_i^\top(\varepsilon + \delta\theta^*)|/n \quad \vee \quad |(\boldsymbol{X\gamma})^\top(\varepsilon + \delta\theta^*)|/n \le \lambda_{01}\}$, as ε and δ from bivariate normal distribution, $\varepsilon + \delta\theta^*$ follows a normal distribution, we have $2|(\varepsilon + \delta\theta^*)^\top\hat{M}(\hat{\theta} - \theta^*)/n| \leqslant \lambda_{01}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1|\hat{\theta} - \theta^*| + \lambda_{01}|\hat{\theta} - \theta^*|$, and $2|(\varepsilon + \delta\theta^*)^\top\boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})/n| \leqslant \lambda_{01}\|\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}\|_1$. The above inequality holds with high probability if $\lambda_{01} \asymp \sqrt{\log p/n}$.

For the remaining component, we have

$$\begin{aligned}(\hat{M}(\hat{\theta} - \theta^*))^\top(G - \hat{M})\theta^*/n &= (\theta^* - \hat{\theta})\hat{\boldsymbol{\gamma}}^\top\hat{\Sigma}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \qquad (31)\\ &\leqslant |\hat{\theta} - \theta^*|(\|\hat{\Sigma}\hat{\boldsymbol{\gamma}}\|_\infty\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \qquad (32)\\ &\leqslant |\hat{\theta} - \theta^*|O(\|\hat{\boldsymbol{\gamma}}\|_1)\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \qquad (33)\\ &\leqslant |\hat{\theta} - \theta^*|O(s_{0_\gamma})\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \qquad (34)\end{aligned}$$

as $|\hat{\boldsymbol{\gamma}}| = O(s_{0_\gamma})$ from [45].

where $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 = O(\lambda_\gamma * s_{0_\gamma})$ with $\lambda_\gamma \asymp \sqrt{\log p/n}$ in Lasso estimated $\hat{\boldsymbol{\gamma}}$. While

$$\begin{aligned}(\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}))^\top(G - \hat{M})\theta^*/n &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^\top\hat{\Sigma}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) && (35)\\ &\leqslant \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^\top\hat{\Sigma}\|_\infty \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 && (36)\\ &= O(\|\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}\|_1)\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1. && (37)\end{aligned}$$

And $\lambda\|\boldsymbol{\beta^*}\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}\|_1$, we have

$$\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n \leq \lambda\|\hat{\theta} - \theta^*\|_1 + \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}\|_1$$

for some constant $\lambda = O(\lambda_\gamma s_{0_\gamma}^2)$.

We need to show the following to complete the proof:

$$\|(\theta^* - \hat{\theta}, {\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}})\|_1^2 \leqslant \|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n \tag{38}$$

for some constant ϕ. Hence we assume the following compatibility condition. That is, there exist ${\phi_0}^2$ such that for all Δ satisfying $\|\Delta_{S_0^c}\| \leqslant 3\|\Delta_{S_{0_\beta}}\|$, it holds that

$$\Delta^\top \frac{(X\boldsymbol{\gamma}, X)^\top(X\boldsymbol{\gamma}, X)}{n}\Delta/\phi_0^2 \geq \|\Delta_{S_{0_\beta}}\|_2^2.$$

where $\Delta = [(\theta^* - \hat{\theta}, \boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})]$. Now we want to show the above condition still hold when M is replaced with $\hat{M}$. Specifically, there exists $\phi_0'^2$ such that

$$\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/(n\phi_0'^2) \geq \|\Delta_{S_{0_\beta}}\|_2^2. \tag{39}$$

$$\begin{aligned}&\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n\\ &= \|\boldsymbol{X\gamma}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}) + (\hat{M} - \boldsymbol{X\gamma})(\theta^* - \hat{\theta})\|_2^2/n\\ &= \|\boldsymbol{X\gamma}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})(\theta^* - \hat{\theta})\|_2^2/n + 2(\theta^* - \hat{\theta})(\hat{M} - \boldsymbol{X\gamma})^\top\boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})/n +\\ &2(\theta^* - \hat{\theta})(\hat{M} - \boldsymbol{X\gamma})^\top\boldsymbol{X\gamma}(\theta^* - \hat{\theta})/n + \|(\hat{M} - \boldsymbol{X\gamma})(\theta^* - \hat{\theta})\|_2/n,\end{aligned}$$

where the cross product terms can be diminished if s_{0_γ} is $o_p(\sqrt{\lambda_\gamma})$ from (34) and (37). Finally, we have the compatibility condition (39).

