Robust Statistical Approaches Dealing with High-Dimensional Observational Data

Huichen Zhu

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Robust Statistical Approaches Dealing with High-Dimensional Observational Data

Huichen Zhu

The theme of this dissertation is to develop robust statistical approaches for the high-dimensional observational data. The development of technology makes data sets more accessible than any other time in history. Abundant data leads to numerous appealing findings and at the same time, requires more thoughtful efforts. We are encountered many obstacles when dealing with high-dimensional data. Heterogeneity and complex interaction structure rule out the traditional mean regression method and expect a novel approach to circumvent the complexity and obtain significant conclusions. Missing data mechanism in high-dimensional data is complicated and is hard to manage with existing methods. This dissertation contains three parts to tackle these obstacles: (1) a tree-based method integrated with the domain knowledge to improve prediction accuracy; (2) a tree-based method with linear splits to accommodate the large-scale and highly correlated data set; (3) an integrative analysis method to reduce the dimension and impute the block-wise missing data simultaneously.

In the first part of the dissertation, we propose a tree-based method called conditional quantile random forest (CQRF) to improve the screening and intervention of the onset of mentor disorder incorporating with rich and comprehensive electronic medical records (EMR). Our research is motivated by the REactions to Acute Care and Hospitalization (REACH) study, which is an ongoing prospective observational cohort study of the patient with symptoms of a suspected acute coronary syndrome (ACS). We aim to develop a robust

and effective statistical prediction method. The proposed approach fully takes the population heterogeneity into account. We partition the sample space guided by quantile regression over the entire quantile process. The proposed CQRF can provide a more comprehensive and accurate prediction. We also provide theoretical justification for the estimate quantile process.

In the second part of the dissertation, we apply the proposed CQRF to REACH data set. The predictive analysis derived by the proposed approach shows that for both entire samples and high-risk group, the proposed CQRF provides more accurate predictions compared with other existing and widely used methods. The variable importance scores give a promising result based on the proposed CQRF that the proposed importance scores identify two variables which have been proved to be critical features by the qualitative study. We also apply the proposed CQRF to Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study data set. We show that the proposed approach improves the personalized medicine recommendation compared with existing treatment recommendation method. We also conduct two simulation studies based on the two real data sets. Both simulation studies validate the consistent property of the estimated quantile process.

In the second part, we also extend the proposed CQRF with univariate splits to linear splits to accommodate a large number of highly correlated variables. Gene-environment interaction is a widely concerned topic since the traits of complex disease is always difficult to understand, and we are eager to find interventions tailored to individual genetic variations. The proposed approach is applied to a Breast Cancer Family Registry (BCFR) study data set with body mass index (BMI) as the response variable, several nutrition intake factors, and genotype variables. We aim to figure out what kind of genetic variations affect the heterogeneous effect of the environmental factors on BMI. We devise a criterion which measures the relationship between the response variable and gene variants conditioning on the environmental factor to determine the optimal linear combination split. The variable importance score is also calculated by summing up the criterion across all splits in the

random forest. We show in the results that top-ranked genes prioritized by the proposed importance scores make the effect of the environmental factors on BMI differently.

In the third part, we introduce an integrative analysis approach called generalized integrative principal component analysis (GIPCA). The heterogeneous data types and the presence of block-wise missing data pose significant challenges to the integration of multi-source data and further statistical analyses. There is not literature can easily accommodate data of multiple types with block-wise missing structure. The proposed GIPCA is a low-rank method which conducts the dimension reduction and imputation of block-wise missing data simultaneously to data with multiple types. Both simulation study and real data analysis show that the proposed approach achieves good missing data imputation accuracy and identifies some meaningful signals.

Key Words: Quantile Regression; Random Forest; Post-Trauma Stress Disorder; Personalized Medicine; Gene-Environmental Interaction; Block-wise Missing Imputation, Exponential Family; Exponential Principal Component Analysis; Joint and Individual Variation Explained; Multi-view Data.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation concludes five years of doctoral study and research. I am deeply grateful to many extraordinary scholars, who helped me and guided me through the journey of my five-year research.

Foremost, I would like to thank my two advisors, Professor Ying Wei, and Professor Gen Li. I started doing research with Professor Ying Wei during my second year, and with Professor Gen Li during my third year. They both are knowledgeable and generously share their knowledge from both theoretical and practical point of view, which helped me much to obtain a big picture of my research work. I am so thankful for their patience and motivation, especially at the beginning of work. Without their encouragement and inspiring advice, I could not come this far. I would also like to thank my collaborator Professor Eric F. Lock in the University of Minnesota. With his invaluable suggestion and perceptive suggestions, my research work becomes a more perfect one.

I would also like to express my sincere gratitude to my dissertation committee member, Professor Ying Kuen Cheung, Professor Yifei Sun, and Professor Ian Kronish, who offered valuable advice and insightful suggestion to improve this thesis. A special thanks to Professor Ian Kronish, for providing the really interesting data set which inspired the framework of my first two projects.

Last but not least, I thank my family members and friends for their love and support.

To my parents and my husband

Chapter 1

Introduction

1.1 Background and Overview

This century is the century of *data*. With advances in data collection and technological development, data acquisition becomes much easier and cheaper. Observational data sets tend to consist of a large number of variables, where hyper-informative details are collected about each subject. In a data set, a single observation contains thousands of variables while there are only a limited number of observations available. We encounter different obstacles when analyzing high-dimensional data sets in various fields. For example, complex interaction mechanism and heterogeneity are significant challenges when we analyze the clinical data set. Complicated missing structure and multiple data types in multi-source high-dimensional data are issues we can not avoid. Classical statistical methods are not designed to cope with data sets with such high dimensionality and perplexing structures. More efforts are needed to develop robust approaches to handle such data sets.

Clinical medicine relies on strong research evidence to validate the efficacy of clinical practices or clinical care. However, large-scale randomized clinical trials are expensive and unfeasible for most of the time. The existence of expansive electronic health record

(EHR) data helps clinicians and investigators to find clinical evidence and supports for medical decisions. EHR data sets are always with high dimension variables, which contains a patient's medical history, diagnoses, medications, treatment plans, etc. Incorporating with EHR data, feasible strategies would be able to make decisions for individual patients based on data from similar patients and similar scenarios. To do so, heterogeneity and complex interaction structure contained in the high-dimensional EHR data are inevitable issues. Traditional statistical methods such as regression approaches are unfeasible to handle large-scale EHR data sets, and mean-based approaches are not able to consider heterogeneity. Hence, we aim to develop an approach which derives robust and accurate inference and prediction when we analyze the EHR data set.

Multi-source data with a large number of dimensions is one particular type of high-dimensional data. Such data sets are encountered in many fields. Despite recent developments on the integrative dimension reduction of such data, most existing methods cannot easily accommodate data of multiple types (e.g., binary or count-valued). Moreover, multi-source data often have block-wise missing structure, i.e., data in one or more sources may be entirely unobserved for a sample. The various data types and the presence of block-wise missing data pose significant challenges to the integration of multi-source data and further statistical analyses. For instance, administrative data is a type of data which has not been aware of recently but is worth to dig into it. Administrative data set always comes in the form of fragments from different government agencies, which leads to block-wise missing structures. Thus, we aim to develop a low-rank method which conducts the dimension reduction and imputation of multi-source block-wise missing data simultaneously and allows different data types for various sources.

1.2 Introduction to Conditional Quantile Random Forest

Our research is motivated by the REactions to Acute Care and Hospitalization (REACH) study, which is an ongoing prospective observational cohort study of the patients admitted

Columbia-New York Presbyterian Hospital with symptoms of a suspected acute coronary syndrome(ACS). For many, ACS is a frightening experience associated with significant pain and the fear of dying (Whitehead et al., 2005; von Känel et al., 2011). In such patients, ACS can induce long-lasting symptoms of Post-Traumatic Stress Disorder (PTSD) (Abbas et al., 2009), a psychiatric disorder that results in substantial psychological distress and impaired functioning. Besides, ACS-induced PTSD has been shown to increase the risk of adverse prognosis. Given the adverse psychological impact of ACS-induced PTSD and its potential to increase cardiovascular risk, there is an urgent need to develop prediction and intervention tools to prevent the onset of PTSD in ACS survivors.

Therapies for preventing PTSD are usually labor-intensive and costly and it would be infeasible to deliver them to all ACS survivors. Thus, developing strategies to predict PTSD and potential intervention efficacy for individual patients is essential and will make therapies more cost-effective. Prior efforts to predict PTSD have been limited to regression models with cross-sectional data (Brewin et al., 2000; Ozer et al., 2003; Meli et al., 2017; Edmondson, 2014). The REACH study includes rich and comprehensive electronic medical records (EMR) of the enrolled patients with ACS symptoms, as well as a rich set of self-reported data from baseline interviews, and all the patients were followed up at one month to assess their PTSD symptom. Figure 1.1 displays the histogram of one-month PSTD scores among the REACH patients. As shown in the figure, the PTSD score ranges from 17 to 89, and its distribution is highly skewed. In a preliminary analysis, we used a mean-based random forest and regression model to predict the one-month PTSD based on the individual EMR and baseline interview. We found that both approaches have poor prediction accuracy for the high-risk PTSD patients (See details in Table 3.1 in Chapter 3). Hence, we aim to develop a robust and effective statistical prediction method to improve the screening and prevention approach in clinical and public health practice.

Decision trees are classical statistical learning methods that recursively partition the sample space into mutually exclusive sub-spaces with the distinctive means of an outcome of interest. Through visualization, trees provide a simple tool for screening and treatment

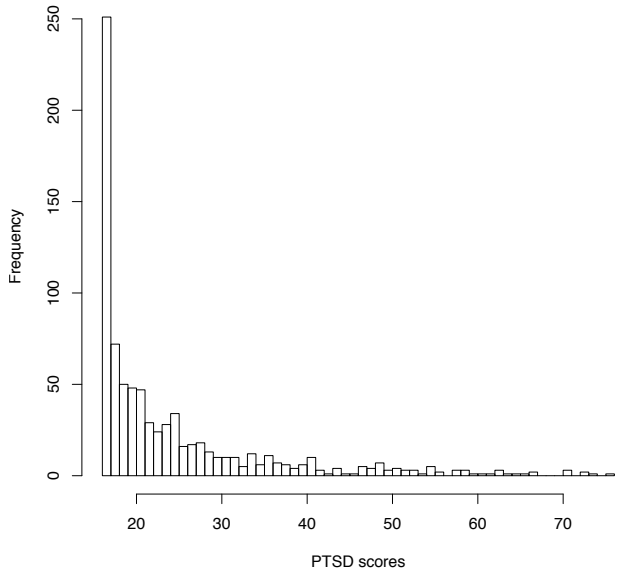


Figure 1.1: Histogram of Patients' PTSD score

decisions, with many applications in clinical research and public health studies. Unlike traditional regression methods, which assume that a global linear additive structure holds for any subset of the population, the tree methods, by design, can incorporate a large number of potential predictors and uncover complex associations and interactions. These are essential considerations in precision health applications that deal with large amounts of data from EMRs, genomics, and mobile data. Thus, decision trees and their extension to random forests would be advantageous to achieve our goal.

Among all, the first thing we need to consider to construct a tree-based model or a random forest, which is an assembly of trees, is the way how the sample space is partitioned. Several splitting criteria have been introduced to compose a tree-structured model. Both Automatic interaction detection (AID) (Morgan and Sonquist, 1963) and Classification and Regression Tree (CART) (Breiman et al., 1984) partitioned the sample space by choosing the optimal split which minimizes the total sum of squared error (SSE) of the two child nodes. Fast and Accurate Classification Tree (FACT) (Loh and Vanichsetakul, 1988) applied the F ratio from analysis of variance (ANOVA) as a criterion to choose the splitting variable. Conditional Inference Trees (CTree) Hothorn et al. (2006) chose the

splitting variable by the p-values of a permutation test. However, these algorithms identify sub-spaces only depending on distinctive means, which only partially reveal the heterogeneity. Breiman et al. (1984) and Galimberti et al. (2011) considered least absolute deviation (LAD) and M-estimates to partition the sample space. More recently, Chaudhuri and Loh (2002) and Bhat et al. (2015) proposed to partition the space by minimizing the total sum of the quantile losses at a specific quantile level. The quantile loss is a counterpart of SSE concerning a quantile level. Prior information is required to determine the quantile level in those methods. Otherwise, it may result in an inadequate consideration of heterogeneous. Alternative splitting criterion, which gives sufficient considerations of heterogeneity, is desired to partition the sample space by exploring the entire distributions.

The causes of many diseases are complicated. Many clinical papers confirmed that there are certain features which directly affect the disease traits. For example, Edmondson (2014) stated that the fear of death underlies both past and future induces the symptoms of PTSD. How to incorporate this domain knowledge and how to measure their heterogeneous effects on different subjects are necessary to be considered. Chen et al. (2007) proposed a partially linear tree-based regression (PLTR) model by modeling with a combination of linear main effects and a non-parametric tree structure. Though PLTR can be applied to integrate the domain knowledge in the model structure as linear main effects, it only deals with the homogeneous main effects across the whole population and various intercepts among different subpopulations. Methods exploring the heterogeneous treatment effect by adopting the tree-structured model (Wager and Athey, 2017; Su et al., 2009) are other possibilities handling the domain knowledge. However, those methods can not be generalized to other features, especially continuous ones.

Random forests are typically used to produce point predictions only without taking the information about how far those predictions are away from the true ones. Limited literature works on providing quantitative assessments of predictions. Meinshausen (2006) proposed a quantile regression forest (QRF) which infers the full conditional distribution of the response variable. The prediction intervals can be obtained from the conditional

distributions. (Zhang et al., 2019) also proposed prediction intervals based on the empirical distribution of out-of-bag prediction errors. However, the partition algorithm involving in both approaches still follow the classic CART using SSE as the splitting criterion. The heterogeneous of the entire conditional distributions are ignored while the optimal splits are determined. Though these approaches provide more informative and accurate predictions, it may fail to be adapted to the most vulnerable subpopulations.

Although there is a vast literature on decision trees and their extension to random forests (i.e., ensembles of trees to robustify and improve predictions), most algorithms partition the sample space according to various means. The most vulnerable and high-risk groups for specific diseases are often patients with relatively high (or low) values in their biomarkers and phenotype values. Hence, mean-based partitioning may not be useful in identifying the most vulnerable subpopulations, the patients with phenotype values in the tail ends of their distributions. We propose developing a new regression tree framework based on quantile regression (Koenker and Bassett Jr, 1978) that provides a systematic strategy for examining how covariates influence various quantiles of the outcome. Also, quantile regression is robust against outliers and automatically adapts to skewness and heteroscedasticity. Hence, it would be a suitable modeling tool for accurate and robust predictions in health-related studies, since health variables are rarely nicely distributed, often contain outliers, missing data and measurement errors, and subject to complex associations.

We develop a robust and efficient approach for improving the screening and intervention strategies. We use the quantile regression (Koenker and Bassett Jr, 1978) over the entire quantile process to guide the sample partitions at each tree node, which distinguishes the proposed work from the existing approaches in random forests. Using the conditional quantile process for sample partition has the following advantages. Firstly, it complements the mean-based approaches and fully takes the population heterogeneity into account. The investigation of the real data shows that the proposed approach significantly improves prediction for patients at high-risk. Secondly, the use of conditional quantile regression at each split also provides a convenient way to incorporate domain knowledge to improve prediction

and to include treatment assignment to allow direct estimation of individualized treatment effect for precision care. Thirdly, existing methods provide the prediction based on the conditional means or the conditional quantile at a fixed level, whereas incorporating with domain knowledge, the proposed approach provides the prediction based on the conditional quantile process to enhance the accuracy of the prediction.

Precision medicine is another topical issue which has been widely discussed in many clinical fields Treatment responses/effects are often heterogeneous across individuals due to their different genetic profiles, medical histories, co-morbidity conditions, and other related factors. A universe treatment strategy which treats all the patients with the same treatment is not adequate. Thus, precision medicine, which makes the treatment decision individualized, becomes one of the future directions in medical science. The goal of precision medicine is to optimize treatment decisions that take into account patient-level information, like their medical records, baseline measurements, etc.

Tremendous research efforts have been devoted to estimating optimal individualized treatment rules (ITRs) (Lavori and Dawson, 2004). Several approaches estimate the optimal ITR by maximizing the value function through augmented inverse probability weighting (Zhang et al., 2012, 2013). Others, like Zhao et al. (2012), maximize the value function directly by using the support vector machine. There are several approaches which are developed to determine individualized treatment rules (ITRs) using tree-based methods. For example, the virtual twins approach (Foster et al., 2011) estimated individualized treatment effect by utilizing a tree-based model. Laber and Zhao (2015) proposed Minimum Impurity Decision Assignments decision trees (MISDAs) to find an optimal ITR, where the partition rules at each split are to maximize an updated value function incorporating with a purity measurement. As a single tree is subject to high variability, Doubleday et al. (2018) further extended the tree algorithm to a random forest. Each tree recommends one treatment based on an individual patient's profiles, and the treatment that receives the most votes among trees will be the optimal treatment.

Almost all the traditional ITR approaches provide a hard decision that which treatment

the patient should be assigned to is determined. The random forest approach proposed by Doubleday et al. (2018) gives a soft probability, which clinician could depend on to decide along with their own knowledge and experience. However, probability sometimes is not enough when we also care about the treatment effects in practice. If the treatment indicator variable takes the role of the domain knowledge, the proposed CQRF is capable of providing individualized treatment recommendation by the estimated/predicted conditional distributions. Several criteria are proposed to assess the overall treatment effect derived from the conditional distributions among different treatment groups. Both real data and simulation data analyses show the promising results derived from the proposed CQRF.

1.3 Introduction to CQRF with Linear Splits

The genetic influences on disease traits do not only depend on the effect of genetic variants but also the impact of environmental factors, as well as their interaction effects. Gene-environment (G E) is defined as the different influences of genotypes on phenotype under different environment circumstances (Falconer, 1952). G E interactions play critical roles in figuring out disease risk at the individual level. Study of G E interactions helps us to identify the heterogeneity effect of genotypes under different environmental conditions. It also enables us to specify environmental factors which have various effects on subpopulations with different gene variations. Furthermore, understanding gene-environment interactions allows us to give individualized preventive advice before the onset of disease, and offer personalized medicine practice after the onset of disease.

Usually, a genetic data set contains a large number of SNPs. A conventional approach to screen a massive number of SNPs is to conduct univariate association tests on each SNP. The tests' results (i.e., p-values) prioritize SNPs for further study. For example, SNPs are ranked by p-values of some test. Variables with smaller p-values are retained in the data to be analyzed. There are two categories for the statistical models involving a statistical test of G E interactions: one is model-based (Guo, 2000; Chatterjee and Carroll, 2005; Kraft

et al., 2007; Maity et al., 2009), and the other one is model-free (Hahn et al., 2003). Most traditional model-based analysis of G E interactions have strong assumptions between the effect of genetics and environmental factors. For example, a large amount of literature under the regression-based framework assume linear interaction effect. They assume a simple underlying model with interaction term like,

$$Y = {}_0 + {}_1 E + {}_1 G + {}_2 GE + ";$$

where Y is the measurement of disease trait; ${}_0$ is the overall means; ${}_1$ and ${}_1$ is the marginal effect of environmental factors E and genetic variables G correspondingly; ${}_2$ is the interaction effect for G E and $"$ is error term with mean 0 and variance ^{2}. Such strong model assumption, which restricts the relationship between disease traits and environmental factors to be linear, is not feasible under the situations of complex genetic mechanism. Non-parametric modeling (Hahn et al., 2003) is an alternative approach which relief the strong model assumption. However, by applying such an approach, it is hard to interpret the relationship between the response variable of interest and the independent variables.

To address the limitation of the linear model assumption, Ma et al. (2011) introduced a varying coefficient model which replaces the linear G E interaction by a smooth non-linear interaction. The linear model becomes,

$$Y = (E) + (E)G + (X)";$$

where (E) is a smooth function in environmental factors E. The varying coefficient model allows the effect of a gene varied as a function of environmental factors. Several statistical tests are proposed to dissect the G E interactions. Regression model-based approaches assume the homogeneity effect of gene variations conditioning on the environmental factors. However, sometimes the conditional homogeneous assumption is impracticable that the average conditional effect might not be able to be generalized to all population. Furthermore, regression model-based approaches always use univariate test to screen SNPs,

which scarifies computation efficiency if we have thousands of SNPs.

Our proposed CQRF allows the heterogeneous effects of the environmental factors across different quantile levels. The previous proposed CQRF only allows univariate splits in the tree-based algorithm. The univariate split makes the algorithm less efficient if there exists a large number of variables, primarily when they are clustered by some common attributes. For example, genetic linkage makes genotypes of a group of SNPs which are physically near to each other highly correlated. It is better to consider a combination of those SNPs as a potential splitting variable. Considering one SNP a time to partition the sample space is not computationally expedient. To improve our proposed tree-based approach adapted to high-dimensional data, we aim to replace the univariate splits with linear splits. We come up with a conditional quantile correlation, which measures conditional dependence between the response variable and the linear combination of the variables. By maximizing the absolute value of the conditional quantile correlation, we can obtain an optimal linear combination.

Variable importance score is a by-product of the random forest, which measures the influence of each splitting variable within the random forest. It can be utilized to rank the splitting variables as a screening and filtering method like p-values in model-based approaches. The abilities to handle heterogeneity and to accommodate a large number of variables make the random forest a well-suited approach to address the issue with large scale association study. The proposed variable importance score is calculated by averaging the proposed absolute value of the conditional quantile correlation for each variable in each node. The variable importance scores can help us to identify those genes which cause the various effects of the environmental factor.

1.4 Introduction to Integrative Analysis for Block-wise Missing Data

With technological developments, data acquisition becomes more accessible and cheaper. In numerous studies, people collect data from multiple sources on the same group of objects, obtaining the so-called *multi-source* (or multi-view) data. The analysis of multi-source data presents many challenges. One major challenge is the coexistence of heterogeneous data types in different data sources, such as continuous, binary, and count-valued data. For instance, in genomic studies, data at different molecular levels such as RNA sequencing and DNA methylation data are collected from the same samples. The next-generation RNA sequencing data typically take count values, while DNA methylation data are usually in the form of proportions between 0 and 1. In addition to the diversity of data types, another challenge is the presence of *block-wise missing* data. For the same genomic example mentioned above, not all samples are completely observed in both data sets. Some may only have gene expression measurements, while others may only have methylation profiles. For such a missing structure, it is challenging to impute or integrate different data sources in a principled, unified fashion.

Integrative analysis of multi-source data has drawn more attention to the statistical learning literature lately. Many recent approaches have been developed for integrated analysis of multi-source data Tseng et al. (2015). For example, Shen et al. (2009) introduced an integrative clustering model (iCluster), which incorporates all of the data sources in single clustering analysis. It captures the association and shared clustering between different data sets through a joint latent variable model, but does not consider the unique aspects of each data set. Several recent methods strive to identify not only the shared structure across multiple sources (i.e., *joint*) but the structure that is specific to each source (i.e., *individual*). Lock et al. (2013) developed the Joint and Individual Variation Explained (JIVE) method, which is an extension of the principal component analysis (PCA) to the multi-source data. Supervised integrated factor analysis (SIFA, Li and Jung, 2017) is another method which focuses on the integrative dimension reduction of multi-source data. Several other approaches

that capture joint and individual latent structures have been developed, including extensions of partial least squares Löfstedt and Trygg (2011), canonical correlation analysis Zhou et al. (2016a), non-parametric Bayesian modeling Ray et al. (2014), non-negative factorization Yang and Michailidis (2016), common orthogonal basis extraction Zhou et al. (2016b) and simultaneous component analysis Schouteden et al. (2014). However, these approaches either explicitly assume a Gaussian model or are only appropriate for continuous data.

Batch adjustment techniques Leek et al. (2010); Johnson et al. (2007); Fan et al. (2016) also involve the integration of different sources of data. They adjust raw data across different sample sets by removing batch effects caused by different laboratories or other sources of artificial heterogeneity. However, those approaches handling batch effects are designed for Gaussian data only and can not manage block-wise missing structure.

More efforts are needed for the integrative analysis of data with different types (e.g., count and binary), as heterogeneous data are often encountered due to the disparate nature of multi-source data. The iCluster+ approach Mo et al. (2013), which enhanced iCluster, provides a feasible approach to the clustering of multi-source data with both discrete and continuous values. Very recently, Li and Gaynanova (2017) developed a generalized association study (GAS) framework for the multivariate association analysis of heterogeneous multi-source data. However, none of the existing methods can easily accommodate block-wise missing values.

Block-wise missing structure is ubiquitous in multi-source data sets. Some well known missing value imputation approaches, such as Expectation-Maximum (EM), iterative singular value decomposition (SVD), and matrix completion Mazumder et al. (2010) are effective to impute data that are missing at random in a single data set. However, the assumption of missing at random is not valid for block-wise missing data, and most existing imputation methods are not robust when the missing rate is high Xiang et al. (2014). The standard imputation methods are inappropriate and inefficient for block-wise missing data imputation Yuan et al. (2012). In many applications, a common practice to deal with block-wise missing data is to remove the observations with missing entries. However, such a procedure

may significantly reduce the number of observations and lead to a loss of information. The incomplete multi-task feature learning (iMSF, Yuan et al., 2012) framework conducted a consistent feature selection procedure by avoiding direct block-wise missing imputation. A bi-level learning model Xiang et al. (2014) further extended the iMSF approach to performing covariates-level and source-level analyses at the same time. However, both methods bypass the imputation step when encountering data sets with block-wise missing entries, and thus may have limited generalizability in other contexts. Recently, Cai et al. (2016) developed a structured matrix completion (SMC) method for imputing structured missing data using a Schur completion. SMC can potentially be used for block-wise missing data imputation. However, by design, SMC is only suitable for Gaussian data, and cannot easily handle more than two data sets with heterogeneous data types.

We develop a flexible approach for the dimension reduction of multi-source data that allows different sources to have different data types. By assuming each data source comes from one type of distribution in the exponential family, we simultaneously model joint and individual patterns of the underlying natural parameters across data sources. The proposed method can be applied to block-wise missing data and achieve superior imputation performance. We devise a computationally, efficient algorithm for model fitting. We also introduce an adapted Bayesian information criterion (BIC) to select the underlying ranks of the model (i.e., the ranks of latent joint and individual structures in the model).

Chapter 2

Conditional Quantile Random Forest

2.1 Overview

In this chapter, we propose a tree-based learning approach, conditional quantile random forest (CARF) to do depression screening and intervention. In Section 2.2, we propose the tree-based algorithm of CQRF in detail. In Section 2.3, we introduce how we estimate the heterogeneous quantile coefficient and predict the conditional quantile process of each subject based on CARF. In Section 2.4, We state the uniform consistency property of the estimated quantile coefficient function. The proof of the consistency is included in the supplementary materials.

2.2 Conditional Quantile Trees and Forest

In this section, we describe the proposed statistical framework for the conditional quantile random forest. We first introduce a non-parametric interactive quantile model, which serves as the foundation of the proposed work, followed by computation algorithms and theoretical results of the proposed CQRF.

2.2.1 A Non-parametric Interactive Quantile Model

Let $(y_i; \mathbf{x}_i; \mathbf{z}_i);\ i = 1; \cdots; n$ be a random sample from a target population, where y_i is the continuous response variable of interest, $\mathbf{x}_i$ is the p dimensional vector of splitting variables, and $\mathbf{z}_i$ is q dimensional key predictors. As we introduced in the Section 1, we assume that the conditional quantile function of y_i given $(\mathbf{x}_i; \mathbf{z}_i)$ follows a non-parametric interactive quantile model,

$$Q_{y_i}(\ ; \mathbf{x}_i; \mathbf{z}_i) = \ _0(\ ; \mathbf{x}_i) + \mathbf{z}_i^{>} \ _1(\ ; \mathbf{x}_i); \quad 2\ (0; 1): \tag{2.1}$$

Model (2.1) assumes that the quantile function of y is linear in $\mathbf{z}$, whose coefficients varies across $\mathbf{x}$. We call the Model (2.1) a non-parametric interactive quantile model, since one can view $\mathbf{z}_i^{>} \ _1(\ ; \mathbf{x}_i)$ as the non-parametric interactive quantile effect between $\mathbf{x}_i$ and $\mathbf{z}_i$. It generalizes the traditional additive interactions $\sum_{j=1}^{p}\sum_{k=1}^{q} \ _{j;k} x_{i;j} z_{i;k}$, where $\ _{j;k}$ is pairwise interactive effect between the jth x and kth z. In the motivating example of PTSD prediction among ACS patients, $\mathbf{z}$ is the level of fear during the event of ACS, and we allow the impact of fear varies by individual patients and vary by quantile levels. This non-parametric interactive quantile model can facilitate a wide range of application without restricting to certain parametric forms of $\mathbf{x}$-$\mathbf{z}$ interactions. When $\mathbf{z}_i$ is a binary treatment assignment, coefficient $\ _1(\ ; \mathbf{x}_i) = Q_{y_i}(\ ; \mathbf{z}_i = 1; \mathbf{x}_i) \quad Q_{y_i}(\ ; \mathbf{z}_i = 0; \mathbf{x}_i)$ is the difference of the conditional quantiles between two treatments and it measures the heterogeneous quantile treatment effect. In the studies on gene-environment interactions, one can choose $\mathbf{z}_i$ as an environmental exposure of interest, $\mathbf{x}_i$ is individual genetic profiles. This way, $\ _1(\ ; \mathbf{x}_i)$ captures the gene-environmental interaction effects. When $\mathbf{z}_i$ is an empty set, Model (2.1) is reduced to a traditional tree structure, where all the covariates are used as splitting variables. In the subsequent sections, we propose computation algorithm to construct a CQRF, and how one could use the construct CQRF to estimate the individualized quantile coefficient $\ _1(\ ; \mathbf{x})$, and conduct predictions accordingly.

