CAUSAL INFERENCE WITH MENDELIAN RANDOMIZATION FOR LONGITUDINAL DATA

By

Jialin Qu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2022

ABSTRACT

CAUSAL INFERENCE WITH MENDELIAN RANDOMIZATION FOR LONGITUDINAL DATA

By

Jialin Qu

Mendelian Randomization (MR) uses genetic variants as instrumental variables (IVs) to examine the causal relationship between an exposure and an outcome in observational studies. When confounding factors exist, the correlation between a predictor variable and an outcome variable does not imply causation. IV regression has been a popular method to control the confounding effect for causal inference. According to Mendel's first and second laws of inheritance, genetic variants can be considered as valid IVs. Popular MR methods include the ratio estimator, the inverse-variance weighted estimator and the two stage estimator. However, all these methods are based on cross-sectional data. In practice, data in the observational studies can be collected over time, the so-called longitudinal data. Longitudinal data makes it possible to capture changes within subjects over time and thus offers advantages to causal modeling to establish causal relationships. However, causal inference method that can control the time-varying confounding effect is largely lacking in literature. In this dissertation, we explore MR analysis for longitudinal data by proposing different causal models and assuming different casual mechanisms. The proposed methods are strongly motivated by a real study to examine the causal relationship between hormone secretion and emotional eating disorder in teen girls.

We start with a concurrent model which assumes current outcome is only affected by current exposure. Coefficients of both genetic variants (i.e., IVs) and exposure are considered as time-varying effects. We apply the quadratic inference function approach in a two-step IV regression framework and focus on statistical testing to infer causality. Through extensive simulation studies, we show that the proposed method can well protect type I error and has reasonable testing power.

In Chapter 3, we generalize the concurrent model to a more complex case and propose a time lag model to investigate time delayed causal effects. In the time lag model, we assume current

outcome at time t is affected by previous exposures measured up to $t - s$ time points, where the time lag $\triangle t$ can be determined by a rigorous model selection procedure based on data. Similar to the concurrent model, we assume the effects of genetic variants on exposure and the effects of exposure on outcome both are time-varying. We propose different tests for point-wise and simultaneous testing to assess the causal relationship.

In Chapter 4, We further generalize the time lag model to the case where the cumulative effect of previous t exposures contributes to the outcome at time t, under a sparse functional data analysis framework. The causal relationship is examined under the functional principal component regression framework with sparse functional data. Simulation results show that the type I error is well controlled.

We apply our models to the emotional eating disorder data to examine if hormone secretion during the menstrual cycle in teen girls has a causal effect on emotional eating behavior and identify interesting results. This thesis work represents the very first exploration in MR analysis with longitudinal data.

To my parents and my maternal grandparents.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of many people. I would like to thank them for their help and support during the production of this dissertation and my journey toward PhD.

I am extremely thankful to my supervisor Dr. Yuehua Cui for his noble guidance, support with full encouragement and enthusiasm. He always provides me with constructive insights and strong supports that sharpen my thinking and bring my work to a higher level. I have benefited greatly from your wealth of knowledge and meticulous editing. I am extremely grateful that you took me on as a student and continued to have faith in me over the years.

Thank you to my committee members, Dr. Honglang Wang, Dr. Haolei Weng, Dr. Jianrong Wang and Dr. Marianne Huebner. Your comments and suggestions are extremely beneficial for my research. I would like to acknowledge Dr. Honglang Wang for patiently answering my questions and providing me his valuable suggestions.

I would also like to thank all the professors and staff members in the Department of Statistics and Probability for their help. I truly enjoy the wonderful courses and I appreciate the time they spent helping me in many occasions.

Most importantly, I am grateful for my parents whose constant love and support keep me motivated and confident. Deepest thanks to my boyfriend, Dr. Peide Li. Thanks for your accompany. I am forever thankful for the love and support throughout the entire journey.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1 INTRODUCTION 1
- 1.1 Overview 1
- 1.2 Causal Inference and Mendelian Randomization 2
- 1.3 A Review of Instrumental Variable Estimation 7
  - 1.3.1 Ratio Estimator 7
  - 1.3.2 Inverse-variance Weighted Estimator 8
  - 1.3.3 Weighted Median Estimator 9
  - 1.3.4 Two-stage Estimator 10
  - 1.3.5 Control Function Estimator 11
  - 1.3.6 Likelihood-based Methods 12
- 1.4 Causal Modeling Methods for Longitudinal Data 14
- 1.5 Motivation and Organization 19

CHAPTER 2 CAUSAL INFERENCE WITH TIME-VARYING CONFOUNDING: A MENDELIAN RANDOMIZATION APPROACH 21
- 2.1 Introduction 21
- 2.2 The Concurrent Model 24
  - 2.2.1 Estimating the time-varying SNP effect 25
  - 2.2.2 Estimation and testing of the time-varying exposure effect 28
- 2.3 Simulation Study 29
- 2.4 Case Study: Albert Twin Data 33
  - 2.4.1 Albert Twin data 33
  - 2.4.2 Concurrent Model Application 39
- 2.5 Conclusion and Discussion 40

CHAPTER 3 MENDELIAN RANDOMIZATION WITH TIME LAG EFFECT 43
- 3.1 Introduction 43
- 3.2 Time Lag Model 45
  - 3.2.1 Estimation of the time-varying SNP effect 46
  - 3.2.2 Estimation and testing of the time-varying exposure effect 48
  - 3.2.3 Time lag selection 49
- 3.3 Model Testing 51
  - 3.3.1 Pointwise Testing 51
  - 3.3.2 Simultaneous Test 53
- 3.4 Simulation Study 54
  - 3.4.1 Selection Performance for p and q 54
  - 3.4.2 Performance of the Coefficient Estimation 57

3.4.3 Performance of Simultaneous Testing . . . 62
3.5 Case Study: Albert Twin Data . . . 64
3.6 Conclusion and Discussion . . . 67

CHAPTER 4 MENDELIAN RANDOMIZATION FOR LONGITUDINAL DATA WITH CUMULATIVE EFFECT . . . 70
4.1 Introduction . . . 70
4.2 Functional Model . . . 72
4.2.1 Estimation of the time-varying SNP effect . . . 73
4.2.2 Estimation and testing of the functional exposure effect . . . 73
4.2.3 Select the number of eigen-functions . . . 76
4.2.4 Functional F test . . . 77
4.3 Simulation Study . . . 77
4.4 Case Study: Albert Twin Data . . . 80
4.5 Conclusion and Discussion . . . 82

CHAPTER 5 CONCLUSION AND FUTURE WORK . . . 84
5.1 Conclusion . . . 84
5.2 Future Work . . . 85

APPENDICES . . . 87
APPENDIX A APPENDIX FOR CHAPTER 2 . . . 88
APPENDIX B APPENDIX FOR CHAPTER 3 . . . 91
APPENDIX C APPENDIX FOR CHAPTER 4 . . . 93

BIBLIOGRAPHY . . . 98

LIST OF TABLES

Table 2.1 Effect of confounding on the type I error and power. 32

Table 2.2 Effect of within sample correlation on the type I error and power. 33

Table 2.3 Subject characteristics at baseline n=225. 34

Table 3.1 Simultaneous testing simulation result. 63

Table 4.1 List of Type I error and power under different sample size and time points. 80

Table A.1 Simulation Results under different IV strength. 89

Table A.2 Subject Characteristics. 90

Table C.1 Effect of within sample correlation on Type I error and power for functional MR model. 94

Table C.2 Effect of confounding on Type I error and power for functional MR model. . . . 95

Table C.3 Effect of within sample correlation on Type I error and power using functional F test. 96

Table C.4 Effect of confounding on Type I error and power using functional F test. 97

LIST OF FIGURES

Figure 1.1 Diagram of instrumental variable assumptions. 5

Figure 1.2 Unit Causal Graph. 18

Figure 2.1 Instrument Variable. 22

Figure 2.2 Flow diagram of subject selection. 35

Figure 2.3 Diagram of the causal relationship for four models (combinations of hormone levels and eating behavior). 36