Lemma 2. *On set S, if* $\lambda \geq \max(2\lambda_{01})$, *we have* $2\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda(\|\theta^* - \hat{\theta}\|_1 + \|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1)$

Proof: from 28, On S, we have

$$2\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}_{\boldsymbol{G}})\|_2^2/n + 2\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda\|\theta^* - \hat{\theta}\|_1 + \lambda\|\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}\|_1 + 2\lambda\|\boldsymbol{\beta^*}\|_1$$

$$
\begin{aligned}
&2\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}_{\boldsymbol{G},S_{0_\beta}^c}\|_1 \\
&\le \lambda\|\theta^* - \hat{\theta}\|_1 + \lambda\|\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}}\|_1 + 2\lambda\|\boldsymbol{\beta^*}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}_{S_0^c}\|_1 - 2\lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&\le \lambda\|\theta^* - \hat{\theta}\|_1 + \lambda\|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1 + 2\lambda\|\boldsymbol{\beta^*}\|_1 + 2\lambda\|\hat{\boldsymbol{\beta}}_{S_{0_\beta}^c}\|_1 - 2\lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&\le \lambda\|\theta^* - \hat{\theta}\|_1 + \lambda\|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1 + 2\lambda\|{\boldsymbol{\beta^*}}_{S_{0_\beta}}\|_1 - 2\lambda\|\hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1 \\
&\le \lambda\|\theta^* - \hat{\theta}\|_1 + 3\lambda\|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1 \\
&\le 3\lambda(\|\theta^* - \hat{\theta}\|_1 + \|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1),
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&2\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n + \lambda\|(\theta^* - \hat{\theta}, \beta - \hat{\beta})\|_1 \\
&\le 2\|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n + \lambda(\|\theta^* - \hat{\theta}\|_1 + \|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1) + \lambda\|\hat{\boldsymbol{\beta}}_{S_{0_\beta}^c}\|_1 \\
&\le 4\lambda(\|\theta^* - \hat{\theta}\|_1 + \|{\boldsymbol{\beta^*}}_{S_{0_\beta}} - \hat{\boldsymbol{\beta}}_{S_{0_\beta}}\|_1) \\
&\le 4\sqrt{s_0 + 1} * \|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2/\sqrt{(n\phi^2)} \\
&\le \|\hat{M}(\theta^* - \hat{\theta}) + \boldsymbol{X}(\boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_2^2/n + 4\lambda^2(s_0 + 1)/\phi^2
\end{aligned}
$$

Therefore on set S, we obtain $\|(\theta^* - \hat{\theta}, \boldsymbol{\beta^*} - \hat{\boldsymbol{\beta}})\|_1 \le 4\lambda(s_0 + 1)/\phi^2$ for a $\lambda = O(\lambda_\gamma s_{0_\gamma}^2 \vee \sqrt{\log p/n})$.

B.2 Proof of Theorem 1

Let z be the projection of $\hat{M}$ to the $\boldsymbol{X}$,

$$
\boldsymbol{z} = \hat{M} - \boldsymbol{X}\hat{\boldsymbol{b}}, \hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \|\hat{M} - \boldsymbol{X}\boldsymbol{b}\|_2^2/2n + \lambda_z \sum_{i=1}^{p} w_i|b_i|
$$

following debiased lasso estimator, we propose

$$
\begin{aligned}
\tilde{\theta} &= \frac{\boldsymbol{z}^\top y}{\boldsymbol{z}^\top \hat{M}} - \sum_{i=1}^{p} \frac{\boldsymbol{z}^\top \boldsymbol{x}_i \hat{\boldsymbol{\beta}}}{\boldsymbol{z}^\top \hat{M}} \\
&= \frac{\boldsymbol{z}^\top y}{\boldsymbol{z}^\top \hat{M}} - \sum_{i=1}^{p} \frac{\boldsymbol{z}^\top \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{\hat{M}}}{\boldsymbol{z}^\top \hat{M}} - \frac{\boldsymbol{z}^\top \hat{M}\hat{\theta}}{\boldsymbol{z}^\top \hat{M}} + \frac{\boldsymbol{z}^\top \hat{M}\hat{\theta}}{\boldsymbol{z}^\top \hat{M}} \\
&= \hat{\theta} + \frac{\boldsymbol{z}^\top (y - \hat{M}\hat{\theta} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\boldsymbol{z}^\top \hat{M}}
\end{aligned}
$$

$$
\tilde{\theta} - \theta^* = \frac{\boldsymbol{z}^\top \boldsymbol{\epsilon}}{\boldsymbol{z}^\top \hat{M}} + \frac{\boldsymbol{z}^\top \boldsymbol{\delta}\theta^*}{\boldsymbol{z}^\top \hat{M}} + \sum_{i=1}^{p} \frac{\boldsymbol{z}^\top \boldsymbol{x}_i(\beta_i - \hat{\beta}_i)}{\boldsymbol{z}^\top \hat{M}} + \frac{\boldsymbol{z}^\top (\boldsymbol{G} - \hat{M})\theta^*}{\boldsymbol{z}^\top \hat{M}}
$$