Without loss of generosity, we abuse model (2.1) as,

$$Q_{y_i}(\ ;\mathbf{x}_i;\mathbf{z}_i) = \mathbf{z}_i^{>}\quad(\ ;\mathbf{x}_i);\quad 2\ (0;1); \tag{2.2}$$

where $\mathbf{z}_i$ contains both 1 and controlling variables in model (2.1).

2.2.2 Construction of CQRF

Throughout the paper, we denote N as a parental node to be split and denote jNj as the sample size of the node N. We denote $\mathsf{S}_N = fsg$ as the collection of all possible binary split rules determined by a single variable in X and a single cut-off value. We also denote N_L^s and N_R^s as the left and right child nodes from a splitting rule s. Consequently, $N_L^s \setminus N_R^s = ;$ and $N_L^s \ [\ N_R^s = N$.

2.2.2.1 Conditional Quantile Based Splitting Criteria

The fundamental component of a tree/forest algorithm is a splitting criterion that recursively partitions the sample space. In the CQRF framework, we would like to select the *best* partition of the parental node N so that result in the most distinctive conditional distribution of y in the two child nodes. We consider two splitting criteria. Splitting Criterion is based on quantile loss functions, which extends from the splitting criterion based on Residual Sum of Squares (RSS) conducted in CART algorithm. Splitting Criterion II is based on a non-parametric test statistics to the ranks of the response variable between two child nodes, which improves the computation efficiency.

Splitting Criterion I: Let $(y_i;\mathbf{x}_i;\mathbf{z}_i)$ be the sample in a parental node N. We first regress y_i against $\mathbf{z}_i$ in N on a sequence of K evenly spaced quantile levels $0 < \ _1 < \quad < \ _K < 1$. That is, we fit a marginal quantile model

$$Q_{y_i}(\ ;\mathbf{z}_i;\mathbf{x}_i\ 2\ N) = \mathbf{z}_i^{>}\ \ _N(\ ;0);\quad 2\ f\ _1;:::;\ _K g; \tag{2.3}$$

where $_N(\ ;0)$ measures the marginal quantile effect of $\mathbf{z}_i$ in the parental node N.

For any candidate split $s \in S_N$, we define a binary indicator $_i(s)$ indicating which child node $\mathbf{x}_i$ belongs to following the split s. We then expand the Model (2.3) to

$$Q_{y_i}(\ ;\mathbf{z}_i;s;\mathbf{x}_i \in N) = f1 \quad _i(s)g\mathbf{z}_i^{>} \ _{N;L}(\ ;s) + \ _i(s)\mathbf{z}_i^{>} \ _{N;R}(\ ;s); \tag{2.4}$$

where $\mathbf{z}_i^{>} \ _{N;L}(\ ;s)$ and $\mathbf{z}_i^{>} \ _{N;R}(\ ;s)$ represent the conditional quantile functions of Y respectively in the left and right child nodes following the partition s. Note that Model (2.3) is nested within the Model (2.4), maximizing the difference between two conditional quantile functions is equivalent to maximizing the reduction of quantile loss between the two models.

We define $L_N(0;\ ;\ ) = \frac{1}{jNj}\sum_i \ _k(y_i \ \ \mathbf{z}_i^{>} \ )I f\mathbf{x}_i \in Ng$ as the quantile loss function of the null model at the quantile level $\ $ in the node N, where function $\ (u) = u(\ \ I fu < 0g)$ is the quantile regression loss function defined in Koenker and Bassett Jr (1978), and $\widehat{L}_N(0;\ )$ is its sample minimum in $\ $. Similarly, we denote $L_N(s;\ ;\ )$ as the quantile loss function of Model (2.4) following the split s, and $\widehat{L}_N(s;\ )$ is the minimized loss function. With these notations, we define a function $\Delta_N(s)$ in the form of

$$\Delta_N(s) = \sum_{k=1}^{K} w(\ _k)\left\{\frac{\widehat{L}_N(0;\ _k) \ \ \widehat{L}_N(s;\ _k)}{\widehat{L}_N(0;\ _k)}\right\} \tag{2.5}$$

to determine the optimal split. Since $\widehat{L}_N(s;\ _k) \ \ \widehat{L}_N(0;\ _k)$ for any s, the term $f\widehat{L}_N(0;\ _k)$ $\widehat{L}_N(s;\ _k)g=\widehat{L}_N(0;\ _k)$ measures the relative improvement in goodness of fit due to the split s at the $_k$th quantile level. Consequently, $\Delta_N(s)$ measures the overall reduction of quantile loss due to the split s, and we select the optimal split by maximizing $\Delta_N(s)$, i.e.

$$s_N^1 = \arg\max_{s2S_N} \Delta_N(s):$$

The weight function $w(\ _k)$ in (2.5) with the constrain $\sum_{k=1}^{K} w(\ _k) = 1$ allows us to assign

difference importance across quantile levels. This could be useful in applications where a certain population stratification is scientifically more important. In all the simulations and real data analysis presented in the paper, we used constant weight $w(\ _k) = 1=K$ for all quantile levels.

Splitting Criterion II: In the splitting criterion I, we need to fit Model (2.4) for every possible split. It can be computationally intensive when the number of candidate splits s is large, even though one can manage with the help of parallel computing. To reduce the computation burden, we propose an alternative splitting criterion that only requires to fit the null Model (2.3) once in the parent node. For each observation in the parental node N, we define a statistics $\widehat{R}_i = \sum_{k=1}^{K} I\{y_i > \mathbf{z}_i^{\top} \widehat{\ }_N(\ _k; 0)\}$. $\widehat{R}_i$ ranges between 1 and K and identifies the ranks of y_i with respect to the estimated conditional quantile process $\mathbf{z}_i^{\top} \widehat{\ }_N(\ _k; 0)$. If $\widehat{R}_i = k$, then y_i stays between the $\ _k$th and $\ _{k+1}$th conditional quantile functions. For each split s, we can construct the contingency table shown in Table 1, where $n_{l;k}(s) = \sum_{\mathbf{x}_i 2 N} I\{R_i = k\}\ _i(s)$ represents the proportion of y_i settling in the kth quantile interval in the left child node (by the split s) and $n_{r;k}(s) = \sum_{\mathbf{x}_i 2 N} I\{R_i = k\}(1 \quad _i(s))$ represents that in the right node. If a split s does not change the conditional distribution of Y, we would expect the distributions of R_i to be homogeneous between the two child nodes. A Chi-squared test statistic, $\ ^2_N(s)$ derived from the contingency table in Table 2.1 helps us to test the homogeneity of the distributions. The larger $\ ^2_N(s)$ is, the more significant the discrepancy of the conditional distributions of y between the left and right nodes is. Hence the optimal splitting rule at this parental node will be chosen as

$$s^2_N = \arg\max_{s2 S_N} \ ^2_N(s): \tag{2.6}$$

Two-step splitting algorithms One could apply the two criteria to go through every possible split to grow a conditional quantile tree. However, such a greedy search often favor a continuous x simply because it has many more possible splits. To avoid such selection bias,

Table 2.1: Contingency table for proposed splitting

	Quantile Levels			
	$(0; \ _1]$	$(\ _1; \ _2]$	$\cdots$	$(\ _K; 1)$
$_i(s) = 1; \mathbf{x}_i \ 2 \ N$	$n_{1l}(s)$	$n_{2l}(s)$	$\ldots$	$n_{(K+1)l}(s)$
$_i(s) = 0; \mathbf{x}_i \ 2 \ N$	$n_{1r}(s)$	$n_{2r}(s)$	$\ldots$	$n_{(K+1)r}(s)$

researchers often consider two-step splitting algorithms (Loh and Vanichsetakul, 1988; Loh and Shih, 1997), where they first choose a splitting variable based on hypothesis testings and then determine the optimal cut-off value for the selected variable in the first step. We followed the similar fashion, and propose a two-step splitting algorithm, where in the first step, we proposed to use integrated rank-score test (Koenker et al. (2010)) to select the *best* splitting variable.

At each parental node N, we first fit the null Model (2.3) over the entire quantile process, and denote $\widehat{\ }_N(\ ; 0)$ as the resulting estimated quantile coefficient function. We then construct an integrated quantile rank test statistics proposed in Koenker et al. (2010) for each of each candidate splitting variable $X_j \ 2 \ \mathbf{X} = fX_j : j = 1; \ldots; pg$. Let $x_{i;j}$ be the jth covariate of the ith subject, and let $\widetilde{x}_{i;j}\widetilde{\mathbf{z}}_i$ be the projection of $x_{i;j}\mathbf{z}_i$ that is orthogonal to the linear space $\mathbf{z}^>(\mathbf{z}\mathbf{z}^>) \ ^1\mathbf{z}$. The test statistics takes the form

$$S_{N;j} = \sum_{i=1}^{n} \left\{ \int_0^1 \widehat{\ }_i(\ )d'(\ ) \right\} \widetilde{x}_{i;j}\widetilde{\mathbf{z}}_i$$

where $\widehat{\ }_i(\ ) = \quad 1fy_i > \mathbf{z}_i^> \widehat{\ }_N(\ ; 0)g$. By design, the test statistics $S_{N;j}$ is approximately zero if and only if the interactive effect between X_j and Z are zero across the entire quantiles. We then use its quadratic form

$$T_{N;j} = S_{n;j}^> Q_n \ ^1 S_{N;j}; \tag{2.7}$$

where $Q_n = \sum_i \widetilde{x}_{i;j}^2 \widetilde{\mathbf{z}}_i\widetilde{\mathbf{z}}_i^>$, to rank the splitting variable. A large value of $T_{N;j}$ indicates stronger interactive quantile effect. The variable with the largest $T_{N;j}$ is chosen as the optimal splitting variable. Once the optimal splitting variable is selected, we use either

splitting criterion I or II to select the optimal cut-off value.

A simulation study: We do a simple simulation study to validate our proposed criteria and compare them with the mean-based splitting criterion. We consider a simple situation that there only two sub-spaces determined by 1 covariates. The splitting variable is X_1, which is generated from a normal distribution with mean 3 and variance 1, and the cut-off value is 3. We investigate two different distributions of Y conditioning on the key predictors Z, which is a binary treatment indicator, taking the value 0 or 1.

Scenario 1 (Location-scale shift): We first consider a scenario that the conditional distributions in the two sub-spaces are location-scale shift. The response variable is generated from the following quantile process,

$$Q_Y(\ ; X; Z) = \quad 0{:}1 + 5Z1fX_1 \quad 3g + (1 + F_e{}^{1}(\ ))Z1fX_1 > 3g + F_e{}^{1}(\ );$$

where e is independently and identical distributed with cumulative density function F_e. The error term e in the first scenario is generated from Cauchy distribution. The treatment effect in the sub-space $1fX_1 \quad 3g$ is 5 across all quantile levels, while the one in the sub-sapce determined by $1fX_1 > 3g$ increases as quantile level increase, which is $1 + F_e{}^{1}(\ )$. The difference of the treatments between the two sub-spaces is $4 + F_e{}^{1}(\ )$, which is an extremely heavy-tail with unbounded variance.

Scenario 2 (Local): The second scenario we consider is a "local" difference that the two conditional distributions in the treatment group are different in a small interval of . The response variable is generated from the following quantile process,

$$Q_Y(\ ; X; Z) = \quad 0{:}1 + 5Z + \quad (\ )Z1fX_1 > 3g + F_e{}^{1}(\ );$$

where

$$(\ ) = \begin{cases} 5 & \text{if} \quad 0.7 \\ 0 & \text{else.} \end{cases} \tag{2.8}$$

The error term is generated from Chi-square distribution with degree freedom equals to 1. The treatment effects in both two sub-spaces are 5 when the quantile level is smaller than 0.7. As to the sub-space determined by $1\{X_1 > 3\}$, the treatment effect becomes to $5 + 5$ when the quantile level is larger than 0.7, which indicates that the difference of the treatment effects between sub-spaces becomes more significant at the higher quantile level. Under this scenario, the treatment effects in the two sub-spaces are the same at the lower quantile levels. Whereas, the treatment effect becomes large at higher quantile levels in one sub-groups but remains the same in another one.

We also introduce 4 noise variable $X_2 \quad X_5$, which are generated from a normal distribution with mean 0 and variance 1 independently. The sample size is 1000 for each simulation study. We apply the two proposed splitting criteria and the criterion based on RSS to select the optimal splits. We repeat the procedure 100 times. The selection performance is evaluated by counting how many times the criterion choose the true splitting variable and the value of the cut-off point as well as the computing time of the selection procedure. The simulation result in Table 2.2 shows that the two proposed splitting criteria choose the true splitting variable with the true cut-off value for both the scenarios in a high frequency, especially under the second scenario. The proposed two-step splitting algorithm also ensure to select the true splitting in a relatively high frequency. However, the selection result of the mean-based criterion is a random guess when the two distributions in the treatment group are a location-scale shift, which indicates that the splitting criterion partitioning the space according to the distinctive means is not able to identify subgroups under such situation. As to the second scenario, though the mean-based criterion provides better performance, it still can not guarantee to choose the true splitting.

Table 2.2: Simulation results for the selection procedure based on the 100 simulation runs. The frequency that the criterion chooses the true splitting variables, the median and median absolute deviation (MAD) of the selected cut-off values and the median and MAD of the computing time are calculated. MAD is in the parenthesis.

		choose X_1	cut-off value	running time
Location-Scale Shift	I	91%	3.13(0.18)	99.25(0.97)
	II	89%	3.03(0.65)	4.00(0.05)
	Two-step	82%	3.13(0.18)	1.45(0.04)
	RSS	16%	0.64(2.74)	4.29(0.07)
Local	I	100%	3.11(0.06)	92.34(0.30)
	II	100%	3.01(0.05)	3.99(0.02)
	Two-step	95%	3.11(0.06)	9.68(0.1)
	RSS	67%	2.77(0.60)	1.46(0.02)

2.2.3 Conditional Quantile Random Forest

Let $(y_i; \mathbf{x}_i; \mathbf{z}_i); i = 1; \ldots; n_t$ be a training data, where n_t n is the number of samples in the training data. We outline below the proposed algorithms to construct a conditional quantile random forest, which is a collection of conditional quantile trees from sub-sampling (Breiman, 2002).

1. **Sub-sampling the training data** We randomly draw a sub-sample from the training data. The number of sub-sampled data is in a predetermined proportion to the total number of the whole data set, like 80%.

2. **Grow a tree** We grow a conditional quantile tree from the sub-sample in Step 1 by recursively partitioning the sample space until each terminal node meets one of the two conditions: (1) The number of observations is smaller than or equal to $q + 1$ and (2) the splitting criterion can no longer be reduced. At each split, we randomly draw a subset of splitting variables from the set of splitting variables. The number of the chosen variables is a predetermined number. We then select the optimal split among them in the following two steps:

   2.1 Select the optimal The optimal splitting variable x^j is chosen by maximizing the test statistics T_{nj} proposed by Koenker et al. (2010).

   2.2 Select an optimal splitting value from the splitting variable Let c be a

splitting value associated with the selected splitting variable x^j , and $s_{j\ ;c} \in S_N$ be the collection of all binary splits rules in the node N determined by j th variable. We then choose the optimal splitting value that either maximizing $\Delta_N(s_{j\ ;c})$ (Splitting criterion I) or maximizing $\hat{\ }^2_N(s_{j\ ;c})$ (Splitting criterion II) regarding to value c.

3. **Assemble a random forest** Repeating this procedure B times independently, we then have a conditional quantile random forest denoted as $T = \{T_b : b = 1, \ldots, B\}$, where T_b stands for a fully grown conditional quantile tree from the b-th sub-sampling.

2.2.4 Variable Importance

By design, the random forest can comfortably accommodate a large number of predictors, and this flexibility leads to enhanced prediction. From the scientific point of view and especially in medical applications, it is also crucial to identify which variables make meaningful contributions for the prediction and rank them for future investigations. In random forest literature, many methods have been developed to determine the variable importance. The conventional approaches include merely counting the number of times a feature appeared in a forest, calculating a predictor's partition power at each split and averaged across all trees, or more computationally intensive permutation approaches that measure the loss of prediction accuracy after permuting a predictor.

The naive approach and permutation-based approaches (Breiman, 2001) can be readily adapted to the conditional quantile random forest. However, the naive approaches tend to underestimate the importance of some variables which strongly associated with the response variables but "masked" by the selected optimal splitting variable. The permutation-based approaches, on the other hand, are computationally undesirable. A more practical way is to take advantage of the integrated rank-score test statistics (2.7), which we evaluated for each candidate splitting variable at each split. As (2.7) measures how well a predictor could partition a node into two distinctive conditional quantile models. Naturally, when we

aggregate the statistics overall splits, the summary statistics measure the overall importance of the predictor. A similar scheme of variable importance is constructed by Loh (2012). By Equation (2.7), we can calculate the rank score statistics $T_{N_{bl};j}$ in the node N_{bl}, where l refers lth split and b refers to tree T_b in the random forest. In this way, the variable importance score for the jth predictor is,

$$\mathsf{T}_{nj} = \frac{1}{B}\sum_{b=1}^{B}\sum_{l=1}^{L_b} T_{N_{bl};j}; \tag{2.9}$$

where L_b is the total number of splits in tree T_b.

2.3 Estimation and Prediction

In this section, we describe how we obtain the estimation of quantile effect and the predicted quantile process of a new-coming observation based on the constructed random forest. We first introduce how we estimate the quantile effect $(\ ;\mathbf{x})$ in Model (2.2) by a weighted quantile loss function. We then elaborate on the procedure to achieve the predicted quantile process if we have a new observation.

2.3.1 Approximation of the Quantile Process $\mathbf{z}^{>}\ (\ ;\mathbf{x})$

We assume that $(\ ;\mathbf{x})$ are smooth functions of on $(0;1)$. We approximate the $(\ ;\mathbf{x})$ by nature linear splices with a sequence of common internal knots $\Pi = f\ _t :\ _t = t=(t_n + 1); t = 1;2;:::;t_n g$, where t_n is a positive integer that t_n " 1 as n ! 1. Let $=$ $(\ _1(\mathbf{x})^{>};\ _2(\mathbf{x})^{>};:::;\ _{t_n}(\mathbf{x})^{>})^{>}$ be the quantile coefficients at quantile levels Π. A nature linear spline function $(\ ;\mathbf{x})$ is defined as a piecewise linear mapping from $[0;1]$ to R^{q+1}. The function is constrained that $^{0}(0;\mathbf{x}) =\ ^{0}(1;\mathbf{x}) = 1$. With enough internal knots, the difference between $(\ ;\mathbf{x})$ and $(\ ;\mathbf{x})$ is negligible de Boor (1978). Equivalently, we approximate the quantile process $\mathbf{z}^{>}\ (\ ;\mathbf{x})$ by its linear spline approximation $\mathbf{z}^{>}\ (\ ;\mathbf{x})$. Such that, to estimate the conditional quantile process, we only need to estimate $(\ ;\mathbf{x})$ at

quantile levels Π.

2.3.2 Estimating the Heterogeneous Quantile Coefficient (; x)

In the proposed model framework, we allow the quantile effect of Z, denoted as (; x), to vary cross the covariate space $\mathbf{X}$ and the quantile level . When Z only contains intercept, (; x) is the conditional quantile function of Y given $\mathbf{x}$. In this section, we outline the estimation procedure to estimate the quantile coefficient function (; x) at a given patient profile $\mathbf{x}_0$ from a constructed CQRF T. Similary as in Breiman (2004) and Lin and Jeon (2006), the estimation procedure is a classic nearest neighbor algorithm. Predictive analyses and other utilities of CQRF depends on the estimation of (; x) will be introduced afterward.

1. Drop the vector of $\mathbf{x}_0$ into each of the tree in the forest T, and denote $N_b(\mathbf{x}_0)$ be the terminal node where $\mathbf{x}_0$ locates in, and $N_b(\mathbf{x}_0)$ be a collection of all samples in $N_b(\mathbf{x}_0)$.

2. We then define a weight function $!_i(\mathbf{x}_0; b)$ for each observation in the training data by,

$$!_i(\mathbf{x}_0; b) = \frac{I(\mathbf{x}_i \ 2 \ N_b(\mathbf{x}_0))}{j \ N_b(\mathbf{x}_0) \ j}:$$

In the weight function, the numerator is a binary indicator which indicates whether the i-th observation $\mathbf{x}_i$ in the training data is located in $N_b(\mathbf{x}_0)$, and the denominator is the cardinality of $N_b(\mathbf{x}_0)$. If $\mathbf{x}_i$ is contained in the set $N_b(\mathbf{x}_0)$, the value of weight is $1=jN_b(\mathbf{x}_0)j$. Otherwise, the weight is 0. The weight $!_i(\mathbf{x}_0; b)$ measures the contribution of $\mathbf{x}_i$ in estimating (; $\mathbf{x}_0$) based on the tree T_b. We then aggregate $!_i(\mathbf{x}_0; b)$ over the entire forest, and define,

$$!_i(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^{B} !_i(\mathbf{x}_0; b) \tag{2.10}$$

as the aggregated contribution of the i-th observation across the random forest T in

estimating the coefficients.

3. Based on the individual aggregated contributions, we obtain a weighted quantile loss function for a give $\in (0;1)$,

$$L_{T;}(\ ;\mathbf{x}_0) = n^{-1}\sum_{i=1}^{n} !_i(\mathbf{x}_0)\quad (y_i \quad \mathbf{z}_i^{>}\ ): \tag{2.11}$$

We write the estimate of (2.11) at a fixed quantile level by

$$\widehat{\ }_{T;}(\mathbf{x}_0) = \arg\min L_{T;}(\ ;\mathbf{x}_0); \tag{2.12}$$

and obtain the estimates on a sequence of quantile levels $\Pi = f\ _t : \ _t = t=(t_n+1); t = 1;2;:::;t_n g$, where t_n is a positive integer that $t_n \to 1$ as $n \to 1$.

4. The estimation of $(\ ;\mathbf{x}_0)$ is then a natural linear spline expanded from $f\widehat{\ }_{T;\ 1};:::\widehat{\ }_{T;\ t_n} g$, i.e.

$$\widehat{\ }_T(\ ;\mathbf{x}_0) = \begin{cases} \widehat{\ }_{T;\ 1}(\mathbf{x}_0) & < \ _1 \\ \widehat{\ }_{T;\ t_n}(\mathbf{x}_0) & > \ _{t_n} \\ \widehat{\ }_{T;\ b\ t_n=(t_n+1)c}(\mathbf{x}_0) & \text{else} \\ +\frac{\widehat{\ }_{T;\ b\ t_n=(t_n+1)c+1}(\mathbf{x}_0) \quad \widehat{\ }_{T;\ b\ t_n=(t_n+1)c}(\mathbf{x}_0)}{1=t_n}(\quad \frac{b\ t_n c}{t_n+1}) & \end{cases} \tag{2.13}$$

Let $\widehat{\ } = (\widehat{\ }_{T;\ 1}(\mathbf{x}_0)^{>}; \widehat{\ }_{T;\ 2}(\mathbf{x}_0);:::; \widehat{\ }_{T;\ t_n}(\mathbf{x}_0))^{>}$ be the estimated quantile coefficients at quantile levels Π. Then the estimator $\widehat{\ }_T(\ ;\mathbf{x}_0) = \ _b(\ )$ is the nature linear spline with $\widehat{\ }$. We show that $\widehat{\ }_T(\ ;\mathbf{x}_0)$ is a consistency estimator of $(\ ;\mathbf{x}_0)$ under certain conditions in Section 2.4.

2.3.3 Predictive Analysis Based on the Conditional Quantile Forest

Like many other approaches, regular random forest only produces a point prediction, which lacks information which measures how far the prediction is away from the true values. A

point prediction sometimes is not acceptable since it can not provide an accurate assessment of the prediction error. By the implementation of our proposed approach, we can develop a conditional prediction interval, which gives a range of values that contains the unknown continuous response with a specified level of confidence, say $100(1 \quad)\%$.

Suppose the conditional quantile random forest T has already built based on the samples $(y_i; \mathbf{x}_i; \mathbf{z}_i); i = 1; 2; : : : ; n$. We now have a new observation $(\mathbf{x} \ ; \mathbf{z} \ )$. By dropping the new case $\mathbf{x}$ down to each tree T_b in the random forest T, we can identify the terminal node associated with $\mathbf{x}$ in the tree T_b and denote it as $N_b(\mathbf{x} \ )$. Followed the same procedure how we estimate the quantile effect $(\ ; \mathbf{x}_0)$ in Section 2.3.2, we can derive the predicted quantile effect $\widetilde{}_{\mathsf{T}}(\ ; \mathbf{x} \ )$ for the new sample $(\mathbf{x} \ ; \mathbf{z} \ )$. Then, the predicted conditional quantile process of y given $(\mathbf{x} \ ; \mathbf{z} \ )$ is constructed by is,

$$\widetilde{Q}_y(\ ; \mathbf{x} \ ; \mathbf{z} \ ) = \mathbf{z}^{>} \widetilde{}_{\mathsf{T}}(\ ; \mathbf{x} \ ); \quad 2 \ (0; 1):$$

The predicted conditional quantile process contains more information than the predicted mean, which is the prediction result by most regular tree-structure model. It is more versatile than point prediction. For a single predicted value, one could either use predicted conditional median $\widetilde{Q}_y(0.5; \mathbf{x} \ ; \mathbf{z} \ )$, or obtain the predicted conditional mean by averaging over the quantile function, i.e. $E_{\mathsf{T}}(Y j \mathbf{x} \ ; \mathbf{z} \ ) = \int_0^1 \mathbf{z}^{>} \widetilde{}_{\mathsf{T}}(u; \mathbf{x} \ ) du$. If a prediction interval is of interest, it could be constructed based on the predicted conditional quantile process,

$$P I \ (\mathbf{x} \ ; \mathbf{z} \ ) = [\mathbf{z}^{>} \widetilde{}_{\mathsf{T}}(\ =2; \mathbf{x} \ ); \ \mathbf{z}^{>} \widetilde{}_{\mathsf{T}}(1 \quad =2; \mathbf{x} \ )]: \tag{2.14}$$

The prediction interval ideally performs well in terms of conditional coverage at or above the nominal level, which is $P f y \ 2 \ P I \ (\mathbf{x} \ ; \mathbf{z} \ ) g \quad 1 \quad$.

The quantile-based prediction intervals do not pre-assume a parametric likelihood. It does not require the scale or the shape of conditional distribution to be the same across predictor values. Hence it facilitates a wide range of applications where parametric likelihoods are insufficient to fit the data. Moreover, since conditional quantile functions fully capture

the entire distributions of the new y, the resulting predictions are more informative, and lead to exciting applications. In what follows, we illustrate one of its potential applications in precision medicine.

2.3.4 Application on Precision Medicine

Suppose we have a data set $(y_i; \mathbf{z}_i; \mathbf{x}_i); i = 1; 2 : : : ; n$ from a clinical study with l different treatments, where $\mathbf{x}_i$ is the feature variables of the i-th patient and y_i is the resulting outcome of interest. We assume that the treatment assignment $\mathbf{z}_i$ is a $\grave{}$ dimensional vector containing 1 and an $\grave{}$ 1 dimensional vector of binary indicators for $\grave{}$ possible treatments. Once the proposed random forest is built upon the study data, by following the procedure in Section 2.3.2, we can obtain the estimation of the quantile treatment effect $(; \mathbf{x}_i) = (_0(; \mathbf{x}_i); _1(; \mathbf{x}_i); : : : ; \grave{} _1(; \mathbf{x}_i))$ for each participant in the study. The $_0(; \mathbf{x}_i)$ is the expected quantile function of the patient's treatment response under the first treatment plan, $_0(; \mathbf{x}_i) + _1(; \mathbf{x}_i)$ is that under the second treatment plan, and likewise, $_0(; \mathbf{x}_i) + \grave{} _1(; \mathbf{x}_i)$ is that under the $\grave{}$-th treatment plan.

When a new patient comes with covariate $\mathbf{x}$, by dropping down to the random forest, we can get the predicted distribution/quantile functions of his/her potential treatment response under each treatment, which are $\widetilde{}_0(; \mathbf{x}); \widetilde{}_0(; \mathbf{x}) + \widetilde{}_1(; \mathbf{x}); : : : ; \widetilde{}_0(; \mathbf{x}) + \widetilde{} \grave{} _1(; \mathbf{x})$. There are many ways to select an optimal treatment for the patient based on conditional quantile functions. Let $(y_1; : : : ; y\grave{})j\mathbf{x}$ be the responses of the patient under the $\grave{}$ treatments. For instance, if there are two treatments in the study, that is $\grave{} = 2$,

1. One can rank the treatments by the stochastic order of their potential responses. Let $(y_1; y_2)j\mathbf{x}$ be the responses of the patient under the $\grave{}$ treatments. We can then select the treatment whose potential responses is superior to the rest treatments in stochastic order. In practice, we can calculate the probability

$$p = P(y_1 \quad y_2):$$

If $p > 0.5$, we say that y_1 is scholastically smaller y_2, we hence recommend treatment 1. Otherwise, we recommend treatment 2.