Figure 2.4 The observed hormones levels and eating behaviour measurements for the first 100 subjects. 37

Figure 2.5 Distribution of Age and BMI at baseline. 38

Figure 2.6 Coefficients estimation for the relationship between PRO and DEBQ. 40

Figure 2.7 Coefficients estimation for the relationship between PRO and PANAS. 40

Figure 3.1 Boxplot of the time lag selection under different true values of p and q. 57

Figure 3.2 Boxplot of the time lag selection under different true values of p and q. 57

Figure 3.3 Coefficient estimation when $p = 1$. The solid red curve represents the true effect function. The solid black curve and dashed blue curves in each figure represent the estimated effect function and the 95% confidence interval, respectively. 59

Figure 3.4 Coefficient estimation when $p = 2$. The solid red curve represents the true effect function. The solid black curve and dashed blue curves in each figure represent the estimated effect function and the 95% confidence interval, respectively. 61

Figure 3.5 Relationship between pro and DEBQE. 65

Figure 3.6 Relationship between Pro and DEBQE. 66

Figure 3.7 Relationship between Pro and DEBQE. 66

Figure 3.8 Pointwise p-value for DEBQE. 67

Figure 4.1 Smoothed progesterone mean function. 81

Figure 4.2 Plot of -log10(p-value) for DEBQ. 81

Figure 4.3 Plot of -log10(p-value) for PANAS. 82

CHAPTER 1

INTRODUCTION

1.1 Overview

In this dissertation, we studied the causal relationship between an exposure and a response variable with repeated measurements. By experimenting with pea plant breeding, Mendel developed three principles of inheritance that described the transmission of genetic traits, before anyone knew genes existed in the nineteenth century[56]. Mendel's insight greatly expanded the understanding of genetic inheritance, and researchers have been fascinated with the role of genetics played in our lives. Using genetic variants as instrumental variables, Mendelian Randomization is a research method that aims to investigate the causal relationship between modifiable risk factors and disease.

The increasing use of Mendelian Randomization has prompted a huge number of research. The general aim of the Mendelian Randomization approach is the estimation of a causal effect of an exposure on an outcome using (one or more) genetic instruments for the exposure. Ratio estimator[79], inverse-variance weighted estimator[52], weighted median estimator[8], two stage estimator and some nonparametric estimators, etc. are developed for this purpose. Another popular direction is discussing the validity of Mendelian Randomization since its validity is based on three key assumptions: 1) the instrument variable is associated with the exposure, 2) the association between the instrument variable and the outcome is unconfounded, and 3) the instrument variable only affects the outcome via the exposure, known as the exclusion restriction criterion. A primary cause of violation of the exclusion restriction criterion, pleiotropy, where a genetic variant affects the exposure and the outcome through independent pathways and without being mediated by another is fully discussed[7][39][77][41]. Besides, pleiotropy, estimation bias may be also caused by weak instruments[13][22][11]. Most of the Mendelian Randomization focus on one-time measurement, while longitudinal observations capture change within subjects over time and contain more information. Thus, incorporating the time information into the analysis can

potentially improve the causal inference.

In this chapter, we first provided some background information on causal inference and Mendelian Randomization in section 1.2. The commonly used IV estimation methods for MR is then reviewed in section 1.3. In section 1.4, we discussed causal modeling methods for longitudinal data. The goal and organization of this dissertation are offered in section 1.5.

1.2 Causal Inference and Mendelian Randomization

The gold standard method to address both confounding and causality is a randomised controlled trial (RCT). RCT is a trial in which subjects are randomly assigned to one of two groups: one (the experimental group) receiving the intervention that is being tested, and the other (the comparison group or control) receiving an alternative (conventional) treatment[47]. The two groups are then followed up to see if there are any differences between them in outcome. RCTs are the most stringent way of determining whether a cause-effect relation exists between the intervention and the outcome. However, for many research questions, it is impossible or unethical to randomly assign the treatment. For example, it would not be possible nor acceptable to randomly allocate obesity. Even if we could randomly assign the treatment, there are still several challenges in conducting a good quality RCT. RCTs are time consuming and it may take many years before the results are available for analysis. RCTs need a large number of participants in a trial to ensure sufficient statistical power. The general orthodontic trials that look at data from start to the end of orthodontic treatment will run for at least five years. For these reasons, RCTs usually have expensive cost. Moreover, it is hard to generalize RCTs' result due to its low external validity. The intervention may only work for a particular group of people in that context instead of working in the same way for a different group in a different context.

Compared with RCTs, observational studies are more common and becoming a key part of research. In an observational study, no intervention takes place. Observational studies are ones where researchers are looking at the effect of some type of risk factor, diagnostic test, treatment or other intervention, without trying to manipulate who is, or who is not exposed to it. Observational

studies are generally used in hard science, medical, and social science fields. There are two main types of observational studies: cohort studies and case control studies. A cohort is a group of people who are linked in a particular way, for example, a birth cohort would include people who were born within a specific period of time. Cohort studies enroll a population at risk and follow them for a period of time. Individuals who develop the disease in that time are then compared with individuals who remain disease-free[72]. Researchers in case control studies identify individuals with an existing health issue or condition, or "cases", along with a similar group without the condition, or "controls". The two groups are then compared, to see if the case group exhibits a particular characteristic more than the control group. The main challenge in case-control studies is to identify an appropriate control group with characteristics similar to those of the general population at risk for the disease. However, when studies lack control and treatment groups, this will result in the difficulty of inferences, and confounding variables may further complicate the results.

A confounder has long been defined as any third variable that is associated with the exposure of interest, is a cause of the outcome of interest, and does not reside in the causal pathway between the exposure and outcome[53]. When confounders exist, correlation does not mean causation and we cannot conclusively say that any difference observed in mortality (or any other outcome of interest) between the two groups is due solely to the treatment. To remove the influence of confounding factors, there exist many well-developed causal inference methods. Matching is employed to make the multivariate distribution of all covariates X as similar as possible by selecting appropriate control observation(s) for each treatment observation[55]. There are, at least, four primary ways to define the distance measures between individuals for matching: exact, Mahalanobis distance and the propensity score or the linear logits predicted by the logit-model. In many ways, the ideal matching is exact matching, however, the primary difficulty with the exact and Mahalanobis distance measures is that neither works very well when X is high dimensional[45]. Another drawback of exact matches is that it can result in larger bias compared with the matches that are not exact, because requiring exact matches often leads to many individuals not being matched, on the other

hand, inexact matches often make more individuals remain in the analysis[67].

Propensity score (PS) methods are among the most popular approaches for causal inference in clinical and epidemiologic research. A PS is a conditional probability of receiving a treatment/exposure given a set of covariates: $PS = Pr(A = 1|L)$. PS is estimated by specifying a propensity model (ie, a model for an exposure), typically via logistic regression. After estimating PS, there are several alternative approaches to control for the estimated PS. These approaches include stratification, regression adjustment, matching, and inverse probability weighting (IPW)[70]. With similar estimated PS, PS matching creates pairs of exposed and unexposed subjects. By excluding observations from individuals with extremely large or small PS if they lack corresponding pairs, the exposed and unexposed groups in the remaining sample of the matched pairs are expected to have comparable distributions of PS and observed confounders that are used in PS estimation. Then the corresponding causal effect can be calculated by the difference in the conditional expectations: $E[Y^{a=1} - Y^{a=0}]$ under the identifiability assumptions.

Besides matching, weighting is another popular technique to remove the influence of measured confounding factors. Propensity scores can also be used directly as inverse weights in estimates of the average treatment effect, known as inverse probability of treatment weighting (IPTW)[20]. Weights for IPTW are typically defined as a function of PS. The average treatment effect weights for the groups are defined as $w(A, X) = \frac{A}{P(X)} + \frac{1-A}{1-P(X)}$ which result in $\frac{1}{P(X)}$ for the exposed individuals with $A = 1$ and $\frac{1}{1-P(X)}$ for the unexposed individuals with $A = 0$. IPTW essentially duplicates observations from individuals with large weights to create a pseudo-population in which probabilities of receiving the exposure A do not depend on the covariates L included in the PS estimation[70]. This weighting serves to weight both the treated and control groups up to the full sample, in the same way that survey sampling weights weight a sample up to a population[43].