Therefore

$$\begin{aligned} &\frac{|\boldsymbol{z}^\top \hat{M}|}{\|\boldsymbol{z}\|_2}(\tilde{\theta}-\theta^*) &(40)\\ &= \frac{\boldsymbol{z}^\top\boldsymbol{\epsilon}}{\|\boldsymbol{z}\|_2} + \frac{\boldsymbol{z}^\top\boldsymbol{\delta}\theta^*}{\|\boldsymbol{z}\|_2} + \frac{1}{\|\boldsymbol{z}\|_2}\sum_{i=1}^{p} \boldsymbol{z}^\top\boldsymbol{x}_i(\beta_i-\hat{\beta}_i) + \frac{1}{\|\boldsymbol{z}\|_2}\sum_{i=1}^{p} \boldsymbol{z}^\top\boldsymbol{x}_i(\boldsymbol{\gamma}_i-\hat{\boldsymbol{\gamma}}_i)\theta^* &(41)\end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{p} \frac{\boldsymbol{z}^\top\boldsymbol{x}_i(\beta_i-\hat{\beta}_{i,})}{\|\boldsymbol{z}\|_2} &\leq \frac{1}{\|\boldsymbol{z}\|_2}\text{max}_{1\leq i\leq p}|\boldsymbol{z}^\top\boldsymbol{x}_i|\|\boldsymbol{\beta^*}-\hat{\boldsymbol{\beta}}\|_1\\ \frac{1}{\|\boldsymbol{z}\|_2}\sum_{i=1}^{p} \boldsymbol{z}^\top\boldsymbol{x}_i(\boldsymbol{\gamma}_i-\hat{\boldsymbol{\gamma}}_i)\theta^* &\leq \frac{1}{\|\boldsymbol{z}\|_2}\text{max}_{1\leq i\leq p}|\boldsymbol{z}^\top\boldsymbol{x}_i|\|\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}\|_1\theta^*\end{aligned}$$

Based on Algorithm 2 of finding penalty λ_z in [46], $\frac{1}{\|\boldsymbol{z}\|_2}\text{max}_{1\leq i\leq p}|\boldsymbol{z}^\top\boldsymbol{x}_i|$ can be bounded by $C\sqrt{logp}$ for some constant C. For the completeness, we present their algorithm below and we define:

$$\begin{aligned} \hat{\boldsymbol{b}}(\lambda_z) &= \underset{\boldsymbol{b}}{\text{argmin}}\|\hat{M}-\boldsymbol{X}\boldsymbol{b}\|_2^2/2n + \lambda_z\sum_{i=1}^{p} w_i|b_i|\\ \boldsymbol{z}(\lambda_z) &= \hat{M}-\boldsymbol{X}\hat{\boldsymbol{b}}(\lambda_z)\\ \eta(\lambda_z) &= \frac{1}{\|\boldsymbol{z}(\lambda_z)\|_2}\text{max}_{1\leq i\leq p}|\boldsymbol{z}(\lambda_z)^\top\boldsymbol{x}_i|\\ \tau(\lambda_z) &= \|\boldsymbol{z}(\lambda_z)\|_2/|\boldsymbol{z}(\lambda_z)^\top\boldsymbol{x}_i|\end{aligned}$$