2. In medical applications, there are often distinct cut-off values for clinical risks. For example, Systolic blood pressure higher than 130 defines hypertension, BMI larger than 25 defines overweight. We calculate $P(y_1 \quad c)$ and $P(y_2 \quad c)$, where c is the cut-off value. We then recommend the treatment with a smaller probability.

3. One can estimate the expected outcome under each treatment, and select the optimal one with the smaller expected value (if the lower value the outcome has, the less severe the participant is). More specifically, $\mathsf{E}(y \mid z_2 = 0) = \int_0^1 \widetilde{}_0(\ ; \mathbf{x}\ )d$ and $\mathsf{E}(y \mid z_2 = 1) = \int_0^1 \widetilde{}_0(\ ; \mathbf{x}\ ) + \widetilde{}_1(\ ; \mathbf{x}\ )d$ are the expected outcomes for the two treatments respectively.

2.4 Theoretical Result

2.4.1 Uniform consistency of $\widehat{}_{\mathsf{T}}(\ ; \mathbf{x}_0)$

In this section, we establish the uniform consistency of $\widehat{}_{\mathsf{T}}(\ ; \mathbf{x}_0)$ obtained from (2.12) and (2.13). We first outline the sufficient conditions to achieve the uniform consistency. Without the loss of generality, we first assume that,

Condition 1. *The covariates* $(\mathbf{x}; \mathbf{z})$ *is uniform on* $[0; 1]^{p+q+1}$.

Condition 2. *Let* $N_b(\mathbf{x}_0)$ *be the leaf space where* $\mathbf{x}_0$ *locates when it is dropped to the tree* b *in the random forest and* $N_b(x_0)$ *be a set containing all samples in the space* $N_b(\mathbf{x}_0)$. *We assume that* $1 = \min_b \mid N_b(\mathbf{x}_0) \mid \ ! \ 0$ *in probability as* $n \ ! \ 1$. *Let* $\mid N_{\mathsf{T}}(\mathbf{x}_0) \mid = \sum_{b=1}^{B} \mid N_b(\mathbf{x}_0) \mid$. *And we have that* $1 = \mid N_{\mathsf{T}}(\mathbf{x}_0) \mid \ ! \ 0$ *in probability as* $n \ ! \ 1$.

Condition 3. *The probability for a speci c covariate* $\mathbf{x}_j$ *to be chosen as a split variable in the random tree is bounded below from* 0.

Condition 4. *For all* $\in (0;1)$*, there exists* $M(\ )$ *so that* $(\ ;\mathbf{x}_0)$ *is Lipschitz continuous with it. And* $\sup_{\in(0;1)} M(\ )$ *is bounded away from* 1 *as* $n \to 1$*. Such that for all* $\mathbf{x}_1;\mathbf{x}_2 \in [0;1]^p$*,*

$$jj \quad (\ ;\mathbf{x}_1) \qquad (\ ;\mathbf{x}_2)\, jj \quad M(\ )\, jj\, \mathbf{x}_1 \quad \mathbf{x}_2\, jj :$$

Condition 5. *The density of* y *given* $(\mathbf{x}_0;\mathbf{z})$ *at th quantile is* $h_{(\mathbf{x}_0;\mathbf{z})}(\ ) = 1=\mathbf{z}^T \ ^{\emptyset}(\ ;\mathbf{x}_0)$*. The true quantile coe cient for a xed* $\mathbf{x}_0$*,* $(\ ;\mathbf{x}_0)$ *are smooth functions on* $\in (0;1)$ *and,*

$$0 < h_{(\mathbf{x}_0;\mathbf{z})}(\ ) < 1 \text{ and } \lim_{\ \to 0} h_{(\mathbf{x}_0;\mathbf{z})}(\ ) = \lim_{\ \to 1} h_{(\mathbf{x}_0;\mathbf{z})}(\ ) = 0;$$

Condition 1 constrains the support of $(\mathbf{x};\mathbf{z})$ to a compact set. Although we assumed continuous support for $(\mathbf{x};\mathbf{z})$, the uniform convergence also held when $(\mathbf{x};\mathbf{z})$ contains discrete variables. Condition 3 makes sure that each covariate is split infinite times before the node stops splitting when the sample size goes to infinity. Condition 5 assumes that the conditional density $f(y|\mathbf{x}_0;\mathbf{z})$ is continuous and uniformly bounded away from 0 and infinity. When quantile level goes to 0 or 1, the conditional density diminishes to 0.

The following Theorem 1 states the uniform consistency property of the estimated coefficient.

Theorem 1. *Under Conditions 1-5 above and Condition 6 in supplementary materials, for* $t_n \to 1$ *and* $t_n = \min_b |N_b(\mathbf{x}_0)| \to 0$*, and a xed* $\mathbf{x}_0$*, the coe cient estimator derived from the conditional random forest* T *is a consistent estimator of the true quantile coe cient* $(\ ;\mathbf{x}_0)$*,*

$$\sup_{\in[1=(t_n+1);t_n=(t_n+1)]} jj\ \widehat{\ }_{\mathsf{T}}(\ ;\mathbf{x}_0) \qquad (\ ;\mathbf{x}_0)\, jj = o_p(1)$$

as $n \to 1$*.*

Chapter 3

Conditional Quantile Random Forest with its Application and Simulation Study Based on Real Data

3.1 Overview

In this chapter, we apply the proposed CQRF to two mental health data sets and conduct simulation based on the real data sets. In Section 3.2, we introduce the REactions Acute Care and Hospitalization (REACH) study. In Section 3.3, we apply the proposed CQRF to REACH data set to do predictive analysis. In Section 3.4, the result of variable importance derived by CQRF demonstrates some inspiring conclusions. In Section 3.5, we conduct a simulation study based on REACH data set to validate the consistent property of the estimated quantile coefficient function. In Section 3.6, we introduce the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. In Section 3.7, we apply the proposed CQRF to STAR*D data set and illustrate how the proposed approach facilitates the personalized treatment recommendation. In Section 3.8, we conduct a simulation study based on STAR*D data set to evaluate the performance of the individualized treatment acquired

by the proposed CQRF.

3.2 REactions to Acute Care and Hospitalization (REACH) Study

REactions to Acute Care and Hospitalization (REACH) study is an ongoing prospective observational cohort study of the patients admitted Columbia-New York Presbyterian Hospital with symptoms of a suspected acute coronary syndrome(ACS). All the patients enrolled in REACH study were followed up at a month after their Emergency Room (ER) discharges. Rich and comprehensive electronic medical records (EMR) of patients were also collected during their ER stays. The EMR includes the demographic information (e.g., age, gender, ethnicity, race, education, and religion), detailed symptoms (e.g., whether a patient have chest pain), medical history and co-morbidity conditions (e.g., whether the patient has cardiovascular disease history), procedures performed, lab results, the experiences in ER (e.g. waiting time), and mental health assessments (e.g., patients' threat perceptions, PHQ-9 depress scores, cumulative traumatic life events) and other related information.

For many, ACS is a frightening experience associated with significant pain and the fear of dying (Whitehead et al., 2005; von Känel et al., 2011). In such patients, ACS can induce long-lasting symptoms of Post-Traumatic Stress Disorder (PTSD) (Abbas et al., 2009), a psychiatric disorder that results in substantial psychological distress and impaired functioning. Besides, ACS-induced PTSD has been shown to increase the risk of adverse prognosis. Given the adverse psychological impact of ACS-induced PTSD and its potential to increase cardiovascular risk, there is an urgent need to develop prediction and intervention tools to prevent the onset of PTSD in ACS survivors.

All the patients enrolled in REACH study were followed up at a month after their ER discharges. At that follow up, physicians evaluate patients' PTSD symptoms using PCL-5, which is a self-report measure. The resulting PTSD score ranges between 17 and 89. The higher the score, the severer the case. A score of 32 indicates potential PTSD. One-month

PTSD score is the outcome of interest. The goal is to predict the one-month PTSD score based on a patient's EMR. Figure 1.1 plots the history of PTSD in the REACH cohort.

According to the psychiatric theory, PTSD is triggered by the fear of death and re-occurrence of a traumatic event. Following this theory, we selected four variables that relate to the level of fear during the ER stay, and use them as $\mathbf{z}$ in the main model. They are the acute stress disorder symptoms (ASDS) score, perception score, PHQ depression score, and baseline PTSD score. The rest EMR variables are treated as splitting variable $\mathbf{x}$ in the proposed approach. This way, the quantile coefficient $(\ ;\mathbf{x})$ in the interactive quantile model (2.1) is the effect of baseline fear on the development of one-month PTSD given a patient's profile $\mathbf{x}$. Model (2.1) essentially assumes the conditional quantile function of one-month PTSD depends on the level of fear at baseline $\mathbf{z}$, and how well a patient cope with fear $(\ ;\mathbf{x})$.

There is 1002 patients information variable from REACH study. We do some pre-processing to this information that we delete subjects with missing response variable, and we exclude variables with more than 40% missing observations. Finally, we have a data set with 764 subjects, 1 response variable, which is PTSD score 1-month after ER discharge, 4 fear variables as controlling variables in the interactive model, and 90 splitting variables containing demographic information and other related measurements.

3.3 Predictive Analysis

We apply the proposed conditional quantile random forest algorithms to data set from REACH study. 500 trees are contained in the forest. In each parent node, we need to regress the response variable against the controlling variables at a sequence of quantile levels. We set the number of quantile levels to be $K = 10$. For each split, a different random subset of the variables is chosen to be the potential splitting variables. Such that, the correlation between trees in the random forest reduces, while the variance does not increase too much. In the implementation of REACH data set, we randomly choose 1=3 of

the variables in each split.

In Section 2.3.3, by implementing the predictive analysis of the proposed approach, we can get the predicted conditional quantile/distribution functions for new coming subjects. The predictive analysis conducted on every subject in the data set is achieved by 5-fold cross-validation. The data set is divided into five-folds, with equal sample sizes. One fold is left out as testing data iteratively, and the rest folds are used to construct the proposed random forest. Once the random forest is built, the left-out fold of data is dropped down to the forest, and the predicted conditional quantile/distribution functions are derived by following the procedure in Section 2.3.3. In order to evaluate the prediction accuracy, we establish two assessment measurements. One is the length of the prediction interval, which is $\mathbf{z}\ \widetilde{}_{\mathsf{T}}(1\quad =2;\mathbf{x}\ )\quad \mathbf{z}\ \widetilde{}_{\mathsf{T}}(\ =2;\mathbf{x}\ )$ for the prediction interval $PI\ (\mathbf{x}\ ;\mathbf{z}\ )$ in Equation (2.14). The other one is the coverage rate $Pfy\ 2\ PI\ (\mathbf{x}\ ;\mathbf{z}\ )g$. The sample coverage rate is calculated by

$$\widehat{P}fy\ 2\ PI\ (\mathbf{x};\mathbf{z})g = \frac{1}{n}\sum_{i=1}^{n} 1fy_i\ 2\ PI\ (\mathbf{x}_i;\mathbf{z}_i)g:$$

Besides the proposed conditional quantile random forest, we also consider several other approaches. They are,

Marginal Quantile Random Forest (MQRF): Instead of treating fear variables as the controlling variables in the interactive model (2.1), one can also set the controlling variables to be null and consider the fear variables as potential splitting variables. Therefore, in Model (2.2), $\mathbf{z}$ only contains a 1 and the new splitting variables are $\widetilde{\mathbf{x}} = (\mathbf{x};\mathbf{z})$.

Linear Regression with LASSO (LR-LASSO): We regress one-month PTSD against both fear variables and variables containing other information with a least absolute shrinkage and selection operator (LASSO) penalty. 5-fold cross-validations are used to select the penalty term.

Quantile Regression Forest (QRF): We build the classical mean based random forest using both $\mathbf{x}$ and $\mathbf{z}$ as splitting variables. Same as in the quantile forest (Meinshausen, 2006), the prediction interval is based on a weighted marginal quantile estimation.

The resulting coverage rate and the average length of the prediction intervals are presented in Table 3.1. Other than evaluating the prediction performance of the whole data set, we also conduct a subgroup evaluation, where we only consider those high-risk samples. The clinically meaningful cut-off point for PTSD evaluated by PCL-5 is 32. Accordingly, high-risk samples are those participants who have score larger than 32 ($n = 130$). The nominal level for the prediction interval is 95%. Different splitting criterion introduced in Section 2.2.2.1 are conducted to both CQRF and MQRF. Table 3.1 shows that the coverage of the prediction intervals from the linear regression with LASSO penalty fail to reach nominal level of 95% among the entire sample as well as among the high-risk subgroup (coverage = 89.1% among the whole sample, and 68.5% among the high-risk group). The quantile regression forest improved overall coverage and accuracy (coverage = 91.5% and average length = 23.6 among entire data) but did not capture a sufficient number of high-risk patients, either (coverage = 58.8% among the high-risk group).

All the prediction intervals from the quantile forests have reached the 95% nominal coverage level among the data (ranging from 95.4% to 97.1%). The coverage rate among the high-risk groups is slightly lower than the nominal level (85.3%-93.1%), but much higher the ones from the mean-based approaches. We also observed the reduction of uncertainty by using the conditional quantile model. For example, the average length of the prediction intervals of the marginal quantile random forest with splitting rule II is 31.9 among the entire sample, and it was reduced to 27.8 in its conditional counterpart. Finally, the differences between the splitting rules I and II are minimal.

Nonetheless, the average lengths of the prediction intervals are not promising. The proposed approach is versatile because we do not have a strong pre-assumption of the conditional response distribution. But it comes with a cost. Referring to Section 2.3.2

and Section 2.3.3, we restrict the estimation and prediction of the conditional response distribution to a small sample space around splitting variables for each subject. In practice, the data we work with to do estimation or prediction is relatively small, which leads to highly variable estimators and predictors. Although we observed considerable improvement in the prediction of the proposed quantile random forest in terms of the coverage rate, especially among the high-risk population, further improvement is needed in order to provide more informative prediction intervals.

Table 3.1: Coverage and average lengths of the cross-validated prediction intervals using different approaches, among the entire sample and among the high-risk group with one-month PTSD score 32.

	Entire sample ($n = 746$)		High risk (PTSD 32 $n = 130$)	
	Average length	Coverage	Average length	Coverage
CQRF splitting I	28.01	95.81%	40.10	86.92%
CQRF splitting II	27.81	95.39%	39.73	85.38%
MQRF splitting I	31.40	95.95%	42.20	90.48%
MQRF splitting II	31.88	97.11%	44.00	93.08%
LR-LASSO	25.29	92.33%	30.37	71.77%
RF	24.86	93.86%	45.28	70.99%
QRF	30.64	96.37%	41.78	82.44%

3.4 Variable Importance

Regarding the REACH study, electronic medical record information for the participants is rich currently. It might be too costly to collect such rich information. We need to reduce the number of variables in order to make the study less costly but remain the same efficiency. Variable importance is a useful by-product of the random forest, which helps us to single out those variables that influence the value of response the most. It makes it possible that only a part of the information is required to be collected without loss of efficiency. It also gives suggestions to clinicians that if a new coming patient with abnormal values of certain variables, he/she should be paid attention to.

Based on the proposed framework, we can derive two kinds of variable importance.

If the controlling variables are considered in the interactive model (2.1), the importance scores prioritize which variables differentiate the effect of controlling variables on the response variable the most.

If no controlling variable is considered in the interactive mode, the importance scores prioritize which variables differentiate the heterogeneous distributions of the response variable the most.

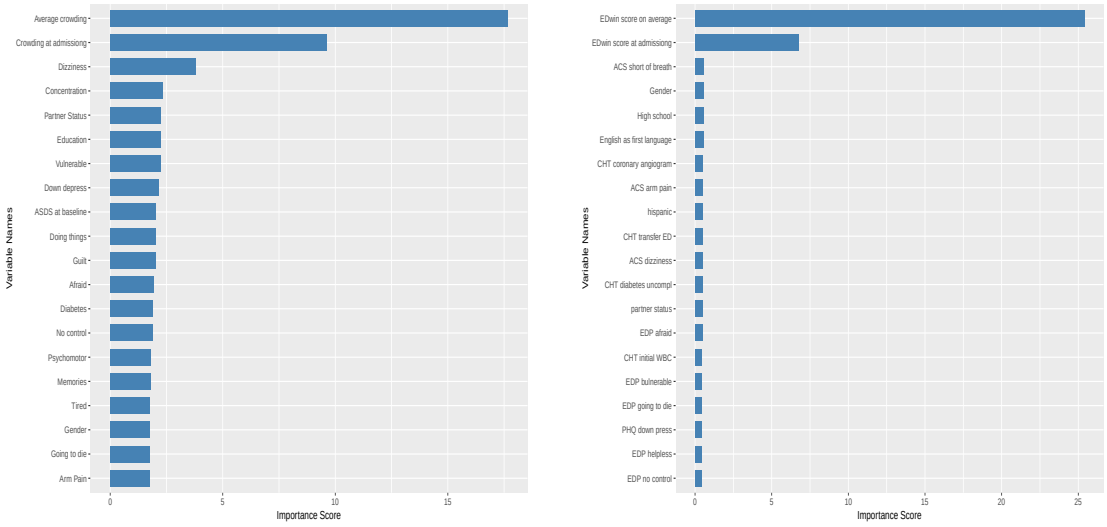
Figure 3.1 shows the variable importance scores for the top 20 variables derived based on the proposed framework. Figure 3.1a is the variable importance scores based on CQRF and Figure 3.1b is the ones based on MQRF. It is easy to observe that two variables are standing out from the other variables. They are measurement about how crowded the emergency room was during the patients' stay and at admission time. There are much literature mentions that emergency room crowding may spur PTSD symptoms (Edmondson et al., 2012, 2013). However, these two variables have not been chosen by other approaches, either by linear regression with LASSO penalty or quantile random forest. Important scores obtained by stratified analysis to REACH data set using regular random forest shows that if we only focus on high-risk samples, those who have PTSD scores larger than 32, measurements of crowding are singled out. But they remain undetectable among samples with lower PTSD scores.

Figure 3.2 illustrates the heterogeneous effect of one of the key predictors we include in the interactive model: PHQ score on the PTSD score at different quantile levels of the PTSD score and crowding measurements. The effect varies in the different quantile levels in both the response variable space and the splitting variable space. The effect of PHQ score on the PTSD score reaches the most significant in the population with a large value of both PTSD score and crowding. Methods like linear regression with LASSO penalty or random forest depending on mean-based importance measurement evaluate how well a variable split the space with distinctive means only in the covariate space. Thus they are not able to pinpoint those variables which lead to a more complex heterogeneity. Nonetheless,

Figure 3.1: Variable Importance

(a) Variable importance of splitting variables $\mathbf{x}$ based on the conditional quantile random forest.

(b) Variable importance of $\mathbf{x}, \mathbf{z}$ based on the marginal quantile random forest.



our proposed variable importance measures how well a variable partition the sample space in terms of the complete conditional distribution of the response variable within each node, considering the heterogeneity in both outcome and covariates space at the same time. The proposed scheme to use rank-score test statistics as importance score is able to identify variables, which contribute to a more complex heterogeneity.

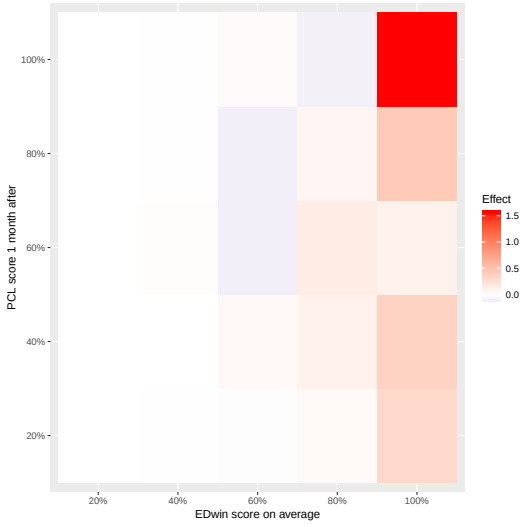
3.5 Simulation based on REACH data set

3.5.1 Simulation Settings

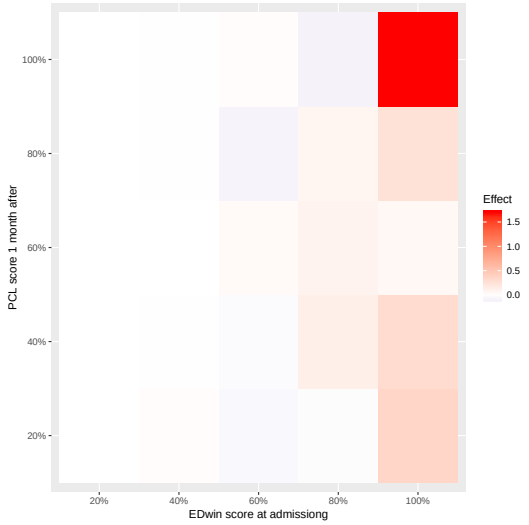
We conduct a simulation study where the data set is a resemble of the real REACH data set to examine the performance of asymptotic property of $\widehat{Q}(\ ; \mathbf{x})$ we state in Section 2.4.1. Following the procedure in Section 2.2.3, we build up one tree by implementing the proposed approach to the REACH data set. Here, in each node, all $\mathbf{x}$ serve as potential splitting variables. Instead of building a fully grown tree, a stopping rule is employed to terminate further splitting resulting in a relatively smaller tree structure. When the number

Figure 3.2: The effect of PHQ score on PCL-C score one month after discharge at different quantile levels of crowding measurements and different quantile levels of PCL-C score.

(a) Crowding on average



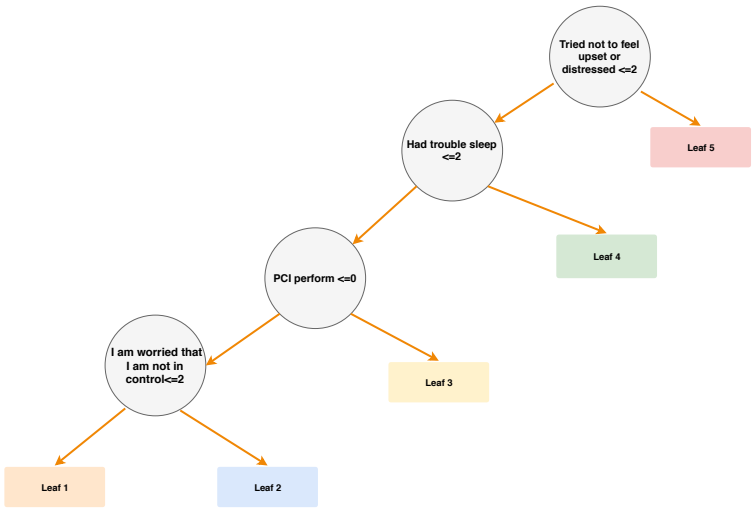
(b) Crowding at admission



of observations in either of two child nodes is smaller than 50, the node is marked as a terminal node. The tree stops growing when all the nodes are marked as a terminal node. Figure 3.3 presents the resulting tree structure, which partition the sample space into five sub-spaces. In each of the terminal nodes, we regress the one-month PTSD against the 4 baseline fear variables at 100 equally spaced quantile levels and we denote the resulting quantile coefficient as $_{l} = (\ _{l\,1};\ _{l\,2};\cdots;\ _{l\,K})$, where $_{k} = k/100; k = 1; 2; \cdots; 100$ and $l = 1; 2; \cdots; 5$. The true quantile coefficients $_{l}(\ ; \mathbf{x}); k = 1; 2; \cdots; 5$ are the nature linear splines with internal knots $_{l}$. Figure 3.4 illustrates the underlying true quantile coefficient functions associated with each controlling variable within the 5 terminal nodes. We notice that those quantile coefficient functions have distinctive patterns across the terminal nodes, which indicates heterogeneous quantile effects among the ACS populations.

We randomly sample simulated covariates and controlling variables $(\mathbf{x}\ ; \mathbf{z}\ )$ from the original REACH data set with replacement. The simulated covariates $\mathbf{x}$ are dropped into the true partition structure in Figure 3.3 and we denote the true conditional quantile process $Q_{y}\ (\ ; \mathbf{x}\ \ 2\ N_{l}; \mathbf{z}\ ) = f\mathbf{z}^{>}\ (\ ; N_{l}) : \mathbf{x}\ \ 2\ N_{l}g$, where $l = 1; 2; \cdots; 5$. For each

Figure 3.3: True tree structure underlying the simulated data set is constructed by the proposed procedure with one single tree.



subject, the simulated outcome is randomly sampled from the underlying true conditional quantile process. In particular, p is a uniformly random number $Unif(0;1)$. Then, the simulated value of response variable is $y = \inf\{y : y > Q_y(p; \mathbf{x}; \mathbf{z})\}$.

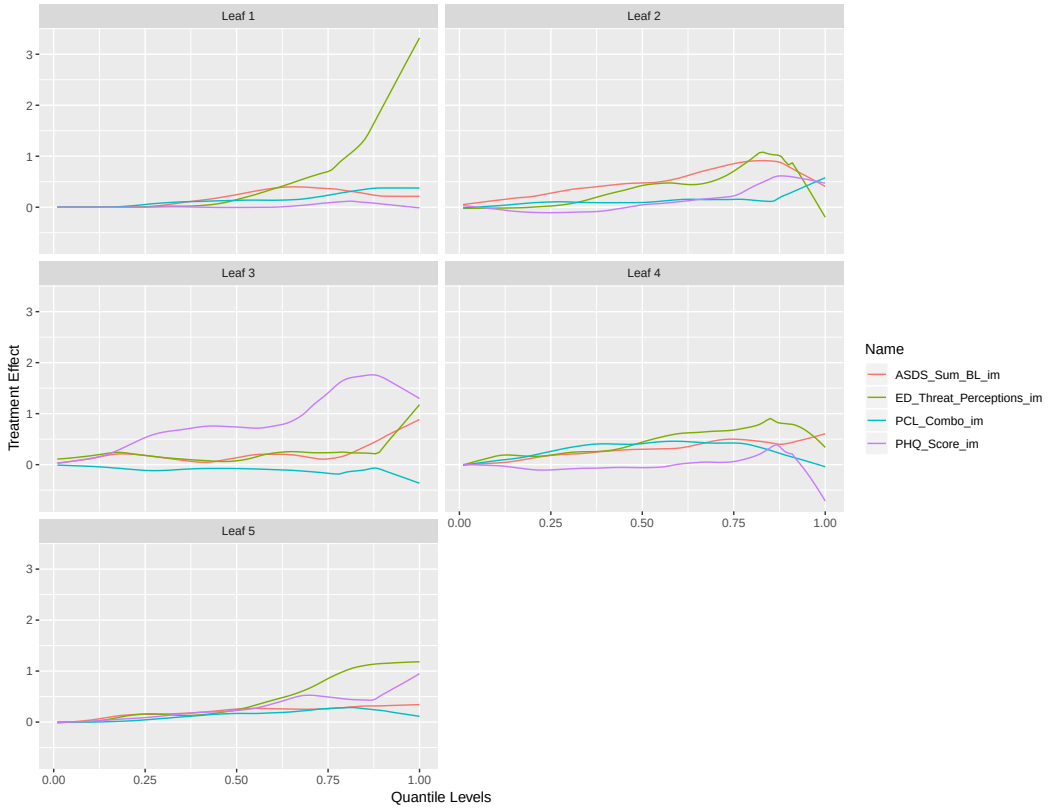
3.5.2 Simulation Result

The proposed conditional quantile random forest is implemented to the simulated data set to obtain the estimated coefficients $\hat{(}; \mathbf{x}_i \mathbf{z}_i)$. The bias for each proposed random forest is calculated by

$$\bar{B}_{T_f} = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 f\hat{}_{T_f}(; \mathbf{x}_i; \mathbf{z}_i) \quad (; \mathbf{x}_i; \mathbf{z}_i)gd; \; f = 1;:::;F; \tag{3.1}$$

where T_f stands for one random forest, F is the total number of iterations we repeat, N is the number of samples in the simulated data set and $\hat{}_{T_f}(; \mathbf{x}_i)$ is the estimated conditional quantile function derivein Section 2.3.2 based on the random forest T_f. We try with $N = 1000$ samples in the simulated data set and $T = 500$ trees in the proposed random forest in each iteration. Table 3.2 shows the bias between the estimated coefficients derived

Figure 3.4: True coefficients at 100 equally spaced quantile levels between 0 and 1 in the 5 leaves based on the true tree structure.



from the propose random forest and the true coefficients. The results include the median and median absolute deviation (MAD) of estimation bias of each controlling variable at different quantile levels $_k = (0{:}1; 0{:}2; \cdots; 0{:}9)$. The biases of the estimated coefficients for the four controlling variables are at 10 ^{2} level. Table 3.2 shows the resulting median of biases and its median absolute deviation. The small bias result implies that the estimated quantile function in Section 2.3.2 are consistent to the true conditional quantile function $(\ ; \mathbf{x}; \mathbf{z})$.