However, even when we know about a confounder, we are unlikely to measure it perfectly, especially for complex situations such as in socioeconomic circumstances. The above mentioned methods can only control measured confounding factors. There will also be confounders we do not know about, have not measured and have not considered. This means there is still some confounding

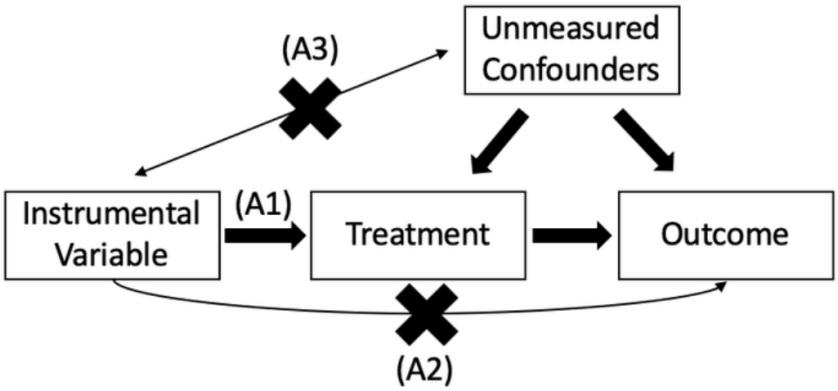


Figure 1.1 Diagram of instrumental variable assumptions.

(residual confounding) in most observational studies. Instrumental variable regression approach is developed to deal with both measured and unmeasured confounding factors.

An Instrumental Variable (IV) is used to control for confounding and measurement error in observational studies so that causal inferences can be made. Specifically, an instrumental variable Z is an additional variable used to estimate the causal effect of variable X on Y. The variable Z is qualified as an instrumental variable (relative to the pair (X, Y)) if it satisfied the following three conditions shown in Figure 1.1: (i) Z is not independent of X. (ii) Z does not have a direct influence on Y which is referred to as the exclusion restriction. (iii) Z is independent of all variables (including error terms) that have an influence on Y that is not mediated by X. Therefore, the effects of the instrumental variable Z on Y only work through its effect on X. Consequently, variable Z is unrelated to the outcome (Y) but is related to the predictor (X) and is not causally affected (directly or indirectly) by X, Y, or the error term U. In this approach, not only one but also multiple IVs and/or causal paths could be used. The instrumental variable approach for controlling unobserved sources of variability is the mirror opposite of the propensity score method for controlling observed variables[2].

Mendelian Randomization (MR) can be regarded as an application of instrumental variable approach to find causal inference. As we discussed before, there are many different options of valid instrument variable as long as it satisfies the above mentioned three conditions. Mendelian

Randomization specifies genetic variants as instrument variables to address causal questions about how modifiable exposures influence different outcomes. Genetic variants are small parts of the genome which can be closely related to human characteristics (e.g. height, weight, blood pressure) and health conditions (e.g. diabetes, coronary heart disease, asthma). According to Mendel's first and second laws of inheritance, genetic variants can be considered as valid instrument variables. Mendel's first law describes that the two alleles at a gene locus segregate from each other during gamete formation and each gamete has an equal probability of containing either allele. Mendel's second law describes the independent segregation of a pair of traits and another pair during gamete formation. Together, the two laws imply that offspring have an equal chance of inheriting an allele from either parent, and that these alleles are inherited independently from one another[40]. Nevertheless, specific assumptions still need to be fulfilled to ensure the validity of the genetic variant as an instrument.

1. The genetic variant is associated with the exposure.

2. The genetic variant is independent of the outcome given the exposure and all confounders (measured and unmeasured) of the exposure-outcome association.

3. The genetic variant is independent of factors (measured and unmeasured) that confound the exposure-outcome relationship.

The most important decision to be made in designing a Mendelian randomization investigation is which genetic variants to include in the analysis[76]. One traditional SNPs selection method of Mendelian randomization studies are implemented using independent genetic variants from across the whole genome. This genome-wide analyses rely on published results from large-scale GWAS studies and genetic variants in each region are pruned for independence. The selected variants are those with the smallest p-value, or a small number of weakly correlated variants with small p-values. To improve the power of the analysis, researchers combine variants from different regions to create a genome-wide set of instruments for MR. Conversely to the more traditional type of genome-wide MR described above, Cis-MR studies have grown in popularity. Genetic variants for

cis-MR are selected from a region containing the protein-encoding gene. The selection of genetic variants is usually performed based on the strength of their associations with the risk factor, while accounting for LD correlation to reduce numerical approximation errors. To include all variants that are associated with the exposure of interest, a given level of statistical significance (typically, a genome-wide significance threshold, such as $p < 5 \times 10^{-8}$) is applied. After selecting valid genetic variants, those instrumental variables are then used for further study. There are several well-developed methods available for MR using instrumental variable estimation.

1.3 A Review of Instrumental Variable Estimation

There are several methods available for instrumental variable estimation. We give brief introduction of those Mendelian Randomization investigations in this section.

1.3.1 Ratio Estimator

The Wald ratio method is the easiest way to calculate the causative effect of an exposure (X) on an outcome (Y). Assume a continuous outcome Y and an dichotomous IV Z which takes values 0 or 1, dividing the population into two genetic subgroups. We define $\bar{y}_j$ for $j = 0, 1$ as the average value of outcome for all individuals with genotype $Z = j$ and define $\bar{x}_j$ similarly for the exposure. Then an average difference in the exposure between the two subgroups is calculated as $\triangle X = \bar{X}_1 - \bar{X}_0$ and an average difference in the outcome can be computed by $\triangle Y = \bar{Y}_1 - \bar{Y}_0$. IV estimates are usually expressed as the change in the outcome resulting from a unit change in the exposure, although changes in the outcome corresponding to different magnitudes of change in the exposure could be quoted instead[12]. If we assume the linear relationship between the exposure and the outcome, the ratio estimator simplifies to

$$\text{Ratio method estimate (dichotomous IV)} = \frac{\triangle Y}{\triangle X} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}.$$

Alternatively, the IV may not be dichotomous, but continuous. Suppose the coefficient of the IV in the regression of exposure X on the IV Z is written as $\hat{\beta}_{X|Z}$, and represents the change in X for a

unit change in Z. Similarly, the coefficient of the IV Z in the regression of outcome Y on the IV Z is written as $\hat{\beta}_{Y|Z}$. Then the ratio estimate of the causal effect is

$$\text{Ratio estimate} = \frac{\hat{\beta}_{Y|Z}}{\hat{\beta}_{X|Z}}.$$

This ratio estimator can be explained by saying that the change in Y for a unit increase in X is equal to the change in Y for a unit increase in Z, scaled by the change in X for a unit increase in Z. The ratio estimator has been named the linear IV average effect[25] since the validity of ratio estimator is restricted to the assumption of monotonicity of the genetic effect on the exposure and linearity of the causal $X \rightarrow Y$ association[1].

We next consider the situation where outcome variable Y is binary instead of continuous. This is very common in epidemiology, where the outcome of interest is disease status and is often dichotomous. In reality, people often use $Y = 1$ to refer an individual who has an outcome of interest or disease, and use $Y = 0$ to describe an individual who does not show the particular phenotypic trait or disease. Similar to the continuous case, the ratio estimate simplifies as with a continuous outcome when the IV Z is dichotomous:

$$\text{Ratio method log relative risk estimate (dichotomous IV)} = \frac{\triangle Y}{\triangle X} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}.$$

However, things become a little complicated under the continuous IV condition. In this case, instead of fitting a linear regression model, a log-linear model or a logistic regression model is generally preferred to estimate the coefficient $\hat{\beta}_{Y|Z}$ in the regression of outcome Y on the IV Z. The ratio estimate is also commonly quoted in its exponentiated form as:

$$\text{Ratio method risk ratio estimate (dichotomous IV)} = R^{1/\triangle X}$$

where R is the estimated risk ratio between the two genetic subgroups.