Algorithm 2 The procedure of computing $\boldsymbol{z}$

```
Input: an upper bound $\eta^*$ for the bias factor, with default value $\eta^* = \sqrt{2\text{log}p}$, tuning
    parameters $\kappa_0 \in [0,1]$ and $\kappa_1 \in (0,1]$;
  1. (very/adjust $\eta$ and compute the corresponding noise factor $\tau$)
     If $\eta(\lambda_z) > \eta^*$ for all $\lambda_z > 0$, $\eta^* \leftarrow (1+\kappa_1)\text{inf}_{\lambda_z>0}\eta(\lambda_z)$;
     $\lambda_z \leftarrow \text{max}\{\lambda_z : \eta(\lambda_z) \leq \eta^*_\lambda\}$, $\eta^* \leftarrow \eta(\lambda_z)$, $\tau^* \leftarrow \tau(\lambda_z)$;
  2. further reduction of the bias factor $\lambda_z \leftarrow \min\{\lambda_z : \tau(\lambda_z) \leq (1+\kappa_0)\tau^*\}$
Output: $\lambda_z$, $\boldsymbol{z} \leftarrow \boldsymbol{z}(\lambda_z)$, $\tau \leftarrow \tau(\lambda_z)$, $\eta \leftarrow \eta(\lambda_z)$
```

Under the null hypothesis $\theta^* = 0$, the bias term in (40) related to $\hat{\boldsymbol{\gamma}}$ is 0, and the left bias $\frac{1}{\|\boldsymbol{z}\|_2}\sum_{i=1}^{p}\boldsymbol{z}^\top\boldsymbol{x}_i(\boldsymbol{\beta^*}-\hat{\boldsymbol{\beta}}) \asymp s_0\sqrt{\log p}\lambda$. If we have $s_0 = o(\sqrt{m}/(s_{0_\gamma}^2\log p) \wedge$

$\sqrt{n}/\log p)$, then $\frac{|\boldsymbol{z}^\top \hat{M}|}{\|\boldsymbol{z}\|_2}\tilde{\theta} \rightarrow N(0,\sigma_\varepsilon^2)$. Now if we replace σ_ε with the consistent estimator $\hat{\sigma}_\varepsilon$, we have $\frac{|\boldsymbol{z}^\top \hat{M}|}{\|\boldsymbol{z}\|_2\hat{\sigma}_\varepsilon}\tilde{\theta} \rightarrow N(0,1)$.

B.3 Proof of Propositioin 2

In the proof of Theorem 1. We have the following requirement of the consistency of $\hat{\boldsymbol{\gamma}}$. Now consider

$$M = \boldsymbol{X}\boldsymbol{\gamma} + \delta \tag{42}$$

$$M' = \boldsymbol{X}'\boldsymbol{\gamma}' + \delta, \tag{43}$$

where (43) is from a external dataset. In the previous proof, we have $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \asymp s_{0_\gamma}\sqrt{\log p/m}$, now let us study the bound of $\|\hat{\boldsymbol{\gamma}'} - \boldsymbol{\gamma}\|_1$, where

$$\begin{aligned}
&\|\hat{\boldsymbol{\gamma}'} - \boldsymbol{\gamma}\|_1 \\
&= \|\hat{\boldsymbol{\gamma}'} - \boldsymbol{\gamma}' + \boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_1 \\
&\leq \|\hat{\boldsymbol{\gamma}'} - \boldsymbol{\gamma}'\|_1 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_1 \\
&= O(s_{0_{\gamma'}}\sqrt{\log p/m}) + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_1.
\end{aligned}$$

Therefore, if $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_1 \lesssim s_{0_{\gamma'}}\sqrt{\log p/m}$ and $s_{0_{\gamma'}} \lesssim s_{0_\gamma}$, the result in Proposition 1 and Theorem 1 still holds.

Bibliography

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[2] E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

[3] J. Chang, C. Y. Tang, and Y. Wu. Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, 41(4):2123–2148, 2013.

[4] J. Chen, W. Chen, N. Zhao, M. C. Wu, and D. J. Schaid. Small sample kernel association tests for human genetic and microbiome association studies. *Genetic epidemiology*, 40(1):5–19, 2016.

[5] M.-Y. Cheng, T. Honda, J. Li, and H. Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics*, 42(5):1819–1849, 2014.

[6] H. Cui, R. Li, and W. Zhong. Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641, 2015. PMID: 26392643.

[7] L. Dowsett, E. Higgins, S. Alanazi, N. A. Alshuwayer, F. C. Leiper, and J. Leiper. Adma: a key player in the relationship between vascular dysfunction and inflammation in atherosclerosis. *Journal of Clinical Medicine*, 9(9):3026, 2020.

[8] 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. PMID: 22279246.

[9] 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.

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

[11] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[12] Y. Fan, Y. Kong, D. Li, and J. Lv. Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*, 2016.

[13] E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology*, 4(2):293–305, 2019.

[14] X. Han. Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data. *The Annals of Statistics*, 47(4):1995–2022, 2019.

[15] N. Hao and H. H. Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014.

[16] X. He, L. Wang, and H. G. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369, 2013.