We also perform 100 simulations with $N = 1000$ for the proposed CQRF and the naive method, where we regress the response variable y with the controlling variables $\mathbf{z}$ at 100 equally spaced quantile levels between 0 and 1 in the whole data set. The four panels illustrate the sample distributions of bias (Equation (3.1)) from the proposed approach and

Table 3.2: Simulation result for simulated data based on random forest with 500 trees when the true quantile coefficients function are fixed. The covariates and controlling variables in the simulated data are randomly sampled from REACH data and response variable is sampled from the true quantile processes. Biases are calculated for each random forest, each controlling variable and 9 equally spaced quantile levels between 0 and 1. The median and the median absolute deviation (MAD) for estimation bias of each controlling variable at different quantile levels are calculated. MAD is in parenthesis.

k	B_b			
0.1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.2	-0.08 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.3	-0.03 (0.00)	-0.02 (0.00)	-0.01 (0.00)	-0.09 (0.01)
0.4	-0.10 (0.05)	0.00 (0.01)	0.00 (0.00)	-0.08 (0.04)
0.5	-0.08 (0.07)	0.01 (0.13)	0.01 (0.06)	-0.03 (0.05)
0.6	-0.09 (0.06)	-0.04 (0.14)	0.04 (0.09)	0.01 (0.05)
0.7	-0.09 (0.07)	-0.03 (0.16)	0.00 (0.11)	0.02 (0.06)
0.8	-0.12 (0.06)	0.01 (0.13)	0.05 (0.10)	-0.05 (0.07)
0.9	-0.06 (0.08)	-0.17 (0.17)	-0.03 (0.14)	-0.06 (0.06)

the naive approach at select quantile levels 0.1, 0.5, and 0.9 for the four controlling variables. Figure 3.5 illustrates that the naive method performs poorly, even at median levels. The proposed approach has a smaller biases with smaller variability.

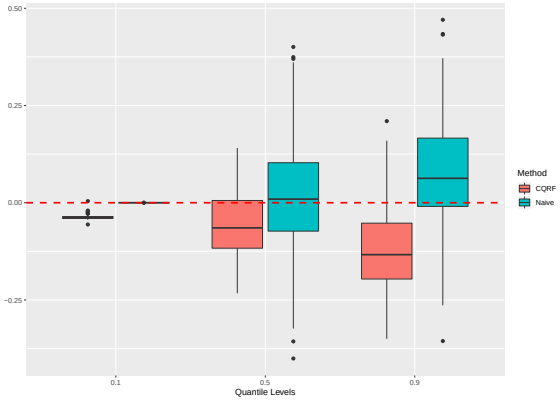
3.6 Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (Rush et al., 2004) was a multi-site, multi-level, randomized clinical trial that simultaneously assesses the efficacy of multiple depression treatments for patients with major depression disorder (MDD) (Chakraborty and Moodie, 2013). The severity of depression is assessed by the *Quick Inventory of Depressive Symptomatology* (QIDS), which is the primary outcome of the study. The range of QIDS score is from 0 to 27. A lower score indicates less severe symptoms. In general, a healthy person should have a QIDS 5. A QIDS score between 6 to 10 indicates a mild depression; QIDS between 11 and 15 is moderate depression; any value higher than 16 is severe depression.

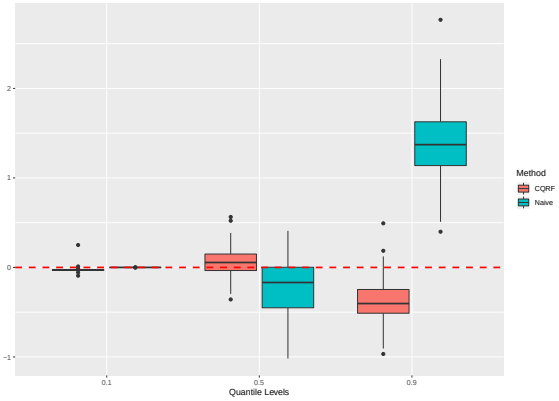
There are four levels of treatments in the study. We only focus on the second level.

Figure 3.5: Box plot of the estimation biases derived by CQRF and naive method. One is the proposed CQRF with 500 trees.

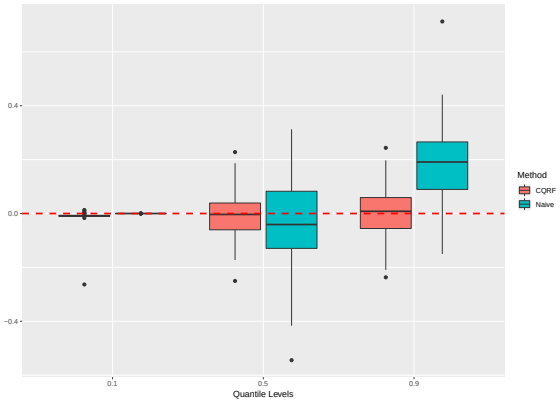
(a) Box plot of the estimation bias for ASDS score at baseline



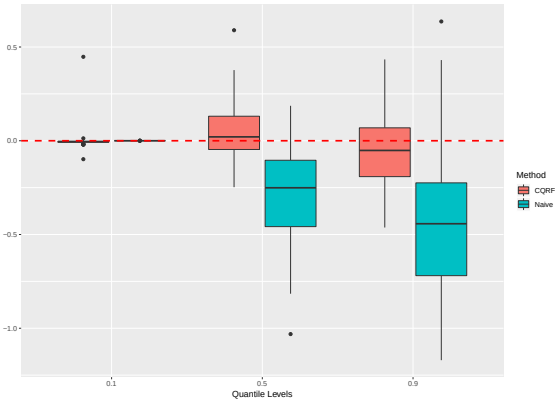
(b) Box plot of the estimation bias for threat perception score



(c) Box plot of the estimation bias for PCL score



(d) Box plot of the estimation bias for PHQ score



At level 1, participants were treated with Citalopram (CIT) for a minimum of 8 weeks. If the participant had a total QIDS score under 5, which was considered to have a clinically meaningful response, he/she would be excluded from the study. Those who were not remitted from the future treatments were randomized to the level 2 of treatment based on their preference or augmenting their QIDS score in the first level.

There are 7 treatments in the study, and they are bupropion (BUP), cognitive psychotherapy (CT), sertraline (SER), venlafaxine (VEN), citalopram (CIT)+BUP, CIT+ buspirone(BUS), and CIT+CT. Participants who preferred to switch treatment were randomly assigned to BUP, CT, SER, or VEN. Participants who preferred augmentation of their treatment were randomly assigned to CIT+BUP, CIT+BUS, or CIT+CT. These treatments fall into two categories. Treatments SER, CIT+BUP, CIT+BUS, and CIT+CT are the treatments with selective serotonin reuptake inhibitors (SSRIs), while the other three treatments (BUP, CT, and VEN) are the ones with non-SSRIs. The study also collected rich patient-level information including gender, race, age, marital status, employ status, family history of depression and mental health status at baseline, such as clinical-rated and self-rated QIDS score, whether patient presented a suicide risk, and how often the patient missed medicine. The response variable is the QIDS score at the end of the level-2 treatment. We also include 36 variables related to patient-level information and 1 dichotomized treatment indicator variable in the data set. There are 557 patients with complete response and treatment variable information at level-2 treatment included in our analysis.

3.7 Personalized Medicine

We apply the proposed conditional quantile regression random forest to estimate the optimal treatment of participants who were enrolled in the level-2 treatment. We first construct a conditional quantile random forest with 500 trees following the algorithms in Section 2.2.3. Suppose $\mathbf{x}$ is the covariate profile of a new patient. Based on the constructed algorithms, we then estimate quantile coefficient $\widehat{(\ ;\mathbf{x}\ )}$. Consequently, if the new patient were assigned

to a non-SSRI treatment, the quantile function of his/her QIDS score at the end level-2 treatment is $\widehat{\ }_0(\ ;\mathbf{x}\ )$, if he/she was assigned to a SSRI treatment, the quantile function of his/her QIDS score at the end level-2 treatment is $\widehat{\ }_0(\ ;\mathbf{x}\ )+\widehat{\ }_1(\ ;\mathbf{x}\ )$. Let $(y_s;y_n)|\mathbf{x}$ be the response of the new patient under the two treatments when the predicted quantile functions are the true quantile functions. Following the three ways to select optimal treatment in Section 2.3.4, we can determine the optimal treatment $\hat{z}$ for the new coming patients with $\mathbf{x}$.

In order to evaluate whether the proposed treatment recommendation produces sensible results, we establish an assessment measurement. The measurement calculates the probability of observing a score smaller or equal to one's true QIDS score based on the predicted conditional quantile function with the optimal treatment chosen by a particular principle. The data set is divided into ten folds. One fold of the data is left out iteratively, while the proposed CQRF is implemented for the remaining nine folds of data. Following the procedure in Section 2.3.3, we can obtain the predicted conditional quantile process for both treatments: $\widetilde{Q}_y\ (\ ;\mathbf{x}\ ;z=1)=\widetilde{\ }_0(\ ;\mathbf{x}\ )+\widetilde{\ }_1(\ ;\mathbf{x}\ )$ and $\widetilde{Q}_y\ (\ ;\mathbf{x}\ ;z=0)=\widetilde{\ }_0(\ ;\mathbf{x}\ )$, where $\mathbf{x}$ is contained in the left-out fold of data. We carry out the three strategies introduced in Section 2.3.4 to select the optimal treatment. The corresponding optimal treatment chosen by these three different ways are denoted as $\hat{z}_o;o=1;2;3$. Then we calculate the measurement by,

$$\hat{p}_{io}=\hat{p}(\mathbf{x}_i;z=\hat{z}_o)=\inf\{\ ;\widetilde{Q}_y\ (\ ;\mathbf{x}_i;z=\hat{z}_{io})>y_i\} \tag{3.2}$$

where y_i is the observed outcome of the ith patient, and $\hat{z}_{io}$ is the selected optimal treatment for ith patient, oth principle. A large probability calculated by (3.2) indicates that the selected treatment will lead to a better outcome (a smaller QIDS score) in high probability.

There are 274 participants suggested changing to another treatment by the first principle. The median of $\hat{p}_{i1}$, if $z_i \neq \hat{z}_{i1}$ is 0:62 and the median absolute deviation (MAD) is 0:43. There are 250 participants suggested changing to another treatment by the second principle.

The median of $\widehat{p}_{i2}$, if $z_i = \widehat{z}_{i2}$ is 0.56 and MAD is 0.48. There are 282 participants suggested to change to another treatment by the third principle. The median of $\widehat{p}_{i3}$, if $z_i = \widehat{z}_{i3}$ is 0.70 and MAD is 0.33. The third principle outperforms the other two for the STAR*D data. It is also the most straightforward principle among the three. Thus, in the following analysis, we decide to use the third principle to choose the optimal treatment.

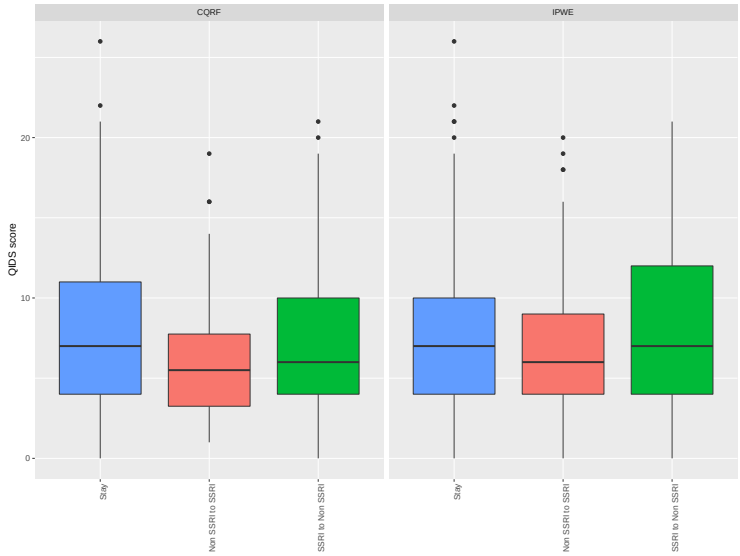
By comparing the treatment assignment conducted by the proposed approach with the true treatment the participants receive, we could classify the samples in the data set into 3 categories: stay in the same assignment indeed received; change from SSRI to non-SSRI treatment; change from non-SSRI to SSRI treatment. We also consider the individualized treatment rules (ITR) random forest by inverse probability weighted estimator (IPWE) (Doubleday et al., 2018) as alternative approaches. Comparing the expected outcomes between treatments, the third way to determine the optimal treatment in Section 2.3.4, is implemented as an assignment rule for the proposed random forest approach. Figure 3.6 illustrates the box plot of true QIDS scores for the two approaches stratified by different categories. The median score of those who are recommended to stay in the same treatment group is 7 for both approaches (median absolute deviation (mad) = 4.4478). The median score of those who are recommended to change to SSRI treatment is 6 (mad = 4.4478) for the proposed CQRF, while the one for random forest by IPWE is 7 (mad = 4.4478). The median score of those who are recommended to change to non-SSRI treatment is 5.5 (mad = 3.7065) for the proposed CQRF, while the one for random forest by IPWE is 6 (mad = 4.4478).

3.8 Simulation based on STAR*D data set

3.8.1 Simulation Settings

In this section, we conduct a simulation study where the simulated data resemble the STAR*D data with clinical-rated QIDS collected at the end of level-2 treatment as our

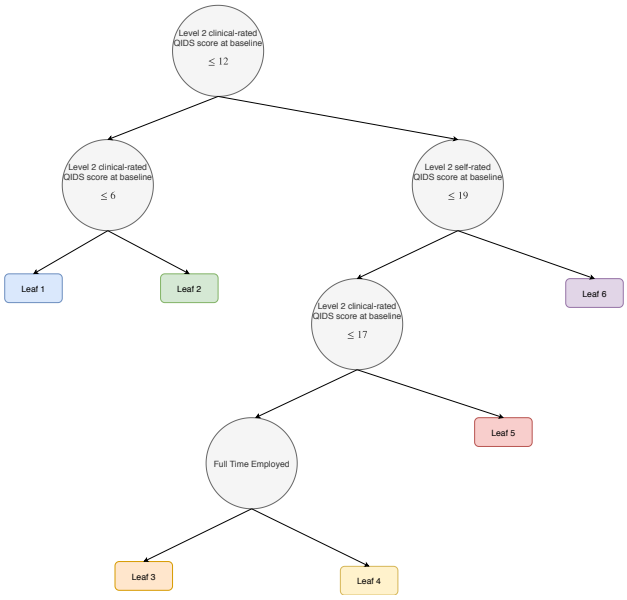
Figure 3.6: The box-plots of the response variable QIDS score based on whether the new assignment makes the patient stay in the same treatment group or change to another group. The left panel is the proposed approach: conditional quantile random forest and the right panel is the random forest with IPWE (Doubleday et al., 2018)



response variable. Similarly, the true partition structure is derived by following the procedure in Section 2.2.3. It is required that each terminal leaf contains at least 30. Figure 3.7 demonstrates the underlying true tree structure of the simulated data. Likewise, within each leaf, we regress clinical-rated QIDS score at the end of level 2 against treatment indicator at 100 equally spaced quantile levels and we denote the resulting quantile coefficient as $_l = (\ _{l;\,1};:::;\ _{l\ K})$ where $_{l\ k} = (\ _0(k{=}100;N_l);\ _1(k{=}100;N_l))$. The true quantile coefficients $_l(\ ;\mathbf{x})$ are the nature linear splines with internal knots $_l$. In Figure 3.8, the true underlying quantile specific treatment effects in 6 different leaves implies a different response to the two treatments by different patients.

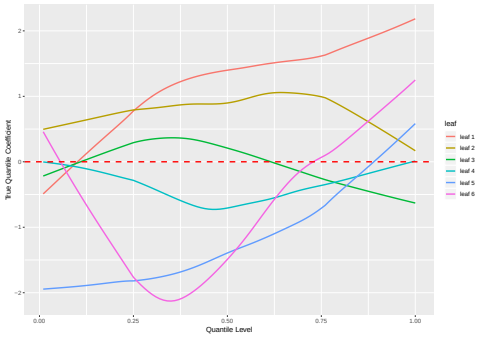
We randomly draw covariates $\mathbf{x}$ from STAR*D data set with replacement. The simulated $\mathbf{x}$ is dropped down to the true tree structure. The patient is randomly assigned to either of the two treatment groups with equal probability, which is denoted as z . The true conditional quantile process is $Q_Y(\ ;\mathbf{x}\ \ 2\ N_l;\mathbf{z}\ ) = f\mathbf{z}^{>}\ (\ ;N_l) : \mathbf{x}\ \ 2\ N_l g$, where $\mathbf{z}\ = (1;z\ )$, $(\ ;N_l) = (\ _0(\ ;N_l);\ _1(\ ;N_l))$ and $l = 1;2;:::;5$. QIDS score is randomly

Figure 3.7: The true underlying tree structure of each simulated data is constructed by the proposed procedure in Section 2.2.3 with one single tree where the size of the terminal nodes are at least 30.



sampled from the corresponding true conditional quantile process. We randomly sample a p form the uniform distribution $Unif(0;1)$. Then the simulated value of the response variable is $y = \inf\{y : y > {}_0(\ ; N_l) + {}_1(\ ; N_l)\}$ if the patient is assigned to SSRI treatment or $y = \inf\{y : y > {}_0(\ ; N_l)\}$ otherwise.

Figure 3.8: True coefficients at 100 equally spaced quantile levels between 0 and 1 in the 6 leaves based on the true tree structure.



3.8.2 Simulation Result

We try the simulated data set with $N = 1000$ sample size with $T = 500$ trees in each proposed random forest and repeat the simulated process 100 times. Similarly, we calculate the bias between the estimated treatment effects and the true treatment effects at different quantile levels from model (3.1) averaging on 100 repetitions.

The predicted CDF is achieved from 5-fold cross-validation. For each repetition, we divide the whole data set into 5 fold. The predicted treatment effect function for each patient in the remaining data set $\widehat{\ }_T(\ ;\mathbf{x}\ )$ is obtained according to Section 2.3.3. Such that, we can get the predicted conditional quantile process $\widehat{Q}_y\ (\ ;\mathbf{x}\ ) = \mathbf{z}_i{}^{>}\,\widehat{\ }_T(\ ;\mathbf{x}\ )$. The treatment assignment is determined by the third principle in Section 2.3.3. Consequently, it is easy to obtain the predicted probability that the value of the response variable is bigger than 11 and 16. The probabilities are $p_1 = \inf\{\ _1 : \widehat{Q}_y\ (\ _1;\mathbf{x}\ ) > 11\}$ and $p_2 = \inf\{\ _2 : \widehat{Q}_y\ (\ _2;\mathbf{x}\ ) > 16\}$. We calculated the median of the probability among all the patients for each repetition.

Figure 3.9 displays the box-plots of the probabilities derived from the proposed approach. We also include the results based on the two alternative approaches ITR with AIPWE and ITR with ITRIPWE (Doubleday et al., 2018). The median probability that the QIDS score is bigger than 11 by the proposed CQRF is 0.158, which is much smaller than 0.257 and 0.258 for ITR random forest. The median probability that the QIDS score is bigger than 16 is 0.059, which also apparently performances better the other two. The results demonstrate that treatment assignment by using the proposed random forest leads to a smaller value of QIDS score, which indicates less severeness of the depression.

We also calculate the bias $\bar{B}_b$ between the estimated treatment effect and the true effect derived fomr Equation 3.1 at quantile levels 0.1, 0.2,..., and 0.9. Table 3.3 shows the resulting median of the biases and its median absolute deviation. The small biases indicate that the estimated treatment effect function is consistent with the true conditional treatment effect function.

Figure 3.9: The box plots of the probability that QIDS score is bigger than 11 and 16 based on the predicted distributions. The box plots are the box plots of the 100 probabilities with 11 and 16 as thresholds. Red ones are box plots of probabilities by using proposed conditional quantile random forest. Green and blue ones are box plots by using ITR with IPWE and AIPWE.

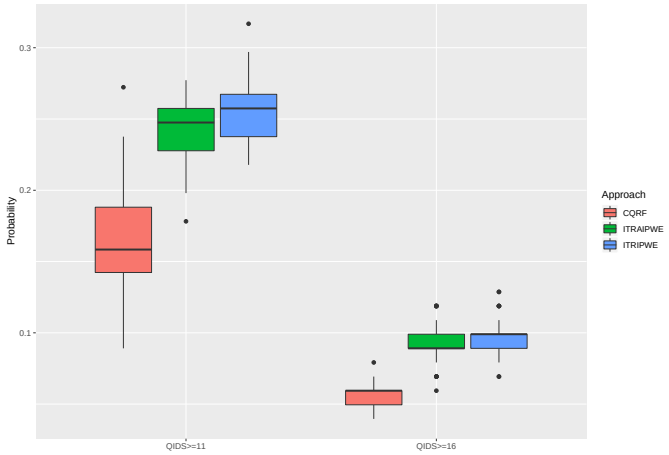


Table 3.3: Simulation result for simulated data based on random forest with 500 trees when the true coefficients at 100 quantile levels are fixed. The covariates and controlling variables in the simulated data are randomly sampled from REACH data and response variable is sampled from the true quantile processes. Biases are calculated for each random forest, each controlling variable and 9 equally spaced quantile levels between 0 and 1. The median and the median absolute deviation (MAD) for estimation bias of each controlling variable at different quantile levels are calculated. MAD is in parenthesis.

k	$\bar{B}_b$
0.1	0.15(0.04)
0.2	0.03(0.10)
0.3	-0.04(0.12)
0.4	-0.06(0.10)
0.5	-0.05(0.06)
0.6	-0.05(0.00)
0.7	-0.10(0.00)
0.8	-0.18(0.02)
0.9	-0.30(0.02)

3.9 Conditional Quantile Random Forest with Linear Splits

Univariate splits are straightforward to understand and easy to be implemented, while linear splits tend to improve the prediction accuracy, particularly when the variables are highly

correlated. Replacing the univariate splits in the proposed CQRF with linear splits will make the proposed approach adjusted to the data set with substantial highly correlated variables, primarily when those splitting variables are clustered by some inherited attributes. For example, genotyped single-nucleotide polymorphisms (SNPs) data in genome-wide association study contains genotypes of SNPs, which are clustered according to the gene regions where they locate. Linear splits enable us to consider all the SNPs locating in the same gene region all together instead of one SNP a time as a splitting variable.

3.9.1 Criterion to Choose the Optimal Linear Split

The non-parametric interaction quantile model described in Section 2.2.3 still holds when univariate splits become linear splits in the algorithm. Here we recall some essential notations. Let $(y_i; \mathbf{x}_i; \mathbf{z}_i);\ i = 1; \cdots; n$ be a random sample from a target population, where y_i is some measurement of disease trait of interest, such as Body Mass Index (BMI), which is an indicator of relative obesity. Splitting variables $\mathbf{x}_i$ contain SNPs data locating in G gene regions. Within each gene region, there are $p_g; g = 1; 2 \cdots; G$ SNPs. In particular, $\mathbf{x}_i$ is the concatenate of $\mathbf{x}_i^g; g = 1; 2 \cdots; G$ and $\mathbf{x}_i^g$ is a p_g dimensional row vector. Thus, $\mathbf{x}_i$ is a $p = \sum_{g=1}^{G} p_g$ dimensional vector. Environmental factor $\mathbf{z}_i$ is a 1 dimensional vector which takes the place of controlling variables in Model (2.1). If $\mathbf{z}_i$ is a binary variable of environmental exposure status, $\quad_1(\ ; \mathbf{x}_i)$ in Model (2.1) catches heterogeneous influences of the environmental factor conditional on the similar gene variants. And also it measures the interaction effect in the distribution level of response variable y. Similary as in Model (2.2), we abuse $\mathbf{z}_i$ to let it contain 1 and a 1 dimensional vector of environmental factor, and $\quad(\ ; \mathbf{x}_i) = (\ _0(\ ; \mathbf{x}_i);\ \ _1(\ ; \mathbf{x}_i))$.

The new proposed splitting criterion tries to find an optimal linear combination $\mathbf{x}_{ig}^{>} \quad_g$ with its corresponding parameter $\quad_g$ and cutoff value c_g within each split. The proposed conditional quantile correlation is motivated by the idea of the quantile partial correction in Li et al. (2015). Let random variable Y be the response variable, $\mathbf{Z}$ be the controlling variable and $\mathbf{X}$ be the covariates. They propose a measurement of the linear relationship

between th conditional quantile of the response variable Y and covariate $\mathbf{Z}$ after adjusting for variable $\mathbf{X}$. We replace the score function of the response variable Y against $\mathbf{X}$ at th quantile with its regression rank score (Koenker et al., 2010). We assume that the random vector $(Y; \mathbf{X}; \mathbf{Z})^{>}$ has the joint density with $\mathsf{E}\, jj\, \mathbf{X}\, jj^2 < 1$ and $\mathsf{E}\, jj\, \mathbf{Z}\, jj^2 < 1$ The conditional quantile correlation is defined as,

$$\begin{aligned} \text{cqcor}\{Y; \mathbf{X}^{>} \quad \mathbf{Z}\, j\, \mathbf{Z}g &= \frac{cov\{\int_0^1 \quad (Y \quad \mathbf{Z}^{>} \tilde{} (\ ))d\ ; \tilde{\mathbf{Z}}_{\mathbf{X}}g}{\sqrt{\text{var}_{Y\, j\mathbf{Z}}\{\int_0^1 \quad (Y \quad \mathbf{Z}^{>} \tilde{} (\ ))d\ g}\sqrt{\text{var}_{\mathbf{X}j\mathbf{Z}}\{\tilde{\mathbf{Z}}_{\mathbf{X}}g}} \\ &= \frac{\mathsf{E}_{Y;\mathbf{X}j\mathbf{Z}}\{\int_0^1 \quad (Y \quad \mathbf{Z}^{>} \tilde{} (\ ))d\ \tilde{\mathbf{Z}}_{\mathbf{X}}g}{\sqrt{\frac{1}{12}\text{var}_{\mathbf{X}j\mathbf{Z}}\{\tilde{\mathbf{Z}}_{\mathbf{X}}g}}; \end{aligned} \tag{3.3}$$

where $(u) = \quad 1\{u \quad 0g$, $\tilde{}(\ ) = \arg\min \quad (y \quad \mathbf{Z}^{>}\ )$ and we project $\mathbf{X}^{>} \quad \mathbf{Z}^{>}$ onto the orthogonal complimentary space of $\mathbf{Z}$: $\tilde{\mathbf{Z}}_{\mathbf{X}} = (\mathbf{I} \quad \mathbf{Z}^{>}(\mathbf{Z}\mathbf{Z}^{>})^{\ 1}\mathbf{Z})\mathbf{X}^{>} \quad \mathbf{Z}^{>}$. By the assumption of the joint density, the values of $\tilde{}(\ )$ are unique. It can be shown that $\mathsf{E}_{Y\, j\mathbf{Z}}(\int_0^1 \quad (Y \quad \mathbf{Z}^{>} \tilde{}_z(\ ))d\ ) = \mathbf{0}$ and $\mathsf{E}_{\mathbf{X}j\mathbf{Z}}(\tilde{\mathbf{Z}}_{\mathbf{X}}) = \mathbf{0}$. According to Koenker et al. (2010) and Gutenbrunner et al. (1993), $var_{y j\mathbf{Z}}\{\int_0^1 \quad (y \quad \mathbf{Z}^{>} \ _1(\ ))d\ g = \int_0^1 \ ^2 d \quad (\int_0^1 \ d\ )^2 = \frac{1}{12}$.

Next, we are going to introduce how we obtain sample conditional quantile correlation between the response variable y_i and the interaction term $\mathbf{x}_i^{>} \quad \mathbf{z}_i^{>}$ within each node. Let $\hat{}(\ ; N) = \arg\min \ \sum_{i=1}^{n} \quad (y_i \quad \mathbf{z}_i^{>}\ )1\{\mathbf{x}_i \ 2\ N g$ be the estimated quantile coefficient at th quantile level. Then the rank score for each sample within the node N is $\hat{b}_{N\, i} = \int_0^1 \quad (y_i \quad \mathbf{z}_i^{>} \hat{}(\ ; N))1\{\mathbf{x}_i \ 2\ N gd\ $. The sample conditional quantile correlation in the node N is,

$$\text{scqcor}_N\{y; diag(\mathbf{x}^{>} \quad)\mathbf{z}\, j\, \mathbf{z}g = \frac{1}{\sqrt{\frac{1}{12}\widehat{\text{var}}_{\mathbf{x}j\mathbf{z};N}\{\tilde{\mathbf{z}}_{\mathbf{x}}g}}\frac{1}{n}\sum_{i=1}^{n}\hat{b}_{N\, i}\tilde{z}_{\mathbf{x}i}1\{\mathbf{x}_i \ 2\ N gg; \tag{3.4}$$

where $\tilde{\mathbf{z}}_{\mathbf{x}} = (\mathbf{I} \quad \mathbf{z}^{>}(\mathbf{z}\mathbf{z}^{>})^{\ 1}\mathbf{z})diag(\mathbf{x}^{>} \quad)\mathbf{z}^{>}$ is an n dimensional column vector and $diag(\ )$ is a function making the input vector as a diagonal matrix whose diagonal entries are the input vector. Each element in $\tilde{\mathbf{z}}_{\mathbf{x}}$ is denoted as $\tilde{z}_{\mathbf{x}i}$ and its variance is $\widehat{\text{var}}_{\mathbf{x}j\mathbf{z};N}\{\tilde{\mathbf{z}}_{\mathbf{x}}g = n^{\ 1}\sum_{i=1}^{n}\tilde{z}_{\mathbf{x}i}^2 1\{\mathbf{x}_i \ 2\ N g$. Previously, we state that $\mathbf{x}_i$ is the concatenate of $\mathbf{x}_i^g$. Thus, we

have $\mathbf{x}_i^{>} = \sum_{g=1}^{G} \mathbf{x}_i^{g>}\ _g$, where $\mathbf{x}_i^g$ is a p_g dimensional row vector and $_g$ is a p_g dimensional column vector.