1.3.2 Inverse-variance Weighted Estimator

For genetic variant j, the Wald ratio estimate is consistent asymptotically if the IV assumptions are satisfied. Furthermore, if the genetic variants are uncorrelated (i.e., in linkage equilibrium)

then the ratio estimates from each genetic variant can be combined into an overall estimate using a formula[46]. We can then generalize the Wald ratio estimator through a meta-analysis process if several genetic variants are correlated with a specific exposure. Burgess et al.[9] proposed the inverse-Variance Weighted estimator shown as follows:

$$\text{Inverse-variance weighted estimate} = \frac{\sum_j \hat{\beta}_{X|Z_j} \sigma_{Yj}^{-2} \hat{\beta}_j}{\sum_j \hat{\beta}_{X|Z_j} \sigma_{Yj}^{-2}},$$

where $\hat{\beta}_{X|Z_j}$ is the coefficient of the genetic variant j in the regression of exposure X on the IV Z_j; σ_{Yj} is the standard error of the gene-outcome association estimate for variant j and $\hat{\beta}_j$ defines the causal effect of the exposure on the outcome which is estimated using the jth variant as the ratio of the gene-outcome association and the gene-exposure association estimates. If all genetic variants satisfy the IV assumptions, then the IVW estimate is a consistent estimate of the causal effect (i.e., it converges to the true value as the sample size increases), as it is a weighted mean of the individual ratio estimates[8].

1.3.3 Weighted Median Estimator

The inverse-variance weighted estimator is efficient when all genetic variants are valid IVs, but it will be biased under the invalid IVs situation. The median ratio estimator suggested by Han [38] aims to deal with this challenge. The median ratio estimator can guarantee a consistent causal effect estimate when up to (but not including) 50% of genetic variants are invalid. Similar to the construction of inverse-variance weighted estimator, if we assume $\hat{\beta}_j$ denote the jth ordered ratio estimate of the causal effect (arranged from smallest to largest) of the exposure on the outcome which is estimated using the jth variant as the ratio of the gene-outcome association and the gene-exposure association estimates. The simple median estimator is defined as the middle ratio estimate. Thus, if the total number of genetic variants is odd ($J = 2k + 1$), then simple median estimator is $\hat{\beta}_{k+1}$; if the total number of genetic variants is even ($J = 2k$), the median is interpolated between the two middle estimates, i.e. $\frac{1}{2}(\hat{\beta}_k + \hat{\beta}_{k+1})$.

However, when the precision of the individual estimates varies considerably, the simple median estimator is inefficient. Bowden et al.[8] considered a weighted median estimator to account for this situation. We define the weight given to the jth ordered ratio estimate be w_j, then the sum of weights up to and including the weight of the jth ordered ratio estimate is denoted as $s_j = \sum_{k=1}^{j} w_k$. In order to make the sum of the weights be equal to 1, all weights are standardized. The weighted median estimator is the median of a distribution having estimate $\hat{\beta}_j$ as its $p_j = 100(s_j - \frac{w_j}{2})$th percentile. The simple median estimator can be thought of as a weighted median estimator with equal weights. The weighted median will provide a consistent estimate if at least 50% of the weight comes from valid IVs.

1.3.4 Two-stage Estimator

Two regression stages are included to construct the two-stage estimator: in the first step, by calculating the fitted values from the regression of the exposure X on the IVs Z, the fitted values of exposure $\hat{X}$ is estimated via the genotypes of the instruments. This fitted values of exposure $\hat{X}$ is independent of the confounders. In the second step, the causal effect estimate is obtained by regressing the outcome on the fitted values of the exposure from the first stage. The causal estimate is this second-stage regression coefficient for the change in the outcome caused by a unit change in the exposure. The models are written as follows:

$$X = \alpha_0 + \alpha_1 Z + \epsilon_1$$

$$Y = \beta_0 + \beta_1 \hat{X} + \epsilon_2$$

When the outcome variable Y is binary, the second-stage regression uses a log-linear or logistic regression model. However, if we apply the non-linear model in the second step regression, the model can not guarantee that residuals from the second-stage regression are uncorrelated with the instrument[33].

The two-stage estimator is consistent for the causal effect when all relationships are linear and there are no interactions between the instrument and unmeasured confounders and between the

exposure and unmeasured confounders. Provided that the genetic variants are uncorrelated, the IVW estimate is asymptotically equal to the two-stage least squares estimate commonly used with individual-level data.

1.3.5 Control Function Estimator

Another method for estimating the causal effect is provided through the control function estimator which is also a two-step approach. The first step is the same as the first step in the two-stage estimator. The exposure X is regressed on the IVs Z. In the second step, When the residuals of the first step are included as an additive covariate, these estimators have been referred to as 2-stage residual inclusion (TSRI) estimators. The models can be constructed as follows:

$$X = \alpha_0 + \alpha_1 Z + \epsilon_1$$

$$h(E(Y)) = \beta_0 + \beta_1 X + \beta_2 \hat{\epsilon}_1$$

where $h(\cdot)$ is the link function for an appropriate generalized linear model. Linear regression is used at the second stage when outcome Y is continuous, while logistic regression is applied for the binary outcome Y.

The standard errors of the second-stage parameter estimates are not correct when calculating two-stage estimator by fitting the 2-stage least square regressions sequentially. The standard error of the coefficient on the first-stage residuals is correct when we apply linear regression to construct control function estimator[83]. It is well known that when using linear regression function $h(\cdot)$, the control function estimator produces an estimate of the causal effect equivalent to the two-stage estimator[23]. Newey[57] developed a correction to the standard errors of the second-stage intercept and causal effect of the probit control function estimator for a binary outcome. The following algorithm is described by Newey[59]:

1. Perform the first-stage linear regression of X on Z to compile matrix $\hat{D}$ and $\hat{\epsilon}_1$, where $\hat{D}$ is defined as $\begin{bmatrix} \hat{\alpha}_1 & 0 \\ \hat{\alpha}_0 & 0 \end{bmatrix}$.

2. Perform a probit regression of Y on Z and $\hat{\epsilon}_1$. Define $\hat{\gamma}$ be the coefficients of Z and the estimated intercept; J_1^{-1} be the variance-covariance matrix of these coefficients. $\hat{\lambda}$ be the coefficient on $\hat{\epsilon}_1$.

3. Fit the second stage of the probit control function estimator by a probit regression of Y on X and $\hat{\lambda}$. Then the estimate of the causal effect of interest is the coefficient on X, $\hat{\beta}_1$.

4. Generate a new variable equal to $X(\hat{\lambda} - \hat{\beta}_1)$. Perform a linear regression of this new variable on Z (also including a constant). Define the covariance matrix from this model as Σ_2. Let $\hat{\Omega} = J_1^{-1} + \Sigma_2$.

5. Calculate $\hat{\beta} = (\hat{D}'\hat{\Omega}^{-1}\hat{D})^{-1}\hat{D}'\hat{\Omega}^{-1}\hat{\gamma}$ and $var(\hat{\beta}) = (\hat{D}'\hat{\Omega}^{-1}\hat{D})^{-1}$.

Palmer et al. [59] further showed that control function estimators with modified standard errors had correct type I error under the null. Researchers should report control function estimates with modified standard errors instead of reporting unadjusted or heteroscedasticity-robust standard errors.