[17] Y. Kong, D. Li, Y. Fan, and J. Lv. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics*, 45(2):897–922, 2017.

[18] J. M. Lee, S. Zhang, S. Saha, S. Santa Anna, C. Jiang, and J. Perkins. Rna expression analysis using an antisense bacillus subtilis genome array. *Journal of bacteriology*, 183(24):7371–7380, 2001.

[19] G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877, 2012.

[20] R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

[21] W. Lin, R. Feng, and H. Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288, 2015.

[22] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, et al. Multiomics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019.

[23] S. Ma, R. Li, and C. L. Tsai. Variable screening via quantile partial correlation. *Journal of the American Statistical Association*, 112(518):650–663, 2017. PMID: 28943683.

[24] Q. Mai and H. Zou. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234, 2013.

[25] Q. Mai and H. Zou. The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497, 2015.

[26] E. Muller, Y. M. Algavi, and E. Borenstein. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj Biofilms and Microbiomes*, 8(1):79, 2022.

[27] D. Nandy, F. Chiaromonte, and R. Li. Covariate information number for feature screening in ultrahigh-dimensional supervised problems. *Journal of the American Statistical Association*, 0(0):1–14, 2021.

[28] W. Pan, X. Wang, W. Xiao, and H. Zhu. A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114(526):928–937, 2019. PMID: 31692981.

[29] E. Şenateş. Chemoprevention of barrett's esophagus and adenocarcinoma. In *Barrett's Esophagus*, pages 189–204. Elsevier, 2016.

[30] R. J. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[31] X. Shao and J. Zhang. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318, 2014.

[32] R. Song, F. Yi, and H. Zou. On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 24(4):1735–1752, 2014.

[33] 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.

[34] Y. Tian and Y. Feng. Rase: A variable screening framework via random subspace ensembles. *Journal of the American Statistical Association*, 0(0):1–12, 2021.

[35] Z. Tong, Z. Cai, S. Yang, and R. Li. Model-free conditional feature screening with fdr control. *Journal of the American Statistical Association*, pages 1–13, 2022.

[36] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. 2014.

[37] S. Vermeire, G. Van Assche, and P. Rutgeerts. Laboratory markers in ibd: useful, magic, or unnecessary toys? *Gut*, 55(3):426–431, 2006.

[38] H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

[39] S. Wang, T. H. McCormick, and J. T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

[40] C. Wen, W. Pan, M. Huang, and X. Wang. Sure independence screening adjusted for confounding covariates with ultrahigh dimensional data. *Statistica Sinica*, 28(1):293–317, 2018.

[41] H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[42] C. Xu and J. Chen. The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, 109(507):1257–1269, 2014. PMID: 25382886.

[43] S. Yachida, S. Mizutani, H. Shiroma, S. Shiba, T. Nakajima, T. Sakamoto, H. Watanabe, K. Masuda, Y. Nishimoto, M. Kubo, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature medicine*, 25(6):968–976, 2019.

[44] B. Yu. Stability. *Bernoulli*, 19(4):1484 – 1500, 2013.

[45] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. 2008.

[46] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

[47] S. Zhang and Y. Zhou. Variable screening for ultrahigh dimensional heterogeneous data via conditional quantile correlations. *Journal of Multivariate Analysis*, 165:1–13, 2018.

[48] S. D. Zhao and Y. Li. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411, 2012.

[49] T. Zhou, L. Zhu, C. Xu, and R. Li. Model-free forward screening via cumulative divergence. *Journal of the American Statistical Association*, 115(531):1393–1405, 2020.

[50] L. P. Zhu, L. Li, R. Li, and L. X. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011. PMID: 22754050.

Vita

Lei Fang

Education

- **M.S. in Statistics**, University of Kentucky May 2020
- **M.S.E in Civil Engineering**, Arizona State University June 2016
- **B.S. in Environmental Engineering**, Ocean Unviersity of China June 2014

Experience

- Teaching Assistant (including primary instructor), Department of Statistics, University of Kentucky, 2018 - 2023

Publications

- **Fang, L.**, Yuan,Q., Ye,C., and Yin, X. ”Variable Screening via Conditional Martingale Difference Divergence”. ***Statistic Sinica***, invited for revision.
- **Fang, L.**, Ye,C., Zhou,M. ”Quantile Restricted Mean Survival Time and Empirical Likelihood”. 2023+.
- **Fang, L.**, Ye,C., Wang,Y. ”Integration of multiview microbiome data for causal discoveries of complex trait-metabolite associations”. 2023+.