We assume that only one $_g$ has non-zero entries every time that $\mathbf{x}_i = \mathbf{x}_i^{g>}\ _g$. We find a corresponding optimal $_g$ each time and repeat the procedure G times. The optimal g with its corresponding $_g$ is chosen by maximizing the absolute value of the resulting sample conditional quantile correlation, that is,

$$\begin{aligned}(g\ ;\ ^{\sim}{}_g\ ) &= \arg\max_{g;\ _g} j\ scqcor_N\ fy; diag(\mathbf{x}^{g>}\ _g)\mathbf{z}\ j\ zg\ j \\ &= \arg\max_{g;\ _g} j\ \frac{1}{\sqrt{\frac{1}{12}\widehat{\mathrm{var}}_{\mathbf{x}jz;N}\ f\tilde{\mathbf{z}}_{\mathbf{x}g}g}}\frac{1}{n}\sum_{i=1}^{n}\hat{b}_{N\,i}\tilde{\mathbf{z}}_{\mathbf{x}gi}\ j; \qquad (3.5)\end{aligned}$$

where $\tilde{\mathbf{z}}_{\mathbf{x}g} = (\mathbf{I}\quad \mathbf{z}^{>}(\mathbf{z}\mathbf{z}^{>})\ ^{1}\mathbf{z})diag(\mathbf{x}^{g>}\ _g)\mathbf{z}^{>}$. Mean and variance are defines similarly as in Model (3.4). Due identifiability issue, we require that $jj\ _g\ jj= 1$. Notably, in practice, we first obtain the optimal $^{\sim}{}_g$ for each $g = 1;:::;G$. We then achieve the optimal g by maximizing the absolute value of correlation among those $^{\sim}{}_g$. After the optimal linear combination is chosen, Splitting Criterion I in Section 2.2.3 is employed to select the optimal splitting value c_g .

3.9.2 Variable Importance

Likewise in Section 2.2.3, we propose to adopt the absolute value of the sample conditional quantile correlation (Equation 3.4) for each g with its optimal $^{\sim}{}_g$ as the variable importance score for each gene region. Let $\mathrm{scqcor}_{\mathbf{T}_b lg}$ be the sample conditional quantile correlation of gth genes at the lth splits in bth tree. The final score is calculated by the mean of all scores in every split, every tree in the random forest,

$$\mathrm{scqcor}_g = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{L_b}\ j\ \mathrm{scqcor}_{\mathbf{T}_b lg}\ j\ :$$

We know from Section 3.9 that the conditional quantile correlation is a rescaled version of $_2$. Therefore, the proposed importance score is a proper measurement of how the interaction term between an environmental factor and genotypes variables affect the quantile process of the response variable.

3.9.3 Application on a Body Mass Index (BMI) Data Set

Obesity is a worldwide epidemic which leads to increased morbidity and mortality. More and more people suffer from it. The most recent report from the Centers for Disease Control and Prevention (CDC) indicates that 18.5% of children (aged 2-19) and 39.8% of adults (aged 20 years and older) have obesity, which defined by a body mass index (BMI) above 95% percentile of sex-specific BMI and 30 kg/m^2, correspondingly Hales et al. (2017). The obesity epidemic is a recent hot research topic which manifests the heterogeneous among the different population based on genetic variability and interaction with the environment factor. It is accepted that among several environmental factors, including dietary nutrients, age, gender ethnicity, etc., gene-nutrition interaction is a primary cause of one's susceptibility of becoming obese. The study of gene-nutrition interactions may help us to understand the obesity traits better and to develop more effective personalized weight loss intervention.

The data set we have is from Breast Cancer Family Registry (BCFR) study. The BCFR was established in 1995 as an international collaborative resource, which helps researchers to study genetic and environmental causes of breast cancer. We obtain a subset from BCRF data set, which contains participants' BMI values, nutrition information, and genotyped SNP information. All participants in the data set are female. BMI is the response variable of interest. Among all the genotyped SNPs collected in BCRF study, we cross-check with the BMI-associated genes identified by Locke et al. (2015) and choose the SNPs whose positions are within the regions of those genes. Finally, we have 40 individual SNPs and SNPs within the regions of 54 genes. Thus, there are 3178 SNPs contained in the data set. We calculate the frequencies of alleles for each SNP and identify the minor allele within the given samples. The genotype data is transferred to a number which counts how many

minor alleles contained. The possible values are $0; 1; 2$, which refer to the number of minor alleles. Environmental factors include age, and calories, protein, fat, calcium, phosphorus, sodium, and various vitamins intake. We first conduct a preliminary screening of those environmental factors. The results of simple linear regressions show that there are 5 factors having significant marginal linear effects on BMI. They are age and calories, fat, calcium, and sodium intake. Accordingly, we have 2653 participants with BMI as the response variable, 3178 genotyped SNPs, and 5 environmental factors in the data set.

We apply the proposed CQRF with linear splits to the data set where each environmental factor serves as the controlling variable and genotyped SNPs contained in different genes serve as potential splitting variables. Figure 3.10 shows the resulting variable importance scores of the top Genes for the 5 different environmental factors. For the environmental factors like age, calories intake, and sodium intake, the importance score dramatically decreases after those genes with outstanding scores, while for other factors, the scores decrease smoothly. In Figure 3.11, we list the gene's names which rank top 20 for each environmental factor. Among all the genes, there are 10 of them ranking top 20 for all of the 5 environmental factors; 9 of them ranking top 20 for all but calories intake. As to calories intake, there are 8 unique genes having the ranking top 20 in terms of importance score.

In order to evaluate the genes chosen by the proposed important score, we follow the following steps for each environmental factor,

1. We rank the genes according to the important scores, and choose the gene with the largest scores.

2. We sum up the number of minor alleles of each SNP in the gene and denote it as $G_i; i = 1; 2; \cdots; n$. The value G_i serves as a measurement of the mutant in this gene for ith observation.

3. We divide the samples into two parts according to the median value of G_i.

4. Within each part, we fit quantile regression models between BMI and the environmental factor at quantile levels $= (1{=}100; 2{=}100; \cdots; 99{=}100)$ and obtain the estimated

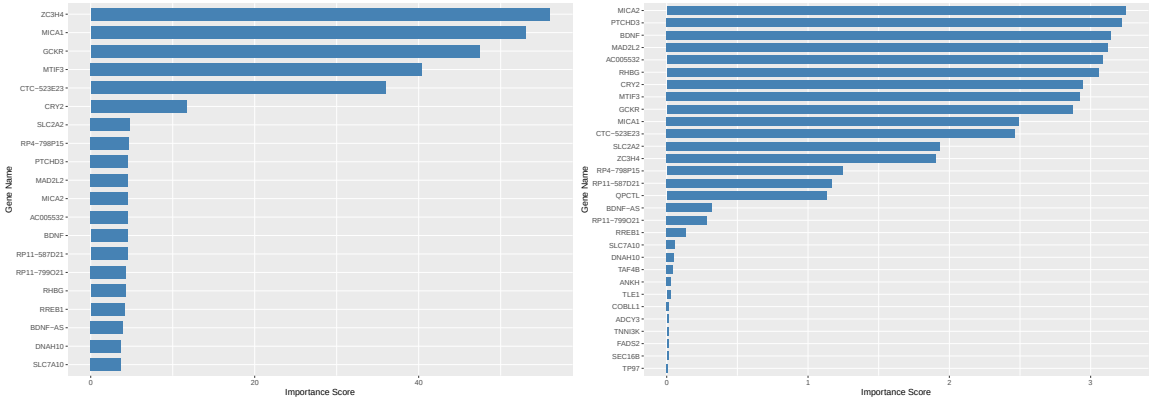
quantile coefficients. The estimated quantile process is acquired by nature linear spline with the estimated coefficients as internal knots.

Figure 3.12 demonstrates that for all 5 environmental factors, the top-ranked gene well separates the samples that they have heterogeneous effects of the environmental factor. Especially for calories intake, when the mutant measurement of gene ENSG00000204520 (MICA2) is above the sample median, the calories intake has a minor effect on the value of BMI. When the mutant measurement is below the sample median, the effect of calories intake becomes more extensive and even more extensive as the value of BMI increases. The result for calories intake shows that slim people have genetic advantages to maintain their weight.

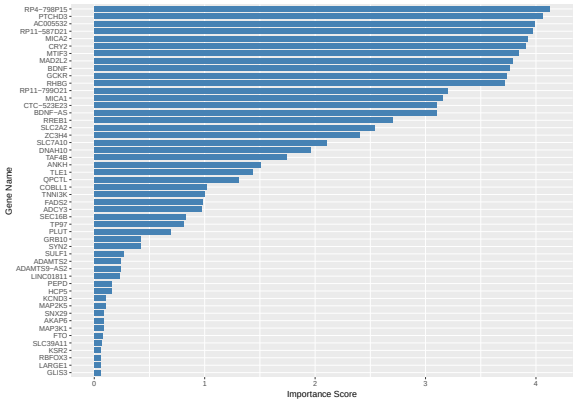
Figure 3.10: Variable Importance: The variable importance scores for different genes are obtained from the proposed CQRF with linear splits.

(a) The Variable Importance Scores for Top Genes When **age** is the environmental factor, which is the controlling variable in the interactive quantile model.

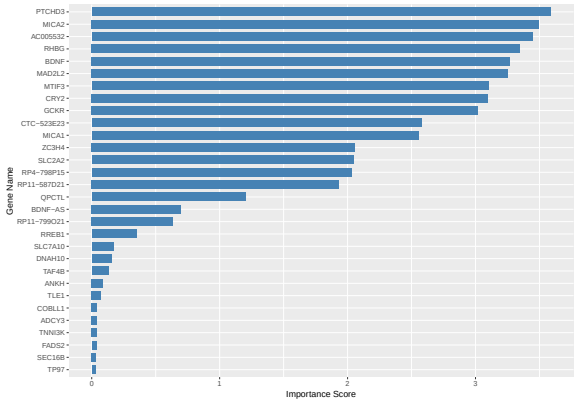
(b) The Variable Importance Scores for Top Genes When **calories** intake is the environmental factor, which is the controlling variable in the interactive quantile model.



(c) The Variable Importance Scores for Top Genes When **fat** intake is the environmental factor, which is the controlling variable in the interactive quantile model.



(d) The Variable Importance Scores for Top Genes When **calcium** intake is the environmental factor, which is the controlling variable in the interactive quantile model.



(e) The Variable Importance Scores for Top Genes When **sodium** intake is the environmental factor, which is the controlling variable in the interactive quantile model.

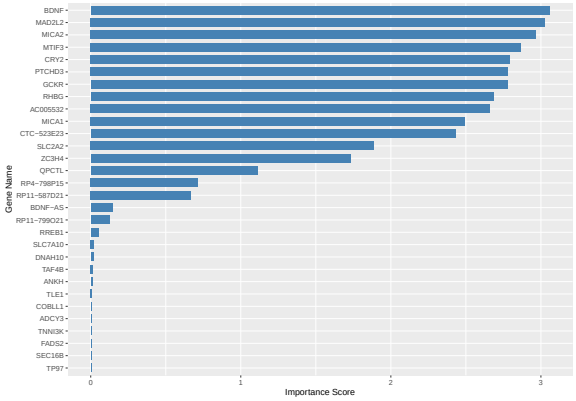


Figure 3.11: The list of genes which rank top 20 in terms of the variable importance for the 5 environmental factors.

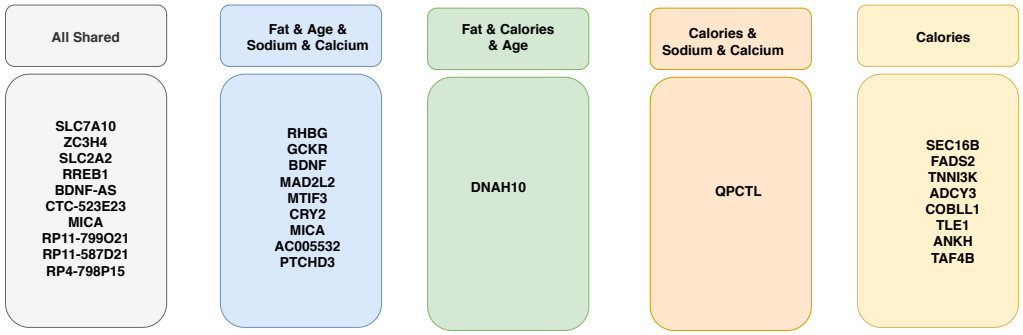
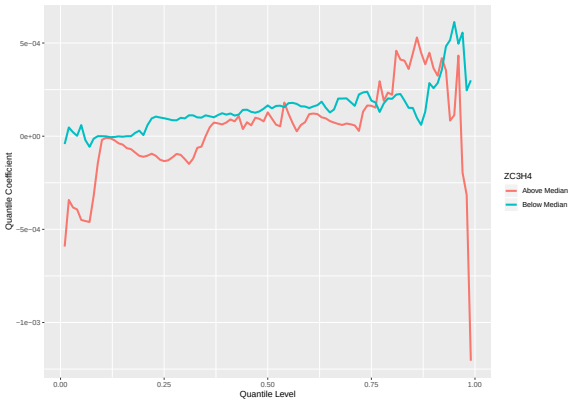


Figure 3.12: The top ranked gene separates the samples into two parts according to the median of G_i. Within each part, the estimated quantile process is obtained.

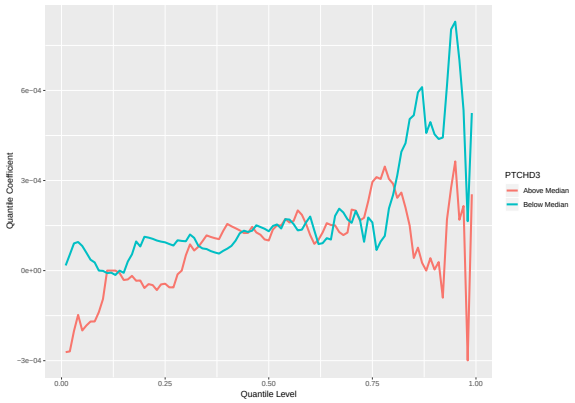
(a) Age



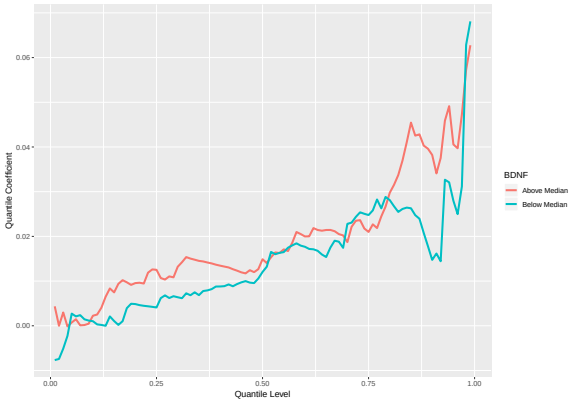
(b) Calories intake

(c) Fat intake

(d) Calcium intake



(e) Sodium intake



Chapter 4

Generalized Integrative Principal Component Analysis for Block-Wise Missing Data

4.1 Overview

In this chapter, we introduce a generalized interactive principal component analysis (GIPCA) approach to for the dimension reduction of multi-source data that allows different sources to have different data types with block-wise missing structure. In Section 4.2, the proposed models and identifiability issues are introduced for non-missing and block-wise missing data. In Section 4.3, we introduce the algorithm and the rank selection. In Section 4.4, we conduct comprehensive simulation studies to evaluate the performance of the proposed method and compare with existing methods. In Section 4.5, we apply the proposed method to a mortality study and discuss the performance of estimation and imputation by comparison with several ad hoc methods.

4.2 Generalized Integrative PCA Model

Let $\mathbf{X}_k$ $(k = 1;:::K)$ be an n p_k data matrix, with n being the number of samples and p_k being the number of variables. Samples are matched across K data sources. Each entry in the data matrix $\mathbf{X}_k$ is a realization of a random variable from an exponential family distribution. The entries of different data matrices may follow different distributions (e.g., Gaussian, Poisson, binomial), while those in the same data matrix are assumed to have the same distributional form. That is, each entry $x_{k;ij}$ in the kth data set is a realization of a random variable following a single-parameter distribution in the exponential family with an underlying natural parameter $_{k;ij}$. The canonical form of the probability density function for each entry can be expressed as,

$$x_{k;ij} \quad f_k(x_{k;ij}j\ _{k;ij}) \ / \ \exp fx_{k;ij}\ _{k;ij} \quad b_k(\ _{k;ij})g:$$

where $b_k(\ _{k;ij})$ is a convex function which defines the distribution. The canonical link function for the generalized linear regression is $g_k(\ _{k;ij}) = b_k^{0\ 1}(\ _{k;ij})$, where $_{k;ij}$ is the mean of $x_{k;ij}$. The entries are assumed independent given the underlying natural parameters. We denote the underlying natural parameters matrix for $\mathbf{X}_k$ as $\mathbf{\Theta}_k$. The natural parameter matrix for all data is denoted as $\mathbf{\Theta} = (\mathbf{\Theta}_1;:::;\mathbf{\Theta}_K)$, which has $p = \sum_{k=1}^{K} p_k$ columns.

4.2.1 Model for Non-Missing Data

We first discuss our proposed model in the context of non-missing (complete) data. For the integrated analysis of multi-source data sets, both shared and individual structure should be considered in the decomposition procedure Lock et al. (2013). The natural parameter matrix $\mathbf{\Theta}_k$; for each data set, is decomposed into joint and individual latent components as follows:

$$\mathbf{\Theta}_k = \mathbf{1}\ _k^T + \mathbf{U}_0\mathbf{V}_k^T + \mathbf{U}_k\mathbf{A}_k^T: \tag{4.1}$$

In Model (4.1), $_k$ is the column means of natural parameters and $\mathbf{1}$ is an n 1 vector of all 1s. Thus, natural parameters within one matrix may have different column means. The second term $\mathbf{U}_0\mathbf{V}_k^T$ represents the shared structure among different data sources, where $\mathbf{U}_0$ is an n r_J joint score matrix among K data sets and $\mathbf{V}_k$ is a p_k r_J joint loading matrix for kth data set, with r_J $\min(n; p_1; p_2; \ldots; p_K)$ being the rank of the joint structure. The individual structure is denoted by $\mathbf{U}_k\mathbf{A}_k^T$, where $\mathbf{U}_k$ is an n r_{A_k} individual score matrix and $\mathbf{A}_k$ is a p_k r_{A_k} individual loading matrix. The individual rank for each data set is r_{A_k}, and r_{A_k} $\min(n; p_k)$.

Equivalently, the decomposition of the natural parameter matrix $\boldsymbol{\Theta}$ can be expressed as follows:

$$\boldsymbol{\Theta} = (\mathbf{1}; \mathbf{U}_0; \mathbf{U}_1; \ldots; \mathbf{U}_K) \begin{pmatrix} {}_1^T & \ldots & {}_K^T \\ \mathbf{V}_1^T & \ldots & \mathbf{V}_K^T \\ \mathbf{A}_1^T & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{A}_K^T \end{pmatrix} = (\mathbf{1}; \mathbf{U}_0; \mathbf{U}_A) \begin{pmatrix} {}^T \\ \mathbf{V}^T \\ \mathbf{A}^T \end{pmatrix} \tag{4.2}$$

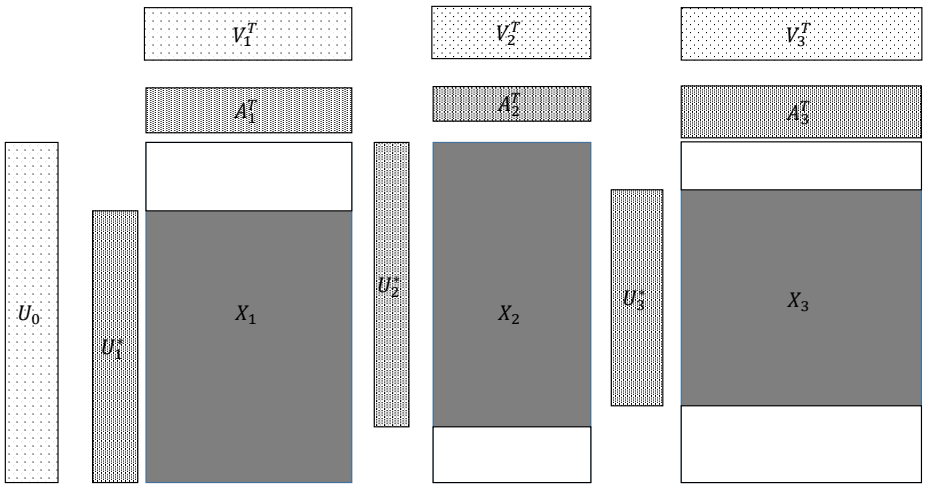
where $= ({}_1^T; \; ; {}_K^T)^T$ is the concatenation of the column means for each data set, $\mathbf{V} = (\mathbf{V}_1^T; \ldots; \mathbf{V}_K^T)^T$ is the concatenation of the joint loading matrices, $\mathbf{U}_A = (\mathbf{U}_1; \ldots; \mathbf{U}_K)$ is the concatenation of the individual score matrices, $\mathbf{A} = diag(\mathbf{A}_1; \ldots; \mathbf{A}_K)$ is a block-wise diagonal matrix, and $\mathbf{0}$ represents any zero matrix with compatible size.

In particular, when there is only one data set, our proposed GIPCA reduces to the decomposition of one natural parameter matrix, which coincides with the exponential family principal component analysis (EPCA, Collins et al., 2002). Under the Gaussian assumption with equal variance, the decomposition of natural parameter matrix reduces to a factorization of the original multi-source data set. Thus, our Model (4.2) is identical to JIVE Lock et al. (2013) in this context. With just two data sets, Model (4.2) coincides with the GAS model Li and Gaynanova (2017) applied to data sets without missing values.

4.2.2 Model with Block-wise Missing Data

Figure 4.1: GIPCA for block-wise missing data

The three big rectangles represent three data sets $\mathbf{X}_k$ ($k = 1; 2; 3$) with block-wise missing values (i.e., blank strips). The horizontal direction (rows) represents samples in the three data sets. And the vertical direction (columns) represents variables. The grey color in the big rectangles means that the data are observed for the corresponding samples in the corresponding sources. The blank rectangles are the block-wise missing entries. Those rectangles on the side are joint score and loading matrices $\mathbf{U}_0$, $\mathbf{V}_k$ and individual score and loading matrices $\mathbf{U}_k$, $\mathbf{A}_k$.



We extend the model described in Section 4.2.1 to allow for block-wise missing structure. Figure 4.1 is an illustrative picture of data sets with block-wise missing. Due to the block-wise missing entries in the data sets, the corresponding rows in the individual score matrices $\mathbf{U}_k$ in Model (4.2) are missing. Thus we denote the submatrix of $\mathbf{U}_k$ containing only rows without block-wise missing as $\mathbf{U}_k$. The joint score matrix $\mathbf{U}_0$ remains the same as in Model (4.2) because for all samples at least one data source is with complete observations, which helps us identify joint structure.

With block-wise missing data, for each data set k, the decomposition of the natural

parameter matrix underlying the observed data becomes

$$\mathbf{\Theta}_k = \mathbf{1} \quad {}_k^T + \mathbf{U}_0^{[k]} \mathbf{V}_k^T + \mathbf{U}_k \mathbf{A}_k^T; \tag{4.3}$$

where $\mathbf{\Theta}_k$ is an $n_k \quad p_k$ matrix (a submatrix of $\mathbf{\Theta}_k$ in Model (4.1)). The joint score matrix $\mathbf{U}_0^{[k]}$ is an $n_k \quad r_J$ submatrix of $\mathbf{U}_0$, where only the rows corresponding to the complete samples in the kth data source are kept. The individual score matrix $\mathbf{U}_k$ is an $n_k \quad r_{A_k}$ matrix. The means $\quad_k$, the joint and individual loading matrices $\mathbf{V}_k$, $\mathbf{A}_k$ remain the same as in Model (4.2). We also note that $\mathbf{1}$ is an $n_k \quad 1$ vector of all 1s. When there is no missing value, Model (4.3) exactly coincides with Model (4.2).

We remark that despite the block-wise missingness, the joint structure in Model (4.3) across data sources is $\mathbf{U}_0\mathbf{V}$. For a sample with block-wise missing values, as long as it has observations in some data sources, it provides information towards the shared structure. Thus, the underlying joint score matrix is complete, regardless of the block-wise missing structure. The mechanism of block-wise missing imputation relies on the joint structure. Such shared information among different data sets informs the missing data for each data source. Specifically, once estimated, the means and the joint structure can be effectively used to impute block-wise missing data.

4.2.3 Identifiability Conditions

In order to ensure identifiability of the estimation, the model parameters should satisfy certain conditions. Following the discussion in Lock et al. (2013); Li and Gaynanova (2017), we provide the identifiability conditions for Model (4.3) as the following.

1. The columns of the score matrices $\mathbf{U}_0$, $\mathbf{U}_k$ are linearly independent and the columns of the means $\quad_k$ and the loading matrices $\mathbf{V}$, $\mathbf{A}_k$ within each data set are linearly independent.
2. All the score matrices are column-centered and the column space of the joint score matrix is orthogonal to the column space of the individual score matrices.

3. All the separate score and loading matrices have orthogonal columns.

The first condition ensures the joint and individual structures are clearly separable. The second orthogonality condition enhances the interpretability by requiring that the means, joint and individual structures are orthogonal to each other. The third condition rules out arbitrary rotations within each subspace. The above conditions guarantee that the model is fully identifiable (up to some trivial order switch and scale change).

4.3 Algorithm

In this section, we explain how we estimate each parameter in Model (4.3). We first assume the ranks for the shared and individual structures are known and devise an iterative algorithm for model fitting. Then we introduce an adaptive BIC procedure for rank selection, which is tailored for the proposed approach.

4.3.1 GIPCA algorithm

The unknown parameters in Model (4.3) are estimated by maximizing the joint log likelihood. Under the assumption that individual measurements are mutually independent given the underlying natural parameter matrix, the maximum likelihood estimators (MLE) are

$$(\hat{\ }; \hat{\mathbf{U}}_0; \ \hat{\mathbf{U}}_1; \quad ; \hat{\mathbf{U}}_K; \ \hat{\mathbf{V}}; \ \hat{\mathbf{A}}) = \arg\max_{\mathbf{\Psi}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{p_k} \log f_k(x_{k;ij} j_{\ k;ij}) \tag{4.4}$$

where $\mathbf{\Psi} = f_{\ k}; \mathbf{U}_0; \mathbf{V}_k; \mathbf{U}_k; \mathbf{A}_k; \ k = 1; : : : ; K g$ is the set of unknown parameters, $\mathbf{\Theta}_k = (\ _{k;ij})$ has the decomposition structure in Model (4.3) and $f_k(\ )$ is the probability density function in each data source.

It is computationally prohibitive to directly maximize the log likelihood because the objective function is not convex with respect to all the parameters. As a remedy, we exploit a block coordinate descent algorithm to estimate the parameters. Namely, we alternatively estimate the joint structure along with the intercept and the individual structures until

converge. More specifically, we

fix $\mathbf{U}_k$ and $\mathbf{A}_k$ for all data sets, and estimate $_k$, $\mathbf{U}_0$ and $\mathbf{V}$;

fix $_k$, $\mathbf{U}_0$ and $\mathbf{V}_k$, and estimate $\mathbf{U}_k$ and $\mathbf{A}_k$ in each data set.

Consequently, the computation is significantly simplified. We shall provide more details below.