1.3.6 Likelihood-based Methods

If we assume the following model which is the same as the model for the two-stage estimator:

$$X = \alpha_0 + \alpha_1 Z + \epsilon_1$$

$$Y = \beta_0 + \beta_1 \hat{X} + \epsilon_2$$

where the error term $\epsilon = (\epsilon_1, \epsilon_2)$ has a bivariate normal distribution $\epsilon \sim N(0, \Sigma)$ and the correlation between ϵ_1 and ϵ_2 is caused by confounding factors. Then we can simultaneously calculate the maximum likelihood estimates of β_1 by full information maximum likelihood method proposed by Davidson and MacKinnon[21]. This method requires the correctly specified regression equations at each step to estimate a consistent estimate of β_1, while in reality, only coefficient β_1 is our interest. To overcome this drawback, the limited information maximum likelihood method is used

by profiling out each of the parameters except β_1 and only using limited information on the structure of the model. The causal effect β_1 is estimated by minimizing the residual sum of squares from the regression of the component of the outcome not caused by the exposure on the IVs.

For a binary outcome Y, we can assume a linear model of association between the logit-transformed probability of an event (π_i) and the exposure (a logistic-linear model), and a Bernoulli distribution for the outcome event, as in the following model

$$x_i \sim N(\mu_i, \sigma_X^2)$$

$$y_i \sim \text{Bernoulli } (\pi_i)$$

$$\mu_i = \alpha_0 + \alpha_1 Z$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \mu_i + \beta_2 (x_i - \mu_i)$$

All coefficients are estimated simultaneously by maximizing the joint likelihood L, which has the following form:

$$L = \prod_{i=1,\cdots,N} (\pi_i^{y_i} (1-\pi_i)^{1-y_i} \frac{1}{\sqrt{2\pi}\sigma_X} \{\exp(-\frac{1}{\sigma_X^2}(x_i - \mu_i)^2)\})$$

Alternatively, model parameters can also be estimated in a Bayesian framework, obtaining posterior distributions from the model by Markov chain Monte Carlo (MCMC) methods[10].

When a single IV is used for analysis, the limited information maximum likelihood method gives the same causal estimate as the Wald ratio method and the two stage least square method. Compared with two-stage method, which performs two regressions sequentially and the output from the first-stage regression is fed into the second-stage regression with no acknowledgement of uncertainty, the likelihood-based methods perform two stages simultaneously and parameters are estimated at the same time. The limited information maximum likelihood is strongly recommended when there exist weak instrument variables, since the median of the distribution of the estimator is close to unbiased even with weak instruments[3].

However, all above mentioned Mendelian Randomization methods are developed based on cross-sectional data. In practice, data in the observation studies are often collected through

longitudinal studies. Unlike in a cross-sectional design, where all measurements are obtained in a fixed time point, the data in a longitudinal design track the same type of information on the same subjects over time. Longitudinal data makes it possible to capture changes within subjects over time and thus gives some advantages to causal modeling in terms of providing more knowledge to establish causal relationships[34]. Aside from this, more data over longer periods of time will allow for more concise and better results. Longitudinal data is considered highly valid for identifying long-term variations and is distinctive in connection with being able to provide useful data about these individual changes. Another advantage is that they are known to have more power than cross-sectional studies when it comes to excluding time-invariant and unobserved individual differences and when it comes to observing a certain event's temporal order, as they use repeated observations at individual levels.

1.4 Causal Modeling Methods for Longitudinal Data

A number of causal modeling methods have been developed for longitudinal data. In recent years, an increasing number of studies have used time-series methods based on the notion of Granger causality. The first formalization of a practically quantifiable causality definition from time series is the concept of Granger causality suggested by Granger[36]. The construction of Granger causality is based on comparing two models: the first one is predicting a stochastic process Y using all the information in the universe, denoted with U; the second one is doing the same using all information in U except for some stochastic process X, which is denoted with $U \backslash X$. Granger causality defines X as the cause of Y if discarding X reduces the predictive power regarding Y, which shows the past values of X contain helpful information for predicting the future value of Y. Granger causality evokes the following two fundamental principles[37]:

1. The effect does not precede its cause in time.

2. The causal series contains unique information about the series being caused that is not available otherwise.

The first principle of temporal precedence of causes is commonly accepted and has also been the basis for other probabilistic theories of causation[35]. By contrast, the second principle is more subtle, as it requires the separation of the special information provided by the former series X from any other possible information[28].

Granger's original argument is based on the identifiability of a unique linear model. This model is known as vector auto-regressive (VAR) model and we state two VAR models here. The first one is called the restricted model and it assumes that Y linearly depends only on past values of itself with linear coefficients γ_i and a time-dependent noise term e_t:

$$Y_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i Y_{t-i} + e_t.$$

Another model to deal with Granger causality is the unrestricted vector auto-regressive model which assumes that Y linearly depends on past values of both X and Y, determined by coefficients α_i, β_i and a time-dependent noise term u_t:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{p} \beta_i X_{t-i} + u_t$$

The unformalized null hypothesis is that the second model does not add information, or provides a better model of Y, when comparing it to the first model. This needs to be formalized into a testable null hypothesis; a common approach to state that the null hypothesis H_0 is that $\beta_i = 0$ for every i.

According to Shojaie and Fox[71], there are a number of implicit and explicit restrictive assumptions required for the VAR model to be an appropriate framework for identifying Granger causal relationships:

- Continuous-valued series. All series are assumed to have continuous-valued observations. However, many interesting data sources—such as social media posts or health states of an individual—are discrete-valued.

- Linearity. The true data generating process, and correspondingly the causal effects of variables on each other, is assumed to be linear. In reality, many real-world processes are non-linear.

- Discrete-time. The sampling frequency is assumed to be on a discrete, regular grid matching the true causal time lag. If the data acquisition rate is slower or otherwise irregular, causal effects may not be identifiable. Likewise, the analysis of point processes or other continuous-time processes is precluded.

- Known lag. The (linear) dependency on a history of lagged observations is assumed to have a known order. Classically, the order was not estimated and taken to be uniform across all series.

- Stationarity. The statistics of the process are assumed time-invariant, whereas many complex processes have evolving relationships (e.g., brain networks vary by stimuli and user activity varies over time and context).

- Perfectly observed. The variables need to be observed without measurement errors.

- Complete system. All relevant variables are assumed to be observed and included in the analysis, i.e., there are no unmeasured confounders. This is a stringent requirement, especially given that early approaches for Granger causality focused on the bivariate case—that is, they did not account for any potential confounders.

The Granger causality can be tested by SSR-based F-test.

$$F = \frac{(RSS_R - RSS_{UR})/p}{RSS_{UR}/(T - 2p - 1)} \sim F_{p,T-2p-1}$$

where RSS_R and RSS_{UR} are the residual sum of squares for the restricted model and unrestricted model respectively; T is time series length and p is the number of lags. Alternatively, one can also use a χ^2 statistic based on likelihood ratio or Wald statistics[19].

There also exist some limitations for Granger causality. For example, it does not provide any insight on the relationship between the variable, hence it is not true causality unlike 'cause and effect' analysis. Besides, Granger causality fails to forecast when there is an interdependency between two or more variables. Moreover, Granger causality test cannot be performed on non-stationary data.

Besides VAR model, Structural Equation Model (SEM) is another popular method to analyze causal relationship for longitudinal data. SEM refers to the complex of multivariate statistical methods aiming to specify, estimate and fit a system of linear equations to a dataset[5]. The SEM consists of two major parts: the first part is a set of equations that give the causal relations between the substantive variables of interest and the second part ties the observed variables or measures to the substantive latent variables. The model can be described as follows:

$$\eta_i = \alpha_\eta + B\eta_i + \Gamma\xi_i + \zeta_i$$

$$y_i = \alpha_y + \Lambda_y\eta_i + \epsilon_i$$

$$x_i = \alpha_x + \Lambda_x\xi_i + \delta_i$$

where i stands for the ith case, η_i is a vector of the latent endogenous variables, α_η is a vector of intercepts, B is a matrix of coefficients that gives the expected effect of the η_i on η_i where its main diagonal is zero, ξ_i is the vector of latent exogenous variables, Γ is the matrix of coefficients that gives the expected effects of ξ_i on η_i, and ζ_i is the vector of equation disturbances that consists of all other influences of η_i that are not included in the equation, y_i is the vector of indicators of η_i and x_i is the vector of indicators of ξ_i[6]. Rahmadi et al.[65] further proposed stable specification search in constrained structural equation modeling to investigate causality on longitudinal data. This approach used exploratory search but allowed incorporation of prior knowledge, e.g., the absence of a particular causal relationship between two specific variables. They represented causal relationships using structural equation models and applied a multi-objective evolutionary algorithm to search for Pareto optimal models.

VAR and SEM framework assume a linear system and independent Gaussian noise. Some other methods, interestingly, take advantage of nonlinearity or non-Gaussian noise to gain even more causal information. Chu et al. [17] considered an additive non-linear time series model by imposing linear constraints only among contemporaneous variables. They showed that for data generated from stationary models of this type, two classes of conditional independence relations among time series variables and their lags could be tested efficiently and consistently using tests

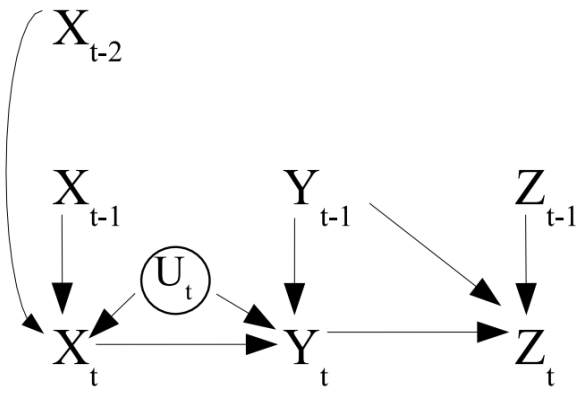


Figure 1.2 Unit Causal Graph.

based on additive model regression.The model is defined as follows:

$$X_{ti} = \sum_{1 \le j \le p, j \ne i} c_{j,i} X_{t,j} + \sum_{1 \le k \le p, 1 \le l \le T} f_{k,i,l}(X_{t-l,k}) + \sum_{m=1}^{q} b_{m,i} U_{t,m} + \epsilon_{t,i}$$

where X_t is a p-dimensional observed time series, U_t is a q-dimensional unobserved time series, $b_{m,i}$'s and $c_{j,i}$'s are constants, and $f_{k,i,l}$'s are smooth univariate functions. This non-linear model can be represented by a directed graph consisting of nodes for $X_{T+1,1}, \cdots, X_{T+1,p}$ and their direct causes, and directed edges between nodes for the direct influences between the corresponding variables. The directed graph is called a unit causal graph and is shown in Fig 1.2. Additive non-linear time series models make it possible to use the additive regression method, which is not subject to the curse of dimensionality, to test conditional independence for nonlinear time series.

Hyv̈arinen et al.[44] considered the general case where causal influences could occur either instantaneously or with considerable time lags and combined the non-Gaussian instantaneous model with autoregressive models. The causal dynamics model are a combination of autoregressive and structural-equation models and is defined as

$$x(t) = \sum_{\tau=0}^{k} B_\tau x(t-\tau) + e(t)$$

Here, $x(t)$ is a single vector collecting the observed time series for all the variables, B_τ denotes the $n \times n$ matrix of the causal effects between the variables with time lag τ and $e_i(t)$ are random

processes modelling the external influences or "disturbances" and are assumed to be mutually independent, and temporally uncorrelated nonGaussian process. To estimate the defined model, they further proposed method which combined classic least-squares estimation of an autoregressive (AR) model with linear non-Gaussian acyclic model (LiNGAM) estimation. They showed that this variant of the non-Gaussian model was identifiable without any other restrictions than acyclicity.

1.5 Motivation and Organization

Although several studies have been done to investigate causal inference for longitudinal data, hardly any of them consider using instrumental variable methods. One advantage of using instrumental variable regression to deal with causal inference analysis is that it can not only remove the effects of measured confounding factors, but also work for the unmeasured confounding factors. While the existence of time-varying confounding effect may fail many existing methods, using IVs can well handle such time-varying confounding in causal inference with longitudinal data.

The thesis is well motivated by a real study to investigate the causal effect of hormone level on emotional eating behavior in teen girls from the Twin Study of Hormones and Behavior across the Menstrual Cycle project [49] from the Michigan State University Twin Registry (MSUTR) [48, 14, 15]. Two hormones were measured, namely estradiol and progesterone. The goal was to evaluate if changes in these two hormones were associated with emotional eating across the menstrual cycle, and further assess if the relationship was causal. Emotional eating was measured with the Dutch Eating Behavior Questionnaire (DEBQ) and negative affect was measured with the Negative Affect scale from the Positive and Negative Affect Schedule (PANAS). The tendency to eat in response to negative emotions is assessed by DEBQ, while negative emotional states like sadness and anxiety are measured by PANAS. Each participant was measured for 45 consecutive days. Data were then grouped into eight menstrual cycle phases, that is, ovulatory phase (1), transition ovulatory to midluteal (2), midluteal phase (3), transition midluteal to premenstrual (4), premenstrual phase including the first day of menstrual cycle (5), remaining days of menstrual cycle, part of follicular phase (6), follicular phase (7) and transition follicular to ovulatory phase

(8), based on profiles of changes in estrogen and progesterone across the cycle [50]. Within each phase, we took two averaged measurements, which ended up with a total 16 data points for each individual. In this project, the exposures will be the two hormone levels measured at 16 time points and the outcome will be the two eating behaviors (DEBQ and PANAS). The goal is to evaluate if there exists causal relationship between hormone levels and eating behaviors and if so, what are the effect mechanisms. The potential effect mechanism may include:

1). concurrent effect, that is, the exposure at time t affects the outcome at time t;

2). time lagged effect, that is, previous s exposures up to $t - s$ time points affect the outcome at time t;

3). and cumulative effect, that is, previous t exposures cumulatively affect the outcome at time t.

To disentangle these three different effect mechanisms, we will propose different models and testing strategies in this thesis under the MR framework. This study represents the very first exploration in MR analysis with longitudinal data.

The rest of dissertation is organized as follows: In chapter 2, we propose a concurrent Mendelian Randomization model which assumes current outcome is only affected by current exposure and linear relation holds at every time points. In chapter 3, we extend concurrent model to time lag model and further assume not only instantaneous causal influences exist but also past exposure values can have causal effect on current outcome. We propose an algorithm to select time lags. Pointwise testing and simultaneous testing are also considered to test the existence of causal relationship. We further consider the functional model setting to investigate Mendelian Randomization in chapter 4, followed by conclusions and further work in chapter 5.

CHAPTER 2

CAUSAL INFERENCE WITH TIME-VARYING CONFOUNDING: A MENDELIAN RANDOMIZATION APPROACH

Mendelian Randomization uses genetic variants as instrument variables to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect. The use of Mendelian Randomization reduces regression bias and provides more reliable estimate of the likely underlying causal relationship between an exposure and a disease outcome. Most current Mendelian Randomization methods are focused on cross-sectional phenotypic traits. Longitudinal studies track the same sample at different time points and have a number of advantages over cross-sectional studies. It would be possible for researchers to learn more about 'cause and effect' relationships when incorporating time information. In this work, we propose a two-stage concurrent Mendelian Randomization analysis under the quadratic inference function (QIF) framework. Our proposed method assumes current outcome is affected by current exposure and coefficients of both genetic variants and exposures are time-varying. Through extensive simulation studies, we show that the proposed method has reasonable type I error control. Application to a real data analysis shows that one hormone has a causal effect on women's emotional eating behavior.

2.1 Introduction

Mendelian Randomization (MR) refers to an analytic approach to which genetic epidemiology can assess the causality of an observed association between a modifiable exposure or risk factor and a clinically relevant outcome by using genetic variants [68]. The choice of the genetic instrumental variables is essential to the success of MR analysis. As depicted in Figure 2.1, a valid instrumental variable must satisfy three core assumptions: 1). it must be associated with exposure of interest; 2). it must not be associated with confounders that confound the relationship between the exposure variable and the disease outcome; and 3). it only affects the outcome through exposure variable i.e it indicates the dependence between genetic variant and outcome given exposure and other

observed confounders in the study. The theoretical underpinnings of the Mendelian randomization approach are that: the genotype is robustly associated with the modifiable (non-genetic) exposure of interest (equivalent to assumption 1 above); the genotype is not associated with confounding factors that bias conventional epidemiological associations between modifiable risk factors and outcomes (assumption 2); and that the genotype is related to the outcome only via its association with the modifiable exposure (assumption 3)[52].