We first estimate the intercept and joint structure with the individual structures fixed. To further alleviate the computational burden, we fix the joint score matrix $\mathbf{U}_0$ to estimate the joint loading matrix $\mathbf{V} = (\mathbf{V}_1^T; \quad ; \mathbf{V}_K^T)^T$ and the intercept $\quad = (\ _1^T; \cdots; \ _K^T)^T$. The estimation of each row in $\mathbf{V}$ paired with the corresponding entry in $\quad$ can be cast as a generalized linear model (GLM) estimation problem. More specifically, let $_{kj}$ be jth column of $\mathbf{\Theta}_k$. We have $_{kj} = \mathbf{1}^T \ _{kj} + \mathbf{U}_0^{[k]} \mathbf{v}_{kj}^T + \mathbf{U}_k \mathbf{a}_{kj}^T$, where $_{kj}$ is the jth entry of $_k$, $\mathbf{v}_{kj}$ is the jth row of $\mathbf{V}_k$, and $\mathbf{a}_{kj}$ is the jth row of $\mathbf{A}_k$. The estimation of $\mathbf{v}_{kj}$ and $_{kj}$ can be obtained by fitting a GLM with the canonical link function, and $\mathbf{U}_k \mathbf{a}_{kj}^T$ being the offset. Similarly, when we fix the joint loading matrix $\mathbf{V}$ to estimate the joint score matrix $\mathbf{U}_0$, again this can be formulated as a GLM problem. Let $_i = (\ _{1i}^T; \cdots; \ _{Ki}^T)^T$, where $_{ki}$ is the column vector of ith row of $\mathbf{\Theta}_k$. We have $_i = (\ _1^T; \cdots; \ _K^T)^T + \mathbf{V}\mathbf{u}_{0i} + (\mathbf{A}_1 \mathbf{u}_{1i}; \cdots; \mathbf{A}_K \mathbf{u}_{Ki})$, where $\mathbf{u}_{0i}$ is the column vector of the ith row in joint score matrix $\mathbf{U}_0$ and $\mathbf{u}_{ki}$ is the column vector of the ith row in individual score matrix $\mathbf{U}_k$. We remark that the standard GLM model fitting procedure cannot be directly applied to the estimation of $\mathbf{u}_{0i}$ since the canonical link functions are different for different data types across multiple sources. To address this, we follow Li and Gaynanova (2017) and adopt an iteratively reweighted least squares algorithm (IRLS, McCullagh and Nelder, 1989) to accommodate heterogeneous link functions.

Then we estimate individual structures with fixed $_k$ and the joint score $\mathbf{U}_0$ and loading $\mathbf{V}$. Based on Model (4.3), the estimation of individual structures is directly separable for each data source. We still exploit the alternating algorithm to estimate $\mathbf{U}_k$ and $\mathbf{A}_k$. Similar to the estimation of the joint structure, the estimation of $\mathbf{U}_k$ can be parallelized as n_k GLMs, and the estimation of $\mathbf{A}_k$ can be parallelized as p_k GLMs. The estimated parameters are

then plugged into the log likelihood in Model (4.4).

The estimates in $\hat{\boldsymbol{\Psi}} = f\hat{}_k; \hat{\mathbf{U}}_0; \hat{\mathbf{U}}_k; \hat{\mathbf{V}}; \hat{\mathbf{A}}_k g$ in each iteration may not meet the identifiability conditions in Section 4.2.3. Some regularization procedure is desired so that the conditions are satisfied and the likelihood values are unchanged. In order to achieve that, after each iteration, we transform the estimated parameters $\hat{\boldsymbol{\Psi}}$ as follows. Define the projection matrix of the column space of $(\mathbf{1}; \mathbf{U}_0)$ as $\mathbf{P}_J$. We want to project the individual score matrices to the orthogonal complement of the column space of $(\mathbf{1}; \mathbf{U}_0)$. However, the individual score matrix $\mathbf{U}_k$ does not have the same dimension as the column space of $(\mathbf{1}; \mathbf{U}_0)$. To address this, we define a new estimated individual score matrix $\hat{\mathbf{U}}_k : n \quad p_k$ based on $\mathbf{U}_k$, where the missing observations are filled with 0. Then we get the projected individual score matrix $(\mathbf{I} \quad \mathbf{P}_J)\hat{\mathbf{U}}_k$. Column-center the submatrix of the projected individual score matrix containing only complete samples and denote it as $\check{\mathbf{U}}_k : n_k \quad p_k$. Apply SVD to the new individual structures $\check{\mathbf{U}}_k\hat{\mathbf{V}}_k^T$, let the left singular vectors absorb the singular values. Let a score matrix $\check{\mathbf{U}}_k : n \quad p_k$ be based on the left singular vector and corresponding block-wise missing rows filled 0. The new joint structure after identifiability modification is the concatenation of K matrices where each is $\mathbf{1}\hat{}_k + \hat{\mathbf{U}}_0\hat{\mathbf{V}}_k^T + \hat{\mathbf{U}}_k\hat{\mathbf{A}}_k^T \quad \check{\mathbf{U}}_k\check{\mathbf{V}}_k{}^T$. The column mean of each new joint structure is $\check{}_k$. Apply SVD to the concatenation of each column-centered joint structure, and let the left singular vectors absorb the singular values. We denote the new score and loading matrices as $\check{\mathbf{U}}_0; \check{\mathbf{V}}$. Consequently, the modified estimators $\check{}_k; \check{\mathbf{U}}_0; \check{\mathbf{U}}_k; \check{\mathbf{V}}; \check{\mathbf{A}}_k$ satisfy all the conditions.

The iterative algorithm terminates when the difference of the log likelihood between the previous step and current step is smaller than a prefixed threshold. Our proposed algorithm is a block coordinate descent algorithm which ensures the log likelihood in each step of the algorithm is non-decreasing. Thus, the algorithm is guaranteed to converge. We summarize the model fitting algorithm with known ranks for our proposed method in Algorithm 1.

After obtaining the estimates of the parameters in $\boldsymbol{\Psi}$, we impute the block-wise missing entries using the shared parameters. More specifically, we use the same procedure as the regularization step mentioned above to get $\hat{\mathbf{U}}_k : n \quad r_{A_k}$. Then, we can have the estimated

Algorithm 1 GIPCA algorithm

Set initial values of each element in Model (4.3) as $\mathbf{\Psi}^{(0)}$
while The convergence criterion does not satisfy **do**
  In the lth iteration:
  Fix the individual structure: $\mathbf{U}_k^{(l-1)}, \mathbf{V}_k^{(l-1)}; k = 1; 2; \cdots; K$,
    Fix $\mathbf{U}_0^{(l-1)}$, estimate each row of $(\ _k; \mathbf{V}_k)$ via GLM.
    Fix $\mathbf{V}^{(l)}$ and $^{(l)}$, estimate each row of $\mathbf{U}_0$ by the adapted IRLS algorithm.
  Fix the means and joint structure $(\ ^{(l)}; \mathbf{U}_0^{(l)}; \mathbf{V}^{(l)})$
    Fix $\mathbf{U}_k^{(l-1)}$, estimate each row of $\mathbf{V}_k$ via GLM.
    Fix $\mathbf{V}_k^{(l)}$, estimate each row of $\mathbf{U}_k$ via GLM.
  Conduct the regularization procedure to satisfy the identifiability conditions.
  Plug the estimated parameters to the log likelihood function.
end while

complete natural parameter matrix, $\hat{\mathbf{\Theta}}_k = \mathbf{1}^T \hat{\ }_k + \hat{\mathbf{U}}_0\hat{\mathbf{V}}_k + \hat{\mathbf{U}}_k\hat{\mathbf{A}}_k$. In particular, the estimated natural parameter matrix for block-wise missing entries is, $\hat{\mathbf{\Theta}}^y{}_k = \mathbf{1}^T \hat{\ }_k + \hat{\mathbf{U}}_0^{[k]^c}\hat{\mathbf{V}}_k$, where $\hat{\mathbf{U}}_0^{[k]^c} : (n \quad n_k) \quad p_k$ is the complement submatrix of $\hat{\mathbf{U}}_0^{[k]}$ in terms of $\hat{\mathbf{U}}_0$. Each vector $\mathbf{1}$ is with the compatible size. By taking the inverse of the link function, the imputed data is $g_k{}^{-1}(\hat{\mathbf{\Theta}}^y{}_k)$.

4.3.2 Rank Estimation: BIC

There are many approaches in the PCA literature to determine the number of principal components or the rank of the latent structure. For example, one may exploit scree plots of eigenvalues to choose the rank that explains a certain proportion of the total variation Jolliffe (1986), or use a hypothesis testing procedure (e.g., Bartlett's test) to determine the rank. There is a large amount of literature on selecting ranks for matrix decomposition with Gaussian assumption. However, there is only a little considering rank estimation for non-Gaussian data. Landgraf and Lee (2015) proposed an approach for binary data based on the percentage of deviance explained by some principal components. As to the rank selection for multi-source data, a permutation testing approach was proposed to JIVE Lock et al. (2013). BIC is another approach that is adapted to JIVE to implement rank selection OConnell and Lock (2016). A two-step cross-validation method Li and Gaynanova (2017)

used the sum of squared Pearson residuals as the criterion to select ranks when modeling heterogeneous data with exponential family distributions assumption. Nevertheless, none of the literature mentioned rank estimation for a multi-source data with block-wise missing entries.

Here, we develop an adapted BIC approach to estimate the joint and individual ranks of the underlying natural parameter matrices for multi-source data. The key to the adapted BIC criterion is to calculate the number of parameters to be estimated in the model. The joint score matrix $\mathbf{U}_0$ has $\sum_{j=n}^{n} {}^{1}_{r_J} j$ entries to estimate since it has centered columns which are orthogonal of each other. For each data set, there are p_k unknown means in Model (4.3). Similarly for the joint and individual loading matrices, they have $\sum_{j^0=p}^{p} {}^{1}_{r_J} j^0$ and $\sum_{j^{00}=p_k}^{p_k} {}^{1}_{r_{A_k}} j^{00}$ parameters to estimate. The individual score matrix $\mathbf{U}_k$ is required to be orthogonal to the joint score matrix and the columns of individual score matrix are centralized and linearly independent of each other. Thus the number of free parameter in the individual score matrix is $\sum_{l=n_k}^{n_k} {}^{(r_J+1)}_{(r_J+r_{A_k})} l$. The number of free parameters in the data sets is,

$$K = \sum_{k=1}^{K} p_k + \sum_{j=n}^{n} {}^{1}_{r_J} j + \sum_{j^0=p}^{p} {}^{1}_{r_J} j^0 + \sum_{k=1}^{K} \sum_{l=n_k}^{n_k} {}^{(r_J+1)}_{(r_J+r_{A_k})} l + \sum_{k=1}^{K} \sum_{j^{00}=p_k}^{p_k} {}^{1}_{r_{A_k}} j^{00}$$

The number observations in data set $\mathbf{X}_k$ is $n_k p_k$. If there is no block-wise missing in the data sets, $n_1 = n_2 = \cdots = n_K = n$. For each combination of $r_J; r_k;\ k = 1; 2; \cdots; K$, a BIC score could be calculated. The value of BIC is calculated as,

$$\text{BIC} = \quad 2l\ (\mathbf{X} j \hat{\mathbf{\Psi}}) + \log(\sum_{k=1}^{k} n_k p_k) K \tag{4.5}$$

where $l\ (\mathbf{X} j \hat{\mathbf{\Psi}})$ is the value of log likelihood given $\hat{\mathbf{\Psi}}$.

In practice, we use a stepwise selection approach to select the ranks via BIC. We first compute BIC for the null model $r_J = 0; r_k = 0; k = 1; 2; \cdots; K$. We add one to or deduct one from each of the ranks at a time and choose the next rank combination with the smallest

BIC value. For instance, assume we have two data sets and start from the BIC score for $r_J = 0; r_k = 0;\ k = 1; 2$. Next, we calculate the BIC value for $r_J = 1\ r_1 = r_2 = 0$, $r_J = r_2 = 0\ r_1 = 1$ and $r_J = r_1 = 0\ r_2 = 1$ and choose the ranks with smallest BIC score. The selection procedure is terminated when the BIC score reaches a local minimum. Then the ranks combination when the procedure is stopped is the estimated ranks.

4.4 Simulation

In this section, we conduct comprehensive simulation studies to validate the proposed method. Since there is no existing method that directly addresses the multi-source multi-type data imputation problem, we come up with two ad hoc approaches to compare with our method.

Ad Hoc 1 (EPCA-PCA): First we estimate a low-rank approximation to the natural parameter matrix of each data set via EPCA. Then, we apply PCA to the concatenated approximations across different data sources.

Ad Hoc 2 (EPCA-SMC): EPCA is first applied to each data set. Then, structured matrix completion (SMC, Cai et al., 2016) is applied to the estimated natural parameter matrices to impute the block-wise missing entries.

The data sets are generated from Model (4.2) and we apply three different methods to the data and impute the block-wise missing entries. For both **Ad Hoc 1** and **Ad Hoc 2** methods, if the Gaussian assumption is satisfied for some data sets, then EPCA step is ignored for such data sets and the original data sets are used in the next step (PCA or SMC). The application of SMC is limited to two data sets when only one of them has block-wise missing entries. Therefore, when we apply SMC to more than one data set has block-wise missing entries, we proceed with one data source at a time. For example, if we have two data sources where both have missing observations, apply SMC approach twice to do imputation for both data sets.

4.4.1 Settings

We set the sample size to be $n = 200$ and the number of variables in each data source to be $p_k = 150$. The joint and individual ranks for the natural parameter matrices are $r_0 = r_1 = r_2 = 2$. Joint and individual score matrices ($\mathbf{U}_0; \mathbf{U}_1; \mathbf{U}_2$) are filled with uniform random numbers $Unif(\ 0.5; 0.5)$ and normalized to have orthonormal columns. In **Scenario 4**, we try 3 data sets with similar settings as the other scenarios. We generate different singular values of joint structure and each individual structure for different scenarios, and the singular values are absorbed by the score matrices.

Scenario 1: **Gaussian-Gaussian** The individual loading matrices $A_1;\ A_2$ are filled with $Unif(\ 0.5; 0.5)$ and normalized to have orthonormal columns. The joint loading matrix $(V_1^T;\ V_2^T)^T$ is generated similarly to have orthonormal columns and is projected to the complement of the column space for the individual loading matrices $diag(A_1^T;\ A_2^T)^T$. The singular values of the joint structure were set to be $(250; 150)$, the singular values of the individual structures to be $(150; 100)$ and $(150; 140)$.

Scenario 2: **Gaussian-Poisson** The procedure to generate individual loading matrices is similar to **Scenario 1**. The joint loading matrices V_1 for Gaussian and V_2 for Poisson are generated from $Unif(\ 1; 1)$, and $Unif(\ 0.25; 0.25)$ respectively. The singular values of the joint structure are set to be $(240; 220)$ for joint, the singular values of the individual structures to be $(90; 80)$ for Gaussian and $(90; 80)$ for Poisson.

Scenario 3: **Gaussian-binomial** The procedure to generate individual loading matrices is similar to **Scenario 1**. The joint loading matrices V_1 for Gaussian, V_2 for binomial are generated from $Unif(\ 0.5; 0.5)$, and $Unif(\ 1.5; 1.5)$ respectively. The singular values of the joint structure are set to be $(240; 220)$ for joint, and the singular values of the individual structures to be $(90; 80)$ for Gaussian and $(100; 80)$ for binomial.

Scenario 4: **Gaussian-Poisson-binomial** The joint loading matrices V_1 for Gaussian, V_2 for Poisson are generated from $Unif(\ 0.5; 0.5)$, and V_3 for binomial is gen-

erated from $Unif(-1.5, 1.5)$. The individual loading matrices A_1 (Gaussian), A_2 (Poisson), A_3 (binomial) are generated from $Unif(-0.5, 0.5)$, $Unif(-0.25, 0.25)$, and $Unif(-1.5, 1.5)$ correspondingly. The singular values of the joint structure are set to be $(300, 280)$, the singular values of the individual structures to be $(150, 120)$ for Gaussian, $(150, 140)$ for Poisson and $(200, 180)$ for binomial.

Scenario 5: **Poisson-binomial** The joint loading matrices V_1 for binomial, V_2 for Poisson are generated from $Unif(-1.5, 1.5)$, and $Unif(-0.5, 0.5)$ respectively.

The means for Gaussian data set in each scenario that contains Gaussian data are generated from $Unif(-0.5, 0.5)$. For Poisson distribution, the inverse of the canonical function makes the realizations skewed to 0 if the natural parameter is a negative number with large absolute value and skewed to a large positive number if the natural parameter is a large positive number. Thus, the scale of the natural parameter matrix for Poisson distribution is required to be smaller in **Scenarios 2, 4, 5**. We also set the means of Poisson distribution to be positive (from $Unif(0, 1)$) to mimic Poisson data in reality. For binomial distribution, we increase the singular values to boost the signal level of binomial data in **Scenarios 3, 4, 5**. The means for binomial data set are generated from $Unif(-1.5, 1.5)$.

When the natural parameters are fixed, data are generated from the corresponding distributions. For Gaussian data, we set the variance for the generated data to be 1. For binomial data, we set the number of trials to be 100. For each simulation, we randomly pick some rows in each data set to be missing. Those rows should not be overlapped over all the data sets to ensure that for each sample, data from at least one data set are without missing. Different missing rates (5% or 10% for rank selection and 5% or 15% for missing imputation) are applied to the generated data when we compare our proposed method with other existing methods. We repeat the procedure multiple times to evaluate the rank selection performance and compare the imputation accuracy of different methods.

4.4.2 Result

When the natural parameter matrix for each scenario is fixed, we apply the rank selection procedure mentioned in Section 4.3.2 to the data generated from corresponding distribution independently for 50 times. BIC criterion (Model (4.5)) is used to estimate ranks for each simulation scenarios with different missing rates. We apply the proposed BIC criterion to all the scenarios with different missing rate 0%; 5%, 10%. The results of rank estimation with different simulation scenarios and missing rates are shown in Table S1 in the supplementary materials.

Overall, the adapted BIC criterion performs well for different settings. The stepwise selection procedure correctly identifies the true ranks for joint structure and individual structures almost all the times for scenarios with two data types with various missing rates. We also apply the selection procedure to **Scenario 4** with three data types: Gaussian, Poisson, and binomial. However, for this scenario the BIC-selected ranks tend to be close to the truth but misallocated; the majority of the 50 simulations select the joint rank to be 3, individual ranks to be 1 for Gaussian and Poisson, and 2 for binomial. This may be because the signal-to-noise ratio for the binomial data is relatively low compared to the other datasets. Alternative approaches to rank selection that can accommodate to multiple (>2) sources of data call for more investigation.

When the natural parameter matrix is fixed, we generate data from corresponding distributions independently for 100 replications. We compare the two ad hoc methods and our proposed method by applying them to the simulated data to estimate elements of $\mathbf{\Psi}$ in Model (4.3). We evaluate the imputation accuracy by the relative Frobenius loss. Mathematically, the relative Frobenius loss is defined as,

$$\text{DiffR}_{Miss} = \frac{||\mathbf{\Theta}_k^y - \hat{\mathbf{\Theta}}_k^y||_F}{||\mathbf{\Theta}_k^y||_F} \tag{4.6}$$

where $\mathbf{\Theta}_k^y$ and $\hat{\mathbf{\Theta}}_k^y$ are true and estimated natural parameter matrices for block-wise missing entries.

The simulation results for two data sets are shown in Table 4.1 and for three data sets are shown in Table S2 in the supplementary materials. For all scenarios, the imputation accuracy of GIPCA outperforms the other ad hoc methods with different distribution combinations and different missing rates. Neither EPCA-PCA nor SMC considers partitioning the joint association from individual structure. We also check the Frobenius norm for the difference between the estimated and true means, the relative Frobenius loss for natural parameter matrix without missing entries. We note that under **Scenario 1**, both EPCA-PCA and GIPCA have similar performance. Under the Gaussian assumption, EPCA-PCA reduces to PCA with the sum of ranks and GIPCA reduces to JIVE without missing entry (Section 4.2.1). Therefore, their estimation accuracy to estimate natural parameter matrix corresponding to samples without block-wise missing is close to each other.

In addition to the simulation settings above, we also explore the scenarios when the signals of the joint and individual structures are distinct. We set the true singular values of the natural parameter matrix of the joint structure relatively small (1=2, 1=5 or 1=10 of the singular values in the original setting in different scenarios). The results are shown in Table S3 in the supplementary materials. The results show that the performance of missing imputation for **Gaussian-Gaussian** and **Gaussian-Poisson** scenarios is relatively robust against the change of singular values. For scenarios involving binomial distributions, the performance is sensitive to the change of signal.

In order to evaluate how sensitive the algorithm is to initial values, we use different initial values and evaluate the estimation performance. Data are generated in the same way in Section 4.1. For each scenario, we fix the simulated data and generate different initial values based on different random seeds. Table S4 in the supplementary materials shows that the performance of missing imputation by the proposed method is stable, which indicates that our algorithm is not sensitive to different initial values.

4.5 Real Data Analysis

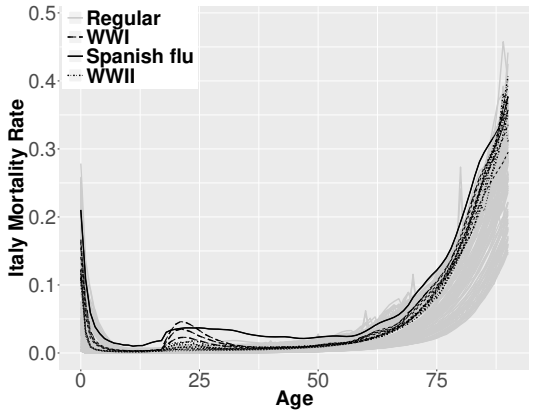
In this section, we apply our proposed method to a mortality study, where the data are publicly available from Human Mortality Database HMD (2011). We focus on exposure-to-risk and population size data sets in two countries, Italy and Switzerland, and analyze the commonality and specificity of the mortality rate patterns in both countries. The chosen data set, exposure-to-risk data set contains realizations of binomial random variables with the number of trials equal to the corresponding entries in population size data set. The Italian data have 143 rows where each row represents a year between 1872 2014; the Switzerland data have 139 rows where each row represents a year between 1876 2014. Since the number of exposure-to-risk becomes quite small at older ages, we only focus on the data at age of 0 90. Therefore, there are 91 columns each for Italy and Switzerland where each column represents an age group. The mortality data are not available for Switzerland in 1872 1875. We use our proposed method to impute the missing mortality rates.

Figure 4.2 illustrates the mortality rates across age groups in different years in each country. The mortality rates are calculated by taking the ratio of the number of exposure-to-risk and corresponding population size. Figure 4.2a and Figure 4.2b are the curve plots which show the mortality rate as a function of age and each curve represents a year. They show that the mortality rate is relatively high at an early age and decreases dramatically after birth time. The death rate remains stable after birth time to age 60 and gradually increases after that time. For Italy, several curves (dashed line and dotted line in Figure 4.2a) have a surge around 20 years old. Those curves are mortality rate curves in the year from 1915 to 1918 and from 1942 to 1944, when World War I (WWI) and World War II (WWII) happened. The two world wars led to a mass death of young adults in these years. A curve (black solid) in Figure 4.2a and Figure 4.2b stands out against the other curves across all age groups. This black solid curve is the mortality rate curve in 1918 when Spanish flu pandemic happened and led to a mass death for people of all age groups. Figure

Figure 4.2: Spaghetti plots and heat maps for the mortality rate over age for Italy and Switzerland. Black solid line represents the Spanish flu pandemic. Dashed lines represent the World War I. Dotted lines represent the World War II. Grey solid lines represent regular years.

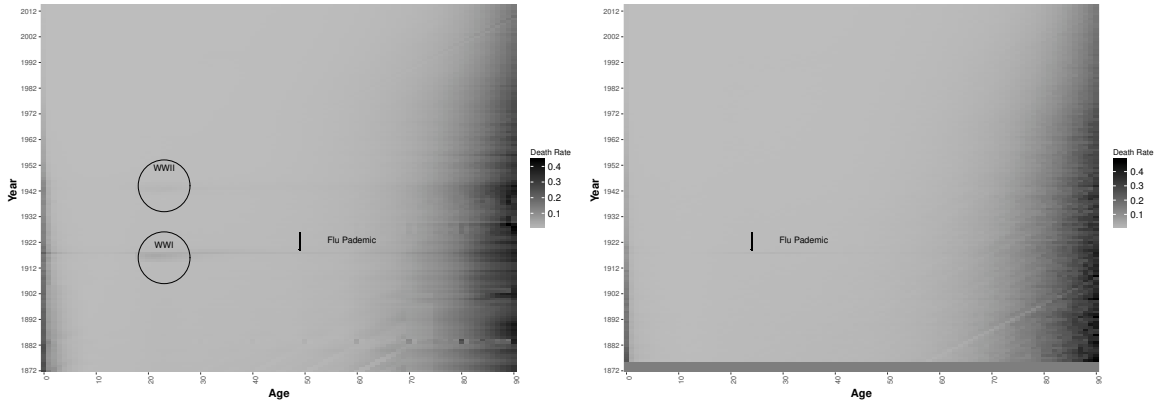
(a) Mortality rate over age for Italy

(b) Mortality rate over age for Switzerland



(c) True Mortality Rate for Italy

(d) True Mortality Rate for Switzerland



4.2c and Figure 4.2d are the heat maps for the true mortality rate of Italy and Switzerland. In the heat map for Italy, the two outlying periods are shown by two horizontal strips in Figure 4.2c. The first strip around age 20 is the period during the time of WWI. Within this period, there is an outlying line across all the age groups, which is the time of Spanish flu pandemic. The second strip around 20 years old is the period during the time of WWII. There is only one thin horizontal line in Figure 4.2d, which is the period during the flu pandemic.

We apply GIPCA to the mortality in both countries. First, we use BIC to estimate the ranks of the underlying structures. By using the stepwise BIC algorithm, we reach to a

rank estimation such that $r_J = 13; r_1 = 7; r_2 = 0$. We check the trajectory of stepwise BIC values. By comparing the BIC values of $r_J = 1; r_1 = 1; r_2 = 0$ and $r_J = 13; r_1 = 7; r_2 = 0$, we figure out that the improvement in BIC for the more complex decomposition is negligible. Thus, we choose $r_J = 1; r_1 = 1; r_2 = 0$, which leads to a simple and intuitive decomposition of the raw data.

The data types for both data sets are binomial. The link functions for both data sets are logit function. Following Algorithm 1 with rank $r_J = 1; r_1 = 1; r_2 = 0$, we get the estimates for the means, joint score and loading matrices, and individual score and loading matrices for Italy and Switzerland. Figure 4.3 visualizes the estimation results. Figure 4.3a and Figure 4.3b are the estimations of the column means for Italy and Switzerland correspondingly. The two figures demonstrate the overall age-dependent component of mortality rate. It decreases for early age groups and after a certain age, it increases exponentially. The pattern agrees with the Gompertz–Makeham law of mortality, which states that the mortality rate consists of an age-independent component and age-dependent component, increasing with age exponentially Gompertz (1825). Figure 4.3c illustrates the estimated left singular vector (score matrix) for joint structure (i.e., the shared time-varying pattern of mortality rates in different countries). The score vector has a clear dip around year 1918, which is the period of Spanish flu pandemic.

Figure 4.3d and Figure 4.3e are the joint loading vector for the two countries. The estimated loading vectors demonstrate that Spanish flu pandemic resulted in a mass death to younger people such as infants and teenagers. The estimated individual score vector for Italy is shown in Figure 4.3f. It has two apparent dips around 1917 and 1943, which correspond to the periods of World War I and World War II. The individual loading vector for Italy is shown in Figure 4.3g. It shows that the population aged around 20 to 25 was mostly affected, probably because they were directly involved in the wars. Switzerland remained neutral during both wars and therefore does not express this mortality pattern.

Next, we evaluate the imputation performance of the proposed method. In particular, we consider three ad hoc methods for imputing missing mortality rates that are commonly

used in practice: mean, adjacent years, and same year imputation. More specifically,

Ad Hoc 1 (Mean Imputation): The missing entries are imputed with the mean of mortality rate at the same age within the same data set.

Ad Hoc 2 (Adjacent Years Imputation): The missing entries are imputed with the average of mortality rate within minus/plus 5 years of the same age group within the same data set.

Ad Hoc 3 (Same Year Imputation): The missing entries are imputed with the mortality rate in the same year of the other data set.

We randomly pick 10 rows (i.e., years) for each data set and set them to be missing. Our proposed method and 3 ad hoc methods are applied to the block-wise missing data. We repeat the procedure 100 times. The estimated mortality rates are the inverse logit of the estimated natural parameter matrix. The block-wise missing entries are imputed by the inverse logit of the corresponding estimated joint structure. We calculate $DiffR_{miss}$ (Model (4.6)) to the imputed and true data and running time for GIPCA and other approaches.

Imputation performance results in Table 4.2 show that GIPCA outperforms the other 3 ad hoc approaches in terms of the relative Frobenius loss. Among the 3 ad hoc methods, the imputation accuracy for **Ad Hoc 3** is the closest to what we have for GIPCA. The better performance validates the assumption we make for **Ad Hoc 3** that the mortality rates are similar between Italy and Switzerland in the same year. On the other hand, it also agrees with the imputation mechanism implemented by GIPCA that we use the joint association to impute the missing entries. **Ad Hoc 2** performs the worst among all the methods. The unsatisfactory results of **Ad Hoc 1** and **Ad Hoc 2** indicate that simply using average across different years within one data set to impute the missing mortality is limited. When we have two or more data sets, which share the same samples, imputing missing entries by taking the advantage of the shared traits among different data sets is better than using the average within one data set.

4.6 Discussion

In this paper, we develop a generalized integrative principal components analysis approach for dimension reduction of data sets from multiple sources with different data types. Our proposed method is also able to deal with multi-source data sets containing block-wise missing entries. We apply the proposed method to mortality data in Italy and Switzerland and identify some meaningful signals, and achieve good missing data imputation accuracy. We also develop a rank selection approach derived from BIC, which accommodates multi-source data of different distributional types.

Based on the result in Section 4.4.2, the stepwise BIC approach performs well in most scenarios. However, when we have data from more than two sources, the accuracy tends to lower. Alternative rank selection methods call for more investigation.

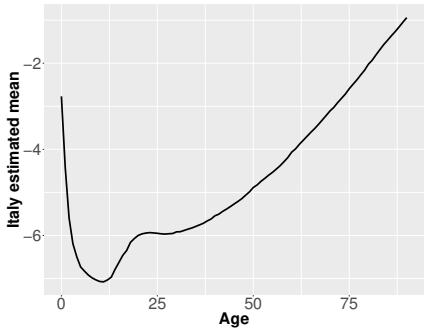
As to the proposed algorithm, although the current GIPCA algorithm only applies to the exponential family distributions, the general idea can be extended to more general non-Gaussian distributions. Extensions to other distributions are future research directions.

Table 4.1: Simulation results for two data sets based on 100 simulation runs when the natural parameter matrices are fixed for each data source. 5%M, 15%M represent the 5% and 15% missing rate correspondingly. The median and the median absolute deviation (MAD) for the relative Frobenius loss under each scenario are calculated. MAD is in parenthesis. The best results are highlighted in bold.

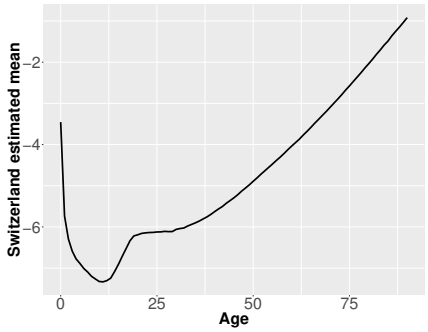
		Adhoc1		**Adhoc2**		**GIPCA**		
		Source1	Source2	Source1	Source2	Source1	Source2
Scenario 1 (5%M)	$Diff R_{Miss}$	8.46 (2.21)	8.78 (2.07)	1.13(0.15)	1.00(0.00)	**0.69 (0.01)**	**0.69 (0.00)**
Gaussian Gaussian	**Running Time**	98.12 (12.06)		3.25 (0.04)		96.66 (21.79)	
Scenario 1 (15%M)	$Diff R_{Miss}$	7.44 (0.45)	7.64 (0.46)	1.36 (0.00)	1.01 (0.00)	**0.65 (0.00)**	**0.72 (0.00)**
Gaussian Gaussian	**Running Time**	98.12 (12.06)		3.25 (0.04)		96.66 (21.79)	
Scenario 2 (5%M)	$Diff R_{Miss}$	8.13 (5.20)	2.87 (1.00)	0.49 (0.01)	1.31 (0.88)	**0.45 (0.00)**	**0.28 (0.01)**
Gaussian Poisson	**Running Time**	238.69 (53.45)		2.23 (0.05)		209.77 (65.14)	
Scenario 2 (15%M)	$Diff R_{Miss}$	9.40 (3.01)	3.64 (0.29)	0.72 (0.08)	0.58 (0.03)	**0.46 (0.00)**	**0.47 (0.00)**
Gaussian Poisson	**Running Time**	245.42 (53.78)		1.84 (0.05)		221 (54.84)	
Scenario 3 (5%M)	$Diff R_{Miss}$	1.11 (0.40)	1.02 (0.32)	0.84 (0.00)	0.99 (0.00)	**0.77 (0.00)**	**0.43 (0.00)**
Gaussian binomial	**Running Time**	464.97 (93.89)		1.82(0.11)		193.94 (110.07)	
Scenario 3 (15%M)	$Diff R_{Miss}$	1.11 (0.40)	1.02 (0.32)	0.84 (0.00)	0.99 (0.00)	**0.77 (0.00)**	**0.43 (0.00)**
Gaussian binomial	**Running Time**	473.5 (99.23)		1.83 (0.11)		189.65 (106.73)	
Scenario 5 (5%M)	$Diff R_{Miss}$	6.1 (2.07)	4.84 (0.88)	0.58 (0.03)	4.58 (1.10)	**0.57 (0.01)**	**0.84 (0.00)**
Poisson binomial	**Running Time**	188.99 (27.49)		0.38 (0.03)		914.98 (177.24)	
Scenario 5 (15%M)	$Diff R_{Miss}$	5.66 (1.66)	5.35 (1.36)	0.63 (0.02)	1.47 (0.21)	**0.61 (0.01)**	**0.86 (0.00)**
Poisson binomial	**Running Time**	191.19 (33.53)		0.30 (0.01)		808.15 (210.31)	

Figure 4.3: Estimated result

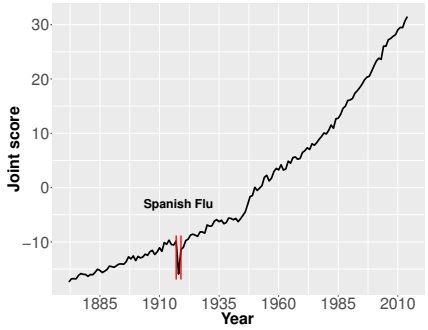
(a) Estimated means for Italy



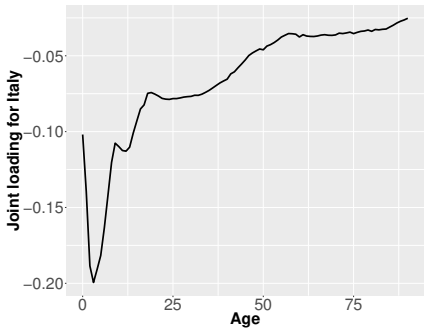
(b) Estimated means for Switzerland



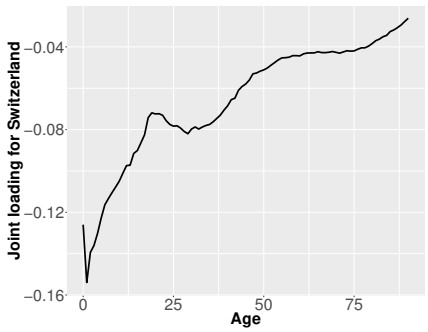
(c) Estimated joint score matrix



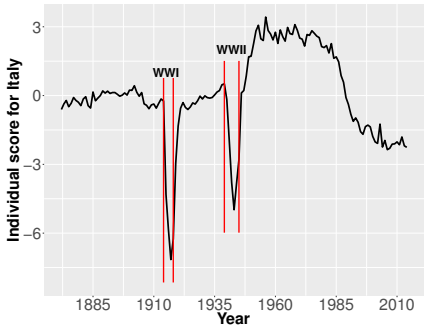
(d) Estimated joint loading matrix for Italy



(e) Estimated joint loading matrix for Switzerland



(f) Estimated individual score matrix for Italy



(g) Estimated individual loading matrix for Italy

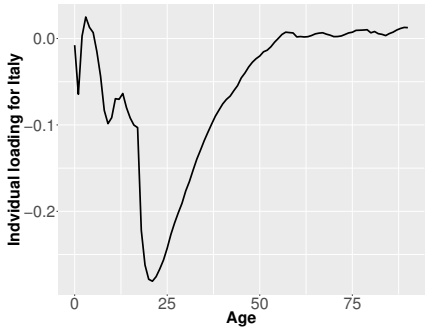


Table 4.2: We randomly pick 10 rows for each data set and set them to be missing. Impute the block-wise missing for Italy and Switzerland using GIPCA and 3 ad hoc approaches. The above procedure is repeated 100 times. The median and the median absolute deviation (MAD, in the parenthesis) for the relative Frobenius loss mentioned in section 4.2 are calculated. The best results are highlighted in bold.

Method	Italy	Switzerland
GIPCA	**0.137 (0.037)**	**0.084 (0.009)**
Ad Hoc 1 Mean Imputation	0.314 (0.056)	0.319 (0.046)
Ad Hoc 2 Adjacent Year Imputation	0.468 (0.140)	0.490 (0.127)
Ad Hoc 3 Same Year Imputation	0.163 (0.035)	0.164 (0.024)

Bibliography

Abbas, C. C., Schmid, J.-P., Guler, E., Wiedemar, L., Begré, S., Saner, H., Schnyder, U., and Von Känel, R. (2009). Trajectory of posttraumatic stress disorder caused by myocardial infarction: a two-year follow-up study. *The International Journal of Psychiatry in Medicine*, 39(4):359–376.

Bhat, H. S., Kumar, N., and Vaz, G. J. (2015). Towards scalable quantile regression trees. *In Big Data (Big Data), 2015 IEEE International Conference*, pages 53–60.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434.

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

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1:58.

Breiman, L. (2004). Consistency for a simple model of random forests.

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

Brewin, C. R., Andrews, B., and Valentine, J. D. (2000). Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults.

Cai, T., Cai, T. T., and Zhang, A. (2016). Structured matrix completion with applications to

genomic data integration. *Journal of the American Statistical Association*, 111(514):621–633.

Chakraborty, B. and Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92(2):399–418.

Chaudhuri, P. and Loh, W. Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, pages 561–576.

Chen, J., Yu, K., Hsing, A., and Therneau, T. M. (2007). A partially linear tree-based regression model for assessing complex joint gene–gene and gene–environment effects. *Genetic Epidemiology: The O cial Publication of the International Genetic Epidemiology Society*, 31(3):238–251.

Collins, M., Dasgupta, S., and Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624.

de Boor, C. (1978). A practical guide to splines (applied mathematical sciences vol 27).

Doubleday, K., Zhou, H., Fu, H., and Zhou, J. (2018). An algorithm for generating individualized treatment decision trees and random forests. *Journal of Computational and Graphical Statistics*, 27(4):849–860.

Edmondson, D. (2014). An enduring somatic threat model of posttraumatic stress disorder due to acute life-threatening medical events. *Social and personality psychology compass*, 8(3):118–134.

Edmondson, D., Richardson, S., Falzon, L., Davidson, K. W., Mills, M. A., and Neria, Y.

(2012). Posttraumatic stress disorder prevalence and risk of recurrence in acute coronary syndrome patients: a meta-analytic review. *PloS one*, 7(6):e38915.

Edmondson, D., Shimbo, D., Ye, S., Wyer, P., and Davidson, K. W. (2013). The association of emergency department crowding during treatment for acute coronary syndrome with subsequent posttraumatic stress disorder symptoms. *JAMA internal medicine*, 173(6):472–475.

Falconer, D. S. (1952). The problem of environment and selection. *The American Naturalist*, 86(830):293–298.

Fan, J., Liu, H., Wang, W., and Zhu, Z. (2016). Heterogeneity adjustment with applications to graphical model inference. *arXiv preprint arXiv:1602.05455*.

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

Galimberti, G., Pillati, M., and Soffritti, G. (2011). Notes on the robustness of regression trees against skewed and contaminated errors. *New Perspectives in Statistical Modeling and Data Analysis, Studies in Classi cation, Data Analysis, and Knowledge Organization*, pages 255–263.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583.

Guo, S.-W. (2000). Gene-environment interaction and the mapping of complex traits: some statistical models and their implications. *Human heredity*, 50(5):286–303.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journaltitle of Nonparametric Statistics*, 2(4):307–331.

Hahn, L. W., Ritchie, M. D., and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, 19(3):376–382.

Hales, C. M., Carroll, M. D., Fryar, C. D., and Ogden, C. L. (2017). Prevalence of obesity among adults and youth: United states, 2015–2016.

HMD (2011). Human mortality database. http://www.mortality.org.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. et al. (2010). Rank tests for heterogeneous treatment effects with covariates. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckova*, pages 134–142. Institute of Mathematical Statistics.

Kraft, P., Yen, Y.-C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human heredity*, 63(2):111–119.

Laber, E. and Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514.

Landgraf, A. J. and Lee, Y. (2015). Dimensionality reduction for binary data through the projection of natural parameters. *arXiv preprint arXiv:1510.06112*.

Lavori, P. W. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1(1):9–20.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733.

Li, G. and Gaynanova, I. (2017). A general framework for association analysis of heterogeneous data. *Annals of Applied Statistics*. to appear.

Li, G. and Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73(4):1433–1442.

Li, G., Li, Y., and Tsai, C.-L. (2015). Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association*, 110(509):246–261.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197.

Löfstedt, T. and Trygg, J. (2011). Onplsa novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics*, 25(8):441–455.

Loh, W.-Y. (2012). Variable selection for classification and regression in large p, small n problems. In *Probability approximations and beyond*, pages 135–159. Springer.

Loh, W. Y. and Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica sinica*, pages 815–840.

Loh, W. Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83(403):715–725.

Ma, S., Yang, L., Romero, R., and Cui, Y. (2011). Varying coefficient model for gene–environment interaction: a non-linear look. *Bioinformatics*, 27(15):2119–2126.

Maity, A., Carroll, R. J., Mammen, E., and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene–environment interactions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):75–96.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research.*, 7:983–999.

Meli, L., Alcántara, C., Sumner, J. A., Swan, B., Chang, B. P., and Edmondson, D. (2017). Enduring somatic threat perceptions and post-traumatic stress disorder symptoms in survivors of cardiac events. *Journal of Health Psychology*, page 1359105317705982.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434.

Ozer, E. J., Best, S. R., Lipsey, T. L., and Weiss, D. S. (2003). Predictors of posttraumatic stress disorder and symptoms in adults: a meta-analysis. *Psychological bulletin*, 129(1):52.

OConnell, M. J. and Lock, E. F. (2016). R. jive for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879.

Ray, P., Zheng, L., Lucas, J., and Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–1376.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M., et al. (2004). Sequenced treatment alternatives to relieve depression (star* d): rationale and design. *Controlled clinical trials*, 25(1):119–142.

Schouteden, M., Van Deun, K., Wilderjans, T. F., and Van Mechelen, I. (2014). Performing disco-sca to search for distinctive and common information in linked data. *Behavior research methods*, 46(2):576–587.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

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

Tseng, G. C., Ghosh, D., and Zhou, X. J. (2015). *Integrating Omics Data*. Cambridge University Press.

von Känel, R., Hari, R., Schmid, J.-P., Saner, H., and Begré, S. (2011). Distress related to myocardial infarction and cardiovascular outcome: a retrospective observational study. *BMC psychiatry*, 11(1):98.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association*, 104(487):1129–1143.

Whitehead, D. L., Strike, P., Perkins-Porras, L., and Steptoe, A. (2005). Frequency of distress and fear of dying during acute coronary syndromes and consequences for adaptation. *The American journal of cardiology*, 96(11):1512–1516.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J., Initiative, A. D. N., et al. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206.

Yang, Z. and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., Initiative, A. D. N., et al. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*, (just-accepted):1–20.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2016a). Group component analysis

for multiblock data: Common and individual feature extraction. *IEEE Trans Neural Netw Learn Syst*, 27(11):2426–2439. Advance online publication.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2016b). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE transactions on neural networks and learning systems*, 27(11):2426–2439.

Appendix 1

Appendices for Chapter 2

Before that, we recall some key notations which are important to establish the theoretical results. In the theoretical result, $T_b; b = 1; 2 : : : ; B$ refers to a single tree in the random forest T, where B is the total number of trees in the forest. Let $N_b(\mathbf{x}_0)$ be the leaf space where the observation $\mathbf{x}_0$ locates in the tree T_b in the random forest T. The estimated quantile specific coefficient is obtained by Equation (2.12), $\widehat{\ }_{\mathsf{T}}(\ ; \mathbf{x}_0) = \arg\min \sum_{i=1}^{n} \quad (y_i \quad \mathbf{z}_i^{>} \quad) \, !_i(\mathbf{x}_0);$ where the weight function is $!_i(\mathbf{x}_0)$ is calculated in equation (2.10). It is equivalent that $\widehat{\ }_{\mathsf{T}}(\ ; \mathbf{x}_0)$ is the solution to

$$S_{n;\mathsf{T}}(\ ; \ ; \mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \quad (y_i \quad \mathbf{z}_i^{>} \quad) \mathbf{z}_i^{>} \, !_i(\mathbf{x}_0; b) = 0; \quad 2 \ (0; 1) \tag{1.1}$$

where $\quad (u)$ is the first derivative function of the quantile loss $\quad (u)$, which is $\quad (u) = 1fu < 0g.$

The expected estimation equation is,

$$\begin{aligned} S_{\mathsf{T}}(\ ; \ ; \mathbf{x}_0) &= \frac{1}{B} \sum_{b=1}^{B} \mathsf{E}_{yj\mathbf{x};\mathbf{z};} \ [\quad (y \quad \mathbf{z}^{>} \quad) \mathbf{z}^{>} \, j \, \mathbf{x} \, 2 \, N_b(\mathbf{x}_0)] \\ &= \frac{1}{B} \sum_{b=1}^{B} \frac{\int_{\mathsf{X}(\mathbf{x}_0;\mathbf{z})} \mathsf{E}_{yj\mathbf{x};\mathbf{z};} \ [\quad (y \quad \mathbf{z}^{>} \quad)] \mathbf{z}^{>} dF_{(\mathbf{x};\mathbf{z})}(\mathbf{x}; \mathbf{z})}{P(\mathbf{x} \, 2 \, N_b(\mathbf{x}_0))}; \quad 2 \ (0; 1) \end{aligned}$$

where $F_{(\mathbf{x};\mathbf{z})}(\mathbf{x};\mathbf{z})$ is the joint distribution of covariates $\mathbf{x}$ and controlling variables $\mathbf{z}$ and $\mathsf{X}(\mathbf{x}_0;\mathbf{z}) = f\mathbf{x};\mathbf{z} : \mathbf{x}\ 2\ N_b(\mathbf{x}_0)g$. The true quantile coefficient $(\ ;\mathbf{x}_0)$ is the solution to

$$S(\ ;\ ;\mathbf{x}_0) \widehat{=} \mathsf{E}_{y;\mathbf{z}j\mathbf{x}_0;}\ [\ \ (y\ \ \mathbf{z}^{>}\ \ )\mathbf{z}] = \mathbf{0};\quad 2\ (0;1): \tag{1.2}$$

In Condition 5, we mention that $(\ ;\mathbf{x}_0)$ is a smooth function regarding to $\ $ on $(0;1)$. In order to prove the consistency of the estimation $\widehat{\ }_{\mathsf{T}}(\ ;\mathbf{x}_0)$ on $\ 2\ (0;1)$ in Theorem 1, we first prove its consistency on a countable grid of quantile on $(0;1)$. Followed by Wei and Carroll (2009), we approximate the quantile coefficient $(\ ;\mathbf{x}_0)$ by natural linear splines with t_n common internal knots that $\Pi = f" = \ _1 < \ _2 < \quad < \ _{t_n} = 1\ \ "g$. Let $\ (\mathbf{x}_0) = (\ (\ _1;\mathbf{x}_0)^{>};\ (\ _2;\mathbf{x}_0)^{>};::::;\ (\ _{t_n};\mathbf{x}_0)^{>})^{>}$ be the set of quantile coefficients at quantile levels Π. The linear approximation of $(\ ;\mathbf{x}_0)$ is,

$$= _{(\mathbf{x}_0)}(\ ) = \begin{cases} (\ _1;\mathbf{x}_0) & < \ _1 \\ (\ _{t_n}\mathbf{x}_0) & > \ _{t_n} \\ (\ _{b\ t_n=(t_n+1)c};\mathbf{x}_0) + \frac{\ _{\mathsf{T}}(\ _{b\ t_n=(t_n+1)c+1};\mathbf{x}_0)\ \ -\ _{\mathsf{T}}(\ _{b\ t_n=(t_n+1)c};\mathbf{x}_0)}{1=t_n}(\quad \frac{b\ t_n c}{t_n+1}) & \text{else} \end{cases}$$

We know that if we have sufficient numbers of knots, that $t_n\ \ !\ \ 1$ when $n\ \ !\ \ 1$ and $"\ \ !\ \ 0$, the difference between $(\ ;\mathbf{x}_0)$ and its spline approximation $= _{(\mathbf{x}_0)}(\ )$ is negligible (de Boor, 1978). Such that, instead of estimating a smoonth function $(\ ;\mathbf{x}_0)$ by solving the estimation equation (1.1) on $\ 2\ (0;1)$, the estimation equation could be reduced to a finite dimension equation by t_n grid internal knots,

$$S_{n;\mathsf{T}}(\ ;\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\Psi_{y_i}(\ )\quad \mathbf{z}_i\ !_i(\mathbf{x}_0;b); \tag{1.3}$$

where $\ = (\ _1^{>};\ _2^{>};::::;\ _{t_n}^{>})^{>}$ and $\Psi_{y_i}(\ ) = (\ _1(y_i\ \ \mathbf{z}_i^{>}\ \ _{t_1});\ _2(y_i\ \ \mathbf{z}_i^{>}\ \ _{t_2});::::;\ _{t_n}(y_i\ \ \mathbf{z}_i^{>}\ \ _{t_n}))^{>}$, which is a $t_n\ \ 1$ vector and contains t_n vectors of $\ _t(y_i\ \ \mathbf{z}_i^{>}\ \ _t)\mathbf{z}_i; t = 1;2:::;t_n;$ and $\Psi_{y_i}(\ )\quad \mathbf{z}_i$ is a $t_n(q+1)$ dimensional vector and $\ $ stands for Kronerker product. The

estimated coefficient

$$\hat{}_{\mathsf{T}}(\mathbf{x}_0) = (\hat{}_{\mathsf{T}}(\ _1;\mathbf{x}_0)^{>};\ \hat{}_{\mathsf{T}}(\ _2;\mathbf{x}_0)^{>};:::;\ \hat{}_{\mathsf{T}}(\ _{t_n};\mathbf{x}_0)^{>})^{>} \tag{1.4}$$

is a $t_n(q+1)$ dimensional vector which is the solution to $S_{n;\mathsf{T}}(\ ;\mathbf{x}_0) = \mathbf{0}$. The expected estimation equation is,

$$S_{\mathsf{T}}(\ ;\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B} \frac{\int_{\mathsf{X}(\mathbf{x}_0;\mathbf{z})} \mathsf{E}_{y|\mathbf{x};\mathbf{z};\ } f\Psi_y(\ )g \quad \mathbf{z}dF_{(\mathbf{x};\mathbf{z})}(\mathbf{x};\mathbf{z})}{P(\mathbf{x}\ 2\ N_b(\mathbf{x}_0))}:$$

The score function (1.2) can be written as,

$$S(\ ;\mathbf{x}_0) = E_{y;\mathbf{z}|\mathbf{x}_0;\ } f\Psi_y(\ )g \quad \mathbf{z}: \tag{1.5}$$

It is obvious that $(\mathbf{x}_0)$ is the solution to $S(\ ;\mathbf{x}_0) = \mathbf{0}$.

Condition 6. *The true coe cient* $(\ ;\mathbf{x}_0)$ *is the unique solution to* $S(\ ;\ ;\mathbf{x}_0) = \mathbf{0}$ *for all* $2\ (0;1)$.

Lemma 1. $!(\mathsf{T};\ (\ ;\ ))$ *is the oscillation of the function* $(\ ;\ )$ *respected to the random forest* T. *For a xed* $\mathbf{x}_0$,

$$!(\mathsf{T};\ (\ ;\ )) = \max_b \sup_{\mathbf{x}_1;\mathbf{x}_2 2 N_b(\mathbf{x}_0)} \sup_{2[1=(t_n+1);t_n=(t_n+1)]} jj\ (\ ;\mathbf{x}_1) \quad (\ ;\mathbf{x}_2)\ jj= o(1);$$

where t_n *is a sequence of positive integers and satis es that* t_n " 1 *and* $t_n\ j\ N_{\mathsf{T}}(\mathbf{x}_0)\ j^{\ 1}\ !\ 0$..

Lemma 2. *Under some regularity conditions, for a xed* $\mathbf{x}_0$, *we have*

$$t_n{}^{1} \sup_{2\mathbf{\Theta}} jj\ S_{n;\mathsf{T}}(\ ;\mathbf{x}_0) \quad S_{\mathsf{T}}(\ ;\mathbf{x}_0)\ jj= o_p(1);\ n\ !\ 1 \tag{1.6}$$

where $\mathbf{\Theta}\ \Pi$ *is a compact set, which is the support of* . *Without loss of generality, we can write is as* $\mathbf{\Theta}\ \Pi = [_t f\ :j \quad (\ _t;\mathbf{x}_0)\ j< 1; t = t_1; t_2;:::; t_n g.$

Lemma 3. *Under Conditions 1-5, for a xed* $\mathbf{x}_0$*, we have*

$$\sup_{\in\Theta} \quad t_n^{\ 1} \, jj \, S_T(\ ;\mathbf{x}_0) \quad S(\ ;\mathbf{x}_0) \, jj = o(1) \tag{1.7}$$

as $n \ ! \ 1$ *.*

1.1 Proof of Lemma 1

Proof. According to Condition 3, we know that for a terminal leaf in a random tree, every covariate in $\mathbf{x}$ is split infinite times. It is certain that

$$\sup_{\mathbf{x}_1;\mathbf{x}_2 2 N_b(\mathbf{x}_0)} jj \, \mathbf{x}_1 \quad \mathbf{x}_2 \, jj = o(1):$$

According to Condition 4, we have

$$\sup_{2[1=(t_n+1);t_n=(t_n+1)]} \sup_{\mathbf{x}_1;\mathbf{x}_2 2 N_b(\mathbf{x}_0)} jj \quad (\ ;\mathbf{x}_1) \qquad (\ ;\mathbf{x}_2) \, jj$$

$$\sup_{2[1=(t_n+1);t_n=(t_n+1)]} M(\ ) \sup_{\mathbf{x}_1;\mathbf{x}_2 2 N_b(\mathbf{x}_0)} jj \, \mathbf{x}_1 \quad \mathbf{x}_2 \, jj = o(1):$$

. □

1.2 Proof of Lemma 2

Proof. Here, for a vector $\mathbf{v}$, we use $jj \, \mathbf{v} \, jj$ to denote its Euclidean norm and $j \, \mathbf{v} \, j$ to denote its component-wise absolute values. By $j \, \mathbf{v} \, j < 1$, we mean that each component of $\mathbf{v}$ is bounded by 1.

To prove Lemma 2, we need to show that

$$Pr(t_n^{\ 1} \sup_{2\Theta} \quad jj \, S_{n;T}(\ ;\mathbf{x}_0) \quad S_T(\ ;\mathbf{x}_0) \, jj > ") \ ! \ 0; \tag{1.8}$$

as $n \to \infty$. We will show (1.8) using Huber's chaining augment (Bickel, 1975). We partition each subspace of the parameter space $\boldsymbol{\Theta} \times \Pi$ at $\theta_t = t\epsilon/(t_n+1)$ into L_n disjoint small cubes Γ_l with diameters less than $q_n = C_1 t_n \epsilon / \min_b \| N_b(\mathbf{x}_0) \|$ for some constant C_1. Let θ_l be the center of the lth cube Γ_l The probability of the left hand side of (1.8) is bounded by the sum of the two probabilities below,

$$P_1 = Pr(t_n^{-1} \max_{l \le L_n} \sup_{\theta \in \Gamma_l} \| S_{n,T}(\theta, \pi; \mathbf{x}_0) - S_{n,T}(\theta_l, \pi; \mathbf{x}_0) + S_T(\theta, \pi; \mathbf{x}_0) - S_T(\theta_l, \pi; \mathbf{x}_0) \| > \epsilon/2);$$

$$P_2 = Pr(t_n^{-1} \max_{l \le L_n} \| S_{n,T}(\theta_l, \pi; \mathbf{x}_0) - S_T(\theta_l, \pi; \mathbf{x}_0) \| > \epsilon/2).$$

Let's first look at,

$$\begin{aligned}
&\max_l \sup_{\theta \in \Gamma_l} \| S_{n,T}(\theta, \pi; \mathbf{x}_0) - S_{n,T}(\theta_l, \pi; \mathbf{x}_0) \| \\
&= \max_l \sup_{\theta \in \Gamma_l} \left\| \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} w_i(\mathbf{x}_0; b) [\Psi_{y_i}(\theta, \pi; \mathbf{x}_0) - \Psi_{y_i}(\theta_l, \pi; \mathbf{x}_0)] \mathbf{z} \right\| \\
&\quad \max_l \sup_{\theta \in \Gamma_l} \left\| \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} w_i(\mathbf{x}_0; b) I\{ \mathbf{x}_i^\top \theta_l \ge y_i \} \| \mathbf{x}_i^\top (\theta - \theta_l) \| g \mathbf{z} \right\| \\
&\quad t_n \max_l \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} w_i(\mathbf{x}_0; b) \| I\{ \mathbf{x}_i^\top \theta_l \ge y_i \} - q_n \| \\
&= t_n o_p(1);
\end{aligned}$$

by the fact that $Pr(\| \mathbf{x}_i^\top \theta_l - y_i \| \le q_n) = 0$ as $n \to \infty$ and $\sum_{b=1}^{B} \sum_{i=1}^{n} w_i(\mathbf{x}_0; b) = 1$.