There exist several methods available for MR analysis. The Wald method[79], or the ratio of coefficients method is the simplest way of estimating the causal effect of the exposure (X) on the outcome (Y). This method uses summarized data and the causal effect can be estimated through dividing the effect of the IV on the outcome (β_{ZY}) by the effect of the IV on the exposure (β_{ZX}):$\beta_{XY} = \beta_{ZY}/\beta_{ZX}$. Another popular method is two-stage least squares method. In the first-stage regression, the exposure is estimated by calculating the fitted value of the exposure on the IVs, and in the second-stage, the outcome is regressed on the fitted values of the exposure from the first stage.

With a single IV, the 2SLS estimate is the same as the ratio estimate when the outcome is continuous or binary. With multiple IVs, the 2SLS estimator may be viewed as a weighted average of the ratio estimates calculated using the instruments one at the time, where the weights are determined by the relative strengths of the instruments in the first-stage regression[1, 3].

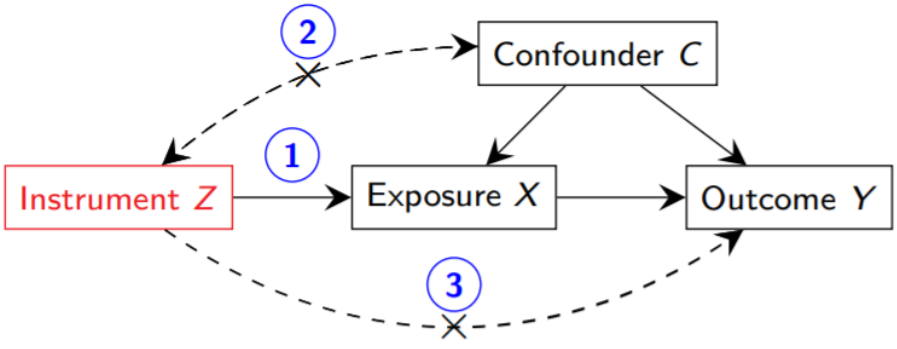


Figure 2.1 Instrument Variable.

However, all the above mentioned methods only use data from a single arbitrary time and assess

cross-sectional causal inference. The existed literature shows single nucleotide polymorphisms (SNPs) may have time-varied effects or the exposure variable may have time-varying effects on the outcome. In the work of Ning et al. [58], the author modeled the time-varied SNP effect for the GWAS analysis based on random regression model and had successfully found some SNPs related with blood pressure for the GWA18 workshop dataset. In these cases, a single measurement is not adequate to capture these time-varying information. In observational studies, researchers often collect longitudinal data, which involves a collection of data at different time points for many study subjects. Since longitudinal data follows changes over time in particular individuals, it would be possible for researchers to learn more about 'cause and effect' relationships. Incorporating this time information can potentially lead to meaningful biological findings. Hogan and Lancaster [42] reviewed and compared two moment-based methods: inverse probability weighting (IPW) and instrumental variables (IV) for estimating causal treatment effects from longitudinal data, where the treatment might vary with time. VanderWeele et al. [78] reviewed some basic principles for causal inference from longitudinal data and discussed the complexities of analysis and interpretation when exposures could vary over time. Newer classes of causal models, including marginal structural models, were considered, which could assess questions of the joint effects of time-varying exposures and could take into account feedback between the exposure and outcome over time.

However, the above mentioned causal models did not consider the time-varying confounding effects. Cao et al. [16] proposed two functional data analysis-based methods to incorporate longitudinal data of a time-varying exposure variable in the MR analysis when the disease outcome was binary. However, instead of selecting valid genetic variants as instrumental variables, they only used genetic variants identified from other research to conduct the first-stage regression. Another limitation of their work is that their proposed new methods are only aimed for hypothesis testing purpose, not for causal effect size estimation.

In this work, we assume the current exposure at time t affects the current outcome at time t. We develop a concurrent causal inference model under a two-stage IV regression framework to assess the causal relationship. The quadratic inference function (QIF) framework proposed by

Qu et al.[63] uses marginal models for the estimation and inference in longitudinal data analysis, and has been popular in longitudinal data analysis. This approach takes into account correlation within subjects and deals directly with both continuous and discrete longitudinal data under the framework of generalized linear models [62]. Qu and Li [62] further extended QIF to varying-coefficient models and proposed a unified and efficient nonparametric hypothesis testing procedure to test whether coefficient functions were time-varying or time invariant. Our concurrent model is built upon the QIF framework, to assess the causal effect between an exposure and an outcome at a particular time point.

The rest of the paper is structured as follows. Section 2.2 describes our time-varying IV methods for Mendelian Randomization. The simulation studies are reported in Section 2.3. In Section 2.4, we apply the model to the eating behavior study. Section 2.5 summarizes the main concludes and discussions.

2.2 The Concurrent Model

In this section, we propose a concurrent model to deal with the situation where confounding factors are time-varying in longitudinal studies. Suppose there are n subjects measured at multiple time points $\{t_j, j = 1, 2, \cdots, T\}$. Let $Y_i(t_j)$ and $X_i(t_j)$ be the time-varying outcome and exposure of subject i recorded at time t_j, respectively. G_i denotes the vector of multiple SNPs of subject i and is time invariant. Denote the data collected as

$$\{Y_i(t_j), X_i(t_j), G_i\}, \quad \text{for } i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, T.$$

If we assume a causal relationship with the order of $G \rightarrow X \rightarrow Y$, i.e., G affects Y only through X, then the following two-stage sequential models could be fitted to dissect the relationship between an exposure and an outcome, i.e.,

$$X(t_j) = \alpha(t_j)G + \epsilon_1(t_j), \tag{2.1}$$

$$Y(t_j) = \beta_0(t_j) + \beta_1(t_j)X(t_j) + \epsilon_2(t_j), \tag{2.2}$$

where $\alpha(t)$ and $\beta(t)$ are coefficient functions; and $\epsilon_1(t)$ and $\epsilon_2(t)$ are model error functions with mean zero. In this model, we assume the effects of genetic variants on exposure and the effects of exposure on outcome both are time-varying. When there are time-varying confounding effects, adjusting the time-varying effect of the genetic effects (i.e., IV effects) can control the time-varying confounding effects, hence leading to the causal inference using the $\hat{X}(t)$ in the 2nd stage regression. The following two sections introduce how to deal with model (2.1) and model (2.2) in details.

2.2.1 Estimating the time-varying SNP effect

In this step, we select IV variables (i.e., SNP variables) and further estimate their time-varying effects on the exposure variable. Genome-wide association studies (GWAS) are providing a rich source of potential instruments for MR analysis. The most common approach for selecting genetic variants for inclusion is LD-pruning. The threshold τ is often taken to be the GWAS significance threshold $\tau = 5 \times 10^{-8}$ in order to reduce the number of false-positive associations arising from the vast number of statistical tests performed. Dudbridge [27] showed using more relaxed threshold might be beneficial. Varying-coefficients models arise naturally when one wishes to examine how regression coefficients change over different groups characterized by certain covariates such as age[32]. Since we assume time-varying coefficients for SNPs and longitudinal exposure measurement is observed, QIF method is a good choice to be applied to estimate and select the IV variables. We first conduct QIF testing for each genetic variant and apply a relax criteria for IV selection. P-values are sorted from the smallest to the largest and 100 most significant SNPs are selected to fit a multiple regression model for further effect estimation and IV selection. Under the assumption that only a few genetic variants are valid instrumental variables with time varying coefficients, we apply some basis functions to approximate the varying coefficients $\boldsymbol{\alpha}(t)$ and insert penalties to choose SNPs with time-varying effects. Any basis system for function approximation could be applied and some popular choices include Fourier basis, polynomial basis, or splines.