A sufficient condition to show $\sup_{\theta \in \Gamma_l} \| S_T(\theta, \pi; \mathbf{x}_0) - S_T(\theta_l, \pi; \mathbf{x}_0) \| = o(1)$ is

$$\max_i \| f(y_i \| \mathbf{x}_0; \theta, \pi) - f(y_i \| \mathbf{x}; \theta_l, \pi) \| = o_p(1)$$

which can be proved by following the proof of (A.9) in Wei and Carroll (2009). It then follows that $P_1 = o(1)$.

Let $\Upsilon_i(l; t; j; b) = \omega_i(\mathbf{x}_0; b) \tau_k(y_i - \mathbf{z}_i^\top \theta_t) z_{ij} 1\{ \theta \in \Gamma_l \}$. Then, a sufficient condition for

$P_2 = o(1)$ is that, δ $_t$ and $\mathbf{z}_{;j}$,

$$Pr\{ \max_{1 \; l \; L_n;1 \; t \; t_n;1 \; j \; q} \frac{1}{B} \sum_{b=1}^{B} \left| \sum_{i=1}^{n} f Y_i(l;t;j;b) \quad E Y_i(l;t;j;b) g \right| > " \}:$$

By Bernstein's inequality, we have,

$$\begin{aligned} & Pr\left\{ \max_{1 \; l \; L_n;1 \; t \; t_n;1 \; j \; q} \frac{1}{B} \sum_{b=1}^{B} \left| \sum_{i=1}^{n} \{ Y_i(l;t;j;b) \quad E Y_i(l;t;j;b) \} \right| > " \right\} \\ & t_n \; q \; L_n \; b \; Pr\left\{ \left| \sum_{i=1}^{n} \{ Y_i(l;t;j;b) \quad E Y_i(l;t;j;b) \} \right| > " \right\} \\ & t_n \; q \; L_n \; b \; \exp\left\{ \frac{\min_b j \, N_b(\mathbf{x}_0) \, j \, "}{2(E\{ \quad _k(y_i \quad \mathbf{z}_i^{>} \quad _t) z_{ij} 1 f \quad 2 \, \Gamma_l g\}^2 + 1=3")} \right\} = o(1): \end{aligned}$$

It then follows that $P_1 = o(1)$ and $P_2 = o(1)$, which implies that (1.8) holds

□

1.3 Proof of Lemma 3

Proof. We know that $_0(\mathbf{x}_0)$ is the solution to $S(\ ;\mathbf{x}_0) = \mathbf{0}$. For a fixed $\mathbf{x}$, we have,

$$\begin{aligned} & jj \, S_T(\ ;\mathbf{x}_0) \quad S(\ ;\mathbf{x}_0) \, jj \\ & = \left|\left| \frac{1}{B} \sum_{b=1}^{B} \frac{\int_{X(\mathbf{x}_0;\mathbf{z})} E_{yj\mathbf{x};\mathbf{z}} f \Psi_y(\ ) g \quad \mathbf{z} dF_{\mathbf{x};\mathbf{z}}(\mathbf{x};\mathbf{z})}{P(\mathbf{x} \; 2 \; N_b(\mathbf{x}_0))} \quad E_{yj\mathbf{x}_0;\mathbf{z}} f \Psi_y(\ ) g \quad \mathbf{z} \right|\right| \\ & \quad \max_b \left|\left| \frac{\int_{X(\mathbf{x}_0;\mathbf{z})} E_{yj\mathbf{x};\mathbf{z}} f \Psi_y(\ ) g \quad \mathbf{z} dF_{\mathbf{x};\mathbf{z}}(\mathbf{x};\mathbf{z})}{P(\mathbf{x} \; 2 \; N_b(\mathbf{x}_0))} \quad E_{yj\mathbf{x}_0;\mathbf{z}} f \Psi_y(\ ) g \quad \mathbf{z} \right|\right| \\ & = \max_b \left|\left| \frac{\int_{X(\mathbf{x}_0;\mathbf{z})} f E_{yj\mathbf{x};\mathbf{z}} f \Psi_y(\ ) g \quad E_{yj\mathbf{x}_0;\mathbf{z}} f \Psi_y(\ ) g g \quad \mathbf{z} dF_{\mathbf{x};\mathbf{z}}(\mathbf{x};\mathbf{z})}{P(\mathbf{x} \; 2 \; N_b(\mathbf{x}_0))} \right|\right| \\ & \quad \max_b \sup_{\mathbf{x} 2 N_b(\mathbf{x}_0)} jj \, E_{yj\mathbf{x};\mathbf{z}} f \Psi_y(\ ) g \quad E_{yj\mathbf{x}_0;\mathbf{z}} f \Psi_y(\ ) g \, jj \quad jj \, \mathbf{z} \, jj \\ & = \max_b \sup_{\mathbf{x} 2 N_b(\mathbf{x}_0)} \left|\left| \int \Psi_y(\ ) df F(yj\mathbf{x};\mathbf{z}) \quad F(yj\mathbf{x}_0;\mathbf{z}) g \right|\right| \quad jj \, \mathbf{z} \, jj : \end{aligned}$$

Thus, if we want to show that $t_n^{1}\, jj\, S_T(\ ;\mathbf{x}_0) \quad S(\ ;\mathbf{x}_0)\, jj = o(1)$, it suffices to show that,

$$\max_b \sup_{\mathbf{x}2N_b(\mathbf{x}_0)} \sup_y \sup_{\mathbf{z}} j\, F(yj\mathbf{x};\mathbf{z}) \quad F(yj\mathbf{x}_0;\mathbf{z})\, j = o(1); \tag{1.9}$$

where $F(yj\mathbf{x};\mathbf{z})$ and $F(yj\mathbf{x}_0;\mathbf{z})$ are two cumulative distribution functions whose quantile functions are $Q_y(\ ;\mathbf{x};\mathbf{z}) = \mathbf{z}^{>}\ (\ ;\mathbf{x})$ and $Q_y(\ ;\mathbf{x}_0;\mathbf{z}) = \mathbf{z}^{>}\ (\ ;\mathbf{x}_0)$. The inverse functions of quantile functions are $_{\mathbf{x};\mathbf{z}}(y) = \inf f\ : \mathbf{z}^{>}\ (\ ;\mathbf{x}) \quad yg$ and $_{\mathbf{x}_0;\mathbf{z}}(y) = \inf f\ : \mathbf{z}^{>}\ (\ ;\mathbf{x}_0)$ yg. Thus, in order to get (1.9), equivalently we need to show that,

$$\sup_y \max_b \sup_{\mathbf{x}2N_b(\mathbf{x}_0)} \sup_{\mathbf{z}} j\ _{\mathbf{x}_0;\mathbf{z}}(y) \quad _{\mathbf{x};\mathbf{z}}(y)\, j = o(1): \tag{1.10}$$

We assume that (1.10) does not hold. It means that there exist a y and an h_n which is always larger than positive constant h satisfying that $_{\mathbf{x}_0;\mathbf{z}}(y\ ) = \ _{\mathbf{x};\mathbf{z}}(y\ ) + h_n$ or $_{\mathbf{x}_0;\mathbf{z}}(y\ ) = \ _{\mathbf{x};\mathbf{z}}(y\ ) \quad h_n$ when $\mathbf{x}\ 2\ N_b(\mathbf{x}_0)$. Without loss of generality, we assume here that $_{\mathbf{x}_0;\mathbf{z}}(y\ ) = \ _{\mathbf{x};\mathbf{z}}(y\ ) + h_n$.

Let $A_n = f(\ _{\mathbf{x}_0;\mathbf{z}}(y);\ _{\mathbf{x};\mathbf{z}}(y)) : 1=(t_n + 1) \quad _{\mathbf{x};\mathbf{z}}(y) \quad _{\mathbf{x}_0;\mathbf{z}}(y) \quad t_n=(t_n + 1)g$ and $B_n = f(\ _{\mathbf{x}_0;\mathbf{z}}(y);\ _{\mathbf{x};\mathbf{z}}(y)) : 0 < \ _{\mathbf{x};\mathbf{z}}(y) < 1=(t_n + 1) \quad _{\mathbf{x}_0;\mathbf{z}}(y)g$. Condition 1 implies that,

$$\begin{aligned}
&\max_b \sup_{\mathbf{x}2N_b(\mathbf{x}_0)} \sup_{2[1=(t_n+1);t_n=(t_n+1)]} \sup_{\mathbf{z}} j\, Q_y(\ ;\mathbf{x};\mathbf{z}) \quad Q_y(\ ;\mathbf{x}_0;\mathbf{z})\, j \\
&= \max_b \sup_{\mathbf{x}2N_b(\mathbf{x}_0)} \sup_{2[1=(t_n+1);t_n=(t_n+1)]} \sup_{\mathbf{z}} j\, \mathbf{z}^{>}\ (\ ;\mathbf{x}) \quad \mathbf{z}^{>}\ (\ ;\mathbf{x}_0)\, j \\
&\quad \max_b \sup_{\mathbf{x}2N_b(\mathbf{x}_0)} \sup_{2[1=(t_n+1);t_n=(t_n+1)]} \sup_{\mathbf{z}} jj\, \mathbf{z}\, jj \quad jj\ (\ ;\mathbf{x}) \quad (\ ;\mathbf{x}_0)\, jj = o(1):
\end{aligned}$$

Thus, for any $" > 0$, there exists a N, when $n > N$,

$$\max_b \sup_{\mathbf{x}2N_b(\mathbf{x}_0)} \sup_{2[1=(t_n+1);t_n=(t_n+1)]} \sup_{\mathbf{z}} j\, Q_y(\ ;\mathbf{x}_0;\mathbf{z}) \quad Q_y(\ ;\mathbf{x};\mathbf{z})\, j \quad ": \tag{1.11}$$

If $(\ _{\mathbf{x}_0;\mathbf{z}}(y\ );\ _{\mathbf{x};\mathbf{z}}(y\ ))\ 2\ A_n\ [\ B_n$, by mean value theorem and Condition 5 that $\mathbf{z}^{>}\ ^{0}(\ ;\mathbf{x};\mathbf{z}) >$

0 for all $\in (0;1)$, there exists a $\in (\ _{x;z}(y\ );\ _{x_0;z}(y\ ))$ and a $M\ > 0$ satisfying,

$$z^{>}\ ^{0}(\ \ ;x;z) = \frac{Q_y\ (\ _{x_0;z}(y\ );x;z)\ \ Q_y\ (\ _{x;z}(y\ );x;z)}{h_n} > \frac{1}{M}: \tag{1.12}$$

Therefore, followed by (1.12) and (1.11), when $n > N$ and $\ _{x_0;z}(y\ ) \in [1=(t_n+1); t_n=(t_n+1)]$, we have,

$$\begin{aligned}
h_n &< M\ \ fQ_y\ (\ _{x_0;z}(y\ );x;z)\ \ \ Q_y\ (\ _{x;z}(y\ );x;z)g \\
&\quad M\ \ f\max_b \sup_{x\in N_b(x_0)} \sup_{\in[1=(t_n+1);t_n=(t_n+1)]} \sup_z j\ Q_y\ (\ _{x_0;z}(y\ );x;z)\ \ \ Q_y\ (\ _{x_0;z}(y\ );x_0;z)\ j \\
&+ j\ Q_y\ (\ _{x_0;z}(y\ );x_0;z)\ \ \ Q_y\ (\ _{x;z}(y\ );x;z)\ jg \\
&\quad M\ \ f''+ j\ Q_y\ (\ _{x_0;z}(y\ );x_0;z)\ \ \ Q_y\ (\ _{x;z}(y\ );x;z)\ jg = M\ \ '' = '' ;
\end{aligned}$$

where $Q_y\ (\ _{x_0;z}(y\ );x_0;z) = Q_y\ (\ _{x;z}(y\ );x;z) = y$ due to the definition of $\ _{x_0;z}(y)$ and $\ _{x;z}(y)$. Thus, for any $''$, there exist a N, when $n > N$ that $h_n < ''$. It contradicts our assumption that h_n is always larger than a positive constant h. Thus, it is sufficient to get that,

$$\max_b \sup_{x\in N_b(x_0)} \sup_{(\ _{x;z}(y);\ _{x_0;z}(y))\in A_n\ [\ B_n} \sup_z j\ \ _{x;z}(y)\ \ \ _{x_0;z}(y)\ j= o(1): \tag{1.13}$$

Similarly, we can have,

$$\max_b \sup_{x\in N_b(x_0)} \sup_{(\ _{x;z}(y);\ _{x_0;z}(y))\in C_n} \sup_z j\ \ _{x;z}(y)\ \ \ _{x_0;z}(y)\ j= o(1); \tag{1.14}$$

where $C_n = f(\ _{x_0;z}(y);\ _{x;z}(y)) :\ _{x;z}(y) < t_n=(t_n+1)\ \ \ _{x_0;z}(y) < 1g$.

Let $D_n = f(\ _{x_0;z}(y);\ _{x;z}(y)) : 0 <\ _{x;z}(y)\ \ \ _{x_0;z}(y) < 1=(t_n+1)g$. Since $t_n\ !\ 1$ as $n\ !\ 1$, it is obvious that

$$\max_b \sup_{x\in N_b(x_0)} \sup_{(\ _{x;z}(y);\ _{x_0;z}(y))\in D_n} \sup_z j\ \ _{x;z}(y)\ \ \ _{x_0;z}(y)\ j= o(1): \tag{1.15}$$

Let $E_n = \{t_n/(t_n+1) < \ _{\mathbf{x},\mathbf{z}}(y) \quad _{\mathbf{x}_0,\mathbf{z}}(y) < 1\}$, by Condition 5, it is obvious,

$$\begin{aligned}
&\max_b \sup_{\mathbf{x}\in N_b(\mathbf{x}_0)} \sup_{(\ _{\mathbf{x},\mathbf{z}}(y),\ _{\mathbf{x}_0,\mathbf{z}}(y))\in E_n} \sup_{\mathbf{z}} \left| \ _{\mathbf{x},\mathbf{z}}(y) \quad _{\mathbf{x}_0,\mathbf{z}}(y) \right| \\
&= \max_b \sup_{\mathbf{x}\in N_b(\mathbf{x}_0)} \sup_{(\ _{\mathbf{x},\mathbf{z}}(y),\ _{\mathbf{x}_0,\mathbf{z}}(y))\in A_n} \sup_{\mathbf{z}} \left| \ _{\mathbf{x},\mathbf{z}}(y) \quad _{\mathbf{x}_0,\mathbf{z}}(y) \right| = o(1). \qquad (1.16)
\end{aligned}$$

So, by (1.13), (1.14), (1.15) and (1.16),we successfully show that $\sup_y \max_b \sup_{\mathbf{x}\in N_b(\mathbf{x}_0)} \sup_{\mathbf{z}} |\ _{\mathbf{x}_0,\mathbf{z}}(y) \quad _{\mathbf{x},\mathbf{z}}(y)| = o(1)$.

□

1.4 Proof of Theorem 1

Proof. According to Lemma 2 and Lemma 3, for fixed covariates $\mathbf{x}_0$, as $n \to \infty$,

$$\begin{aligned}
&t_n^{-1} \sup_{\in\Theta} \| S_{n,T}(\ ;\mathbf{x}_0) \quad S(\ ;\mathbf{x}_0)\| \qquad (1.17)\\
&t_n^{-1} \sup_{\in\Theta} \| S_{n,T}(\ ;\mathbf{x}_0) \quad S_T(\ ;\mathbf{x}_0)\| + t_n^{-1} \sup_{\in\Theta} \| S_T(\ ;\mathbf{x}_0) \quad S(\ ;\mathbf{x}_0)\| = o_p(1).
\end{aligned}$$

For $\delta > 0$, we define a compact set $B(\mathbf{x}_0) = \{ \in \mathbb{R}^{q+1} : \| \quad (\ ;\mathbf{x}_0)\| \quad ; \in (0,1)\}$. And the complementary set of $B(\mathbf{x}_0)$ is $B^c(\mathbf{x}_0)$. We define the distance as,

$$d_n(\ ;\mathbf{x}_0) = t_n^{-1} \min_{\in\Theta\setminus B^c(\mathbf{x}_0)} \{\| S(\ ;\mathbf{x}_0)\| \quad \| S(\ _0(\mathbf{x}_0);\mathbf{x}_0)\|\}.$$

According to Condition 6, $ \ _0(\mathbf{x}_0)$ is the unique solution to $S(\ ;\mathbf{x}_0) = \mathbf{0}$. It follows that $\| S(\ _0(\mathbf{x}_0);\mathbf{x}_0)\| = 0$. Then, we have $d_n(\ ;\mathbf{x}_0) > 0$.

Define the random event that

$$E_n = \{t_n^{-1} \sup_{\in\Theta} \| S_{n,T}(\ ;\mathbf{x}_0) \quad S(\ ;\mathbf{x}_0)\| < d_n(\ ;\mathbf{x}_0)/3\}.$$

Followed by (1.17), we have that $\lim_{n \to \infty} P(E_n) = 1$. Event E_n implies,

$$t_n^{-1} \| S(\hat{\theta}_T(x_0); x_0) \| \leq t_n^{-1} \| S_{n;T}(\hat{\theta}_T(x_0); x_0) \| + d_n(\theta; x_0)/3; \tag{1.18}$$

$$t_n^{-1} \| S_{n;T}(\theta_0(x_0); x_0) \| \leq t_n^{-1} \| S(\theta_0(x_0); x_0) \| + d_n(\theta; x_0)/3. \tag{1.19}$$

We know that $\hat{\theta}_T(x_0)$ is the solution to $S_{n;T}(\theta; x_0) = \mathbf{0}$. Therefore, $\| S_{n;T}(\hat{\theta}_T(x_0); x_0) \| \leq \| S_{n;T}(\theta_0(x_0); x_0) \|$, together with (1.18) and (1.19), it follows that,

$$\begin{aligned} t_n^{-1} \| S(\hat{\theta}_T(x_0); x_0) \| &\leq t_n^{-1} \| S_{n;T}(\theta_0(x_0); x_0) \| + d_n(\theta; x_0)/3 \\ &\leq t_n^{-1} \| S(\theta_0(x_0); x_0) \| + 2d_n(\theta; x_0)/3 \end{aligned}$$

Since $\lim_{n \to \infty} P(E_n) = 1$,

$$\lim_{n \to \infty} P(t_n^{-1} \| S(\hat{\theta}_T(x_0); x_0) \| \leq t_n^{-1} \| S(\theta_0(x_0); x_0) \| + 2d_n(\theta; x_0)/3) \quad \lim_{n \to \infty} P(E_n) = 1.$$

Since $d_n(\theta; x_0) > 0$, according to the definition of the compact set $B(x_0)$, we have $\lim_{n \to \infty} P(\hat{\theta}_T(x_0) \in B(x_0)) = 1$. By construction,

$$\hat{\theta}_T(\theta; x_0) = \begin{cases} \hat{\theta}_{T;1}(x_0) & \theta < 1 \\ \hat{\theta}_{T;t_n}(x_0) & \theta > t_n \\ \hat{\theta}_{T;\lfloor t_n \rfloor = (t_n+1)c}(x_0) + \dfrac{b_{T;\lfloor t_n \rfloor = (t_n+1)c+1}(x_0) - b_{T;\lfloor t_n \rfloor = (t_n+1)c}(x_0)}{1/t_n} \left(\theta - \dfrac{\lfloor t_n c \rfloor}{t_n+1} \right) & \text{else} \end{cases}$$

Since $\theta(\theta; x_0)$ is approximated by the nature splines with value $\theta_0(x_0)$ at internal knots Π, which has the similar construction as $\hat{\theta}_T(\theta; x_0)$, we have

$$\sup_{\theta \in [1/(t_n+1); t_n/(t_n+1)]} \| \hat{\theta}_T(\theta; x_0) - \theta(\theta; x_0) \| = o_p(1)$$

. □

Appendix 2

Appendices for Chapter 4

2.1 Result of Rank Selection by Adapted BIC

We apply the rank selection procedure mentioned in Section 3.2 in the manuscript to the data generated from corresponding distribution independently for 50 times when the natural parameter matrix is fixed. BIC criterion (Model (3.5) in the manuscript) is used to estimate ranks for each simulation scenarios with different missing rates. We apply the proposed BIC criterion to all the scenarios with different missing rate 0%; 5%, 10%. Overall, the adapted BIC criterion performs well for different settings (Table 2.1). The stepwise selection procedure correctly identifies the true ranks for joint structure and individual structures almost all the times for scenarios with two data types with various missing rates. We also apply the selection procedure to scenario **Gaussian-Poisson-binomial**. The proposed adapted BIC criterion is unable to specify the rank combinations correctly (misspecified as $r_J = 3; r_{A_{Gaussian}} = r_{A_{Poisson}} = 1; r_{A_{binomial}} = 2$). This may be because the signal-to-noise ratio for the binomial data is relatively low compared to the other datasets. Alternative approaches to rank selection that can accommodate to multiple (>2) sources of data call for more investigation.

Table 2.1: Rank selection result for Scenario 1 (Gaussian Gaussian), Scenario 2 (Gaussian Poisson), Scenario 3 (Gaussian binomial), and Scenario 5 (binomial Poisson) with different missing rates. The number of correctly specified ranks is out of 50.

	Missing Rate %		
Scenario	0	5	10
Scenario 1	50	50	50
Scenario 2	49	49	50
Scenario 3	50	50	50
Scenario 5	50	50	49

2.2 Simulation Result for Gaussian Poisson binomial Scenario

In addition to the scenarios with data sets from two sources with different data types, we also apply the propose approach to the scenario with three different data types, Gaussian, Poisson and binomial (**Scenario 4**). The ranks for each part are set to be 2. Joint and individual score matrices ($\mathbf{U}_0; \mathbf{U}_1; \mathbf{U}_2; \mathbf{U}_3$) are filled with uniform random numbers $Unif(\ 0{:}5; 0{:}5)$ and normalized to have orthonormal columns.

Scenario 4: **Gaussian-Poisson-binomial** The joint loading matrices V_1 for Gaussian,V_2 for Poisson are generated from $Unif(\ 0{:}5; 0{:}5)$, and V_3 for binomial is generated from $Unif(\ 1{:}5; 1{:}5)$. The individual loading matrices A_1 (Gaussian), A_2 (Poisson), A_3 (binomial) are generated from $Unif(\ 0{:}5; 0{:}5)$, $Unif(\ 0{:}25; 0{:}25)$, and $Unif(\ 1{:}5; 1{:}5)$ correspondingly. The singular values of the joint structure are set to be $(300; 280)$, the singular values of the individual structures to be $(150; 120)$ for Gaussian, $(150; 140)$ for Poisson and $(200; 180)$ for binomial. For such scenario, the imputation accuracy of GIPCA outperforms the other ad hoc methods for three data sets as well.

The means for Gaussian data set in each scenario that contains Gaussian data are generated from $Unif(\ 0{:}5; 0{:}5)$. The means of Poisson distribution to be positive (from $Unif(0; 1)$) to mimic Poisson data in reality. The means for binomial data set are generated from $Unif(\ 1{:}5; 1{:}5)$. For Gaussian data, we set the variance for the generated data to be 1. For binomial data, we set the number of trials to be 100. Similarly, as what is stated in Section 4.1, data are generated from fixed natural parameter matrix. The result in Table

2.2 shows that the proposed GIPCA outperforms the other ad hoc methods.

Table 2.2: Simulation results for **Scenario 4** based on 100 simulation runs when the natural parameter matrices are fixed for each data source. The missing rate is 5%. The median and the median absolute deviation (MAD) for each evaluation criterion under each scenario are calculated. MAD is in parenthesis. The best results are highlighted in bold.

	Scenario 4			
	Gaussian	**Poisson**	**binomial**	**Running Time**
Adhoc1	13.02 (8.06)	8.82 (4.03)	7.10 (2.40)	594.37 (148.9)
Adhoc2	4.60 (1.13)	3.40 (1.31)	3.46 (0.98)	3.38 (0.87)
GIPCA	**0.83 (0.00)**	**0.87 (0.00)**	**0.57 (0.00)**	549.47 (119.39)

2.3 Simulation Results for Negligible Joint Structure

If the joint structure is dominant, due to the reason that the imputation relies on the estimated joint structure, the missing imputation by our proposed approach would be more accurate. If the joint structure is negligible compared with the individual structure, our proposed approach can still handle such a situation. However, since our proposed method directly exploits the joint structure for imputation, the imputation of the missing entries may not be accurate. We explore the scenarios when the signals of the joint and individual structures are comparable in the manuscript (Table 1 in the manuscript and Table 2.2). In addition to the settings in the manuscript, we also explore the scenarios when the signals of the joint and individual structures are distinct. In Table 2.3, we set the true singular values to construct the natural parameter matrix of the joint structure relatively small (1=2, 1=5 or 1=10 of the singular values in the original setting). The results show that the performance of missing imputation for **Gaussian-Gaussian** and **Gaussian-Poisson** scenarios are relatively robust against the change of singular values. For scenarios involving binomial distributions, the performance is sensitive to the change of signal.

2.4 Sensitivity to Initial Values

In order to evaluate how sensitive the algorithm to initial values, we set up different initial values in the proposed algorithm to fit the same two data sets. The data set is generated the same as we described in Section 4.1. For each scenario, we use the same simulated data, but we generate different initial values based on different random seed for the proposed algorithm. Table 2.4 shows that the performance of missing imputation derived by the proposed method is stable, which indicates that our algorithm is not sensitive to initial values.

Table 2.3: Simulation results for two data sets based on 100 simulation runs when the natural parameter matrices are fixed for each data source. A $r_J \quad r_J$ matrix Σ_J is a diagonal matrix whose diagonal elements are the singular values to construct the natural parameter matrix of the joint structure, where r_J is the rank of joint structure. The median and the median absolute deviation (MAD) for each evaluation criterion under each scenario are calculated. MAD is in parenthesis.

		Adhoc1		**Adhoc2**		**GIPCA**	
		Source1	Source2	Source1	Source2	Source1	Source2
Gaussian Gaussian	$DiffR_{Miss}$	11.14 (0.85)	8.41 (0.69)	1.56 (0.00)	1.00 (0.00)	1.15 (0.06)	0.93 (0.01)
Σ_J=2	**Running Time**	100.45 (13.94)		3.26 (0.07)		209.37 (145.26)	
Gaussian Gaussian	$DiffR_{Miss}$	12.22 (1.06)	9.13 (0.77)	1.65 (0.00)	1.00 (0.00)	1.55 (0.01)	1.00 (0.00)
Σ_J=5	**Running Time**	101.51 (14.45)		3.26 (0.04)		138.5 (26.69)	
Gaussian Gaussian	$DiffR_{Miss}$	13.4 (1.37)	8.76 (1.10)	1.67 (0.00)	1.00 (0.00)	1.15 (0.23)	4.51 (5.12)
Σ_J=10	**Running Time**	114.71 (18.98)		3.28 (0.06)		396.54 (295.22)	
Gaussian Poisson	$DiffR_{Miss}$	13.13 (4.98)	6.01 (0.4)	1.23 (0.05)	1.64 (0.01)	0.73 (0.00)	0.74 (0.00)
Σ_J=2	**Running Time**	212.39 (35.78)		1.84 (0.05)		152.27 (35.57)	
Gaussian Poisson	$DiffR_{Miss}$	10.06 (6.39)	7.03 (0.71)	2.81 (0.43)	1.25 (0.02)	0.95 (0.01)	0.99 (0.02)
Σ_J=5	**Running Time**	196.02 (37.53)		1.84 (0.06)		394.18 (138.01)	
Gaussian Poisson	$DiffR_{Miss}$	15.97 (6.19)	8.02 (1.05)	1.17 (0.00)	1.64 (0.01)	1.02 (0.04)	6.91 (2.52)
Σ_J=10	**Running Time**	210.66 (42.03)		1.84 (0.10)		595.02 (196.2)	
Gaussian binomial	$DiffR_{Miss}$	8.06 (3.40)	7.56 (1.72)	1.04 (0.00)	0.99 (0.00)	0.95 (0.01)	0.77 (0.05)
Σ_J=2	**Running Time**	502 (97.77)		1.78 (0.08)		835.26 (483.43)	
Gaussian binomial	$DiffR_{Miss}$	11.09 (5.37)	1.43 (0.32)	1.12 (0.00)	2.05 (0)	1.02 (0.01)	6.54 (0.79)
Σ_J=5	**Running Time**	467.65 (244.06)		1.87 (0.13)		391.85 (87.33)	
Poisson binomial	$DiffR_{Miss}$	11.07 (5.16)	13.26 (6.85)	0.88 (0.06)	5.83 (6.83)	3.63 (2.44)	0.99 (0.03)
Σ_J=2	**Running Time**	190.22 (39.83)		0.39 (0.04)		272.48 (136.3)	

Table 2.4: Simulation results ($DiffR_{Miss}$) for the same two data sets generated by fixed natural parameter matrices for each source and repeated for 100 time with different initial values for the algorithm. The median and the median absolute deviation (MAD) for each evaluation criterion under each scenario are calculated. MAD is in parenthesis.

	Simulated data I	**Simulated data II**
Gaussian Gaussain	0.65 (0.00)	0.72 (0.00)
Gaussian Poisson	0.46 (0.00)	0.47(0.00)
Gaussian binomial	0.77 (0.00)	0.44(0.01)
Poisson binomial	0.59 (0.01)	0.84 (0.00)