We assume that the observations from different subjects are independent, but those within the same subject are correlated. Under the first moment model assumption, the varying-coefficient

models assume the following mean structure:

$$E(X_{it}) = \mu_{it} \text{ and } g(\mu_{it}) = \alpha(t)G_i \tag{2.3}$$

where $g(\cdot)$ is a known link function and α represents a q-dimensional regression coefficients vector.

Suppose we have q genetic instruments in total. For each $l = 1, \cdots, q$, $B_{lv}(t)$ is a set of basis functions of the functional space to which $\alpha_l(\cdot)$ belongs. For simplicity, we assume each $\alpha_l(t)$ has the basis functions $B(u) = (B_1(u), \cdots, B_V(u))$ with the same order M and knots K, where $V = M + K + 1$. Then $\alpha_l(t)$ could be approximated by a linear combination of the basis functions, i.e.

$$\alpha_l(t) \approx \sum_{v=0}^{V} \gamma_{lv} B_v(t), \quad \text{for } l = 0, \cdots, q,$$

where γ_{lv}'s are spline constants and V is associated with the number of basis functions for the coefficient.

Plugging the approximation of $\alpha_l(t)$ into the mean structure, equation (2.3) could be defined as follows:

$$E(x_{it}) = \mu_{it} \text{ and } g(\mu_{it}) = \alpha(t)G_i \approx \sum_{l=0}^{q} \sum_{v=0}^{V_l} \{G_{il} B_{lv}(t)\} \gamma_{lv}. \tag{2.4}$$

Qu and Li [62] considered the q-degree truncated power spline basis with knots $k_1, \cdots, k_{K_l}$, that was

$$1, t, t^q, (t - k_1)_+^q, \cdots, (t - k_{K_l})_+^q,$$

where $z_+^q = z^q I(z \geq 0)$.

The quasi-likelihood equation for longitudinal data is defined as follows:

$$\sum_{i=1}^{n} \dot{\mu}_i' V_i^{-1} (x_i - \mu_i) = 0,$$

where $V_i = var(x_i)$ and is often unknown in practice, $\dot{\mu}_i = \partial \mu_i / \partial \alpha$. If V_i is known, one might use empirical estimator to estimate V_i. However, if the size of V_i is large, there would be many nuisance parameter estimations, and a high risk of numerical error in the inversion of the empirical estimator[63]. Liang and Zeger [54] proposed generalised estimating equations (GEE) method and simplified V_i using $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i was a diagonal marginal variance matrix

and R was a common working correlation. Regardless of whether the working correlation is correctly specified or not, GEE method enables one to estimate regression parameters consistently in longitudinal data analysis. However, the GEE estimator is inefficient when the correlation structure is misspecified. Qu et al.[63] proposed a method of quadratic inference function that did not require more assumptions than does the generalised estimating equation method, but remained optimal even if the working correlation structure was misspecified.

The QIF is derived by observing that the inverse of the working correlation matrix could be approximated by a linear combination of several basis matrices:

$$R^{-1} \approx a_0 I + a_1 M_1 + \cdots a_m M_m,$$

where I is the identity matrix and M_i are symmetric matrices. Plugging expansion into the quasi-likelihood function leads to a linear combination of the elements of the following extended score vector:

$$\bar{g}_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} g_i(\gamma) = \begin{pmatrix} \sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1}(x_i - \mu_i) \\ \sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1/2} M_1 A_i^{-1/2}(x_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1/2} M_m A_i^{-1/2}(x_i - \mu_i) \end{pmatrix},$$

and the quadratic inference with respect to γ is then defined as

$$Q_n(\gamma) = n\bar{g}_n' \bar{C}_n^{-1} \bar{g}_n. \tag{2.5}$$

where $\bar{C}_n = n^{-1}\sum_{i=1}^{n} g_i g_i'$ is the sample covariance matrix.

It is worth noting that the quadratic inference function defined above contains only the regression parameter γ, and only the basis matrices from the working correlation structure are used to formulate this function. There is no need to estimate the nuisance correlation parameter to obtain optimal estimator $\hat{\gamma}$ [73]. When we assume an independent working correlation, or exchangeable correlation for balanced data, the quadratic inference function and the generalised estimating equation have the same estimating functions, resulting in identical estimators[64].

When there exists a high-dimensional regression setup and only a subset of those are important for predicting the response, an overfitted model lowers the efficiency of estimation while an underfitted one leads to a biased estimator. One popular approach is to incorporate some "penalty" to estimate the nonzero parameters and functions simultaneously.

The smoothly clipped absolute deviation (SCAD) penalty is taken into consideration due to its unbiasedness, sparsity, and continuity properties. The derivative of non-convex SCAD penalty is defined as:

$$p'_{\lambda_n}(\theta) = \lambda_n\{I(\theta \leq \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n)\}$$

where $a > 2$, $\theta > 2$ and $p'_{\lambda_n}(0) = 0$. In practice, searching the best pair (a, λ_n) over the two-dimensional grids using some criteria, such as cross-validation and generalized cross-validation is computationally expensive[18]. Fan and Li [29] showed the choice of $a = 3.7$ had a good performance.

To incorporate the within-cluster correlation and select important SNPs, we apply the QIF to estimate γ and exerted group-wise SCAD penalization to equation (2.5) to guarantee that spline coefficient vector of the same nonparametric component is treated as an entire group in model selection. The group-wide penalized quadratic inference function is defined as follow:

$$Q_n^p(\gamma) = Q_n(\gamma) + n\sum_{l=1}^{q} p_\lambda(||\gamma_l||_H) \tag{2.6}$$

where $||\gamma_l||_H = (\gamma_l^T H \gamma_l)^{1/2}$, $H = (h_{ij})_{V\times V}$, $h_{ij} = \int_0^1 B_i(u)B_j^T(u)du$ and p_λ is the SCAD penalty function.

Minimizing the penalized objective function of (2.6), we could get the penalized estimator $\hat{\gamma}$ by

$$\hat{\gamma} = \arg\min Q_n^p(\gamma). \tag{2.7}$$

2.2.2 Estimation and testing of the time-varying exposure effect

In the first step, we obtain an estimate of spline coefficients $\hat{\gamma}$ by minimizing penalized quadratic inference function in (2.6). Then an estimator for $\alpha_l(t)$ is given by $\hat{\alpha}_l(t) = \sum_{v=0}^{V} \hat{\gamma}_{lv} B_v(t)$. The fitted value $\hat{X}(t) = \hat{\alpha}(t)G$ is then used to substitute in the second step.

In the second step, we assume the current response is only affected by the current exposure. Since we also consider a time-varying coefficient in the second step, similar to the first step, we still consider the quadratic inference function for time varying effects. The coefficient $\beta(t)$ can be approximated by a linear combination of the basis functions, i.e.

$$\beta_l(t) \approx \sum_{s=0}^{S} \eta_{ls} B_s(t), \quad \text{for } l = 0, 1$$

where η_s's are spline constants and S is associated with the number of basis functions for the coefficients. The basis functions we use in the second step can be different from the basis functions used for $\alpha(t)$. Minimizing the quadratic inference function with respect to η, we can get the estimator $\hat{\eta}$ by

$$\hat{\eta} = \arg\min Q_n(\eta). \tag{2.8}$$

In Mendelian Randomization, we focus more on testing instead of estimation. For the testing problem, we are interested in testing $H_0 : \beta_1(t) = 0$ versus $H_a : \beta_1(t) \neq 0$. Since the coefficient $\beta_1(t)$ was approximated by a linear combination of the truncated power basis, we could test if time-variant coefficient $\beta_1(t)$ is zero by the following equivalent hypothesis:

$$H_0 : \eta_{1s} = 0, s = 1, 2, \cdots, S \text{ v.s. } H_a\text{: At least one } \eta_s \neq 0$$

The test statistic to test H_0 against H_a is constructed by $T = Q_n(\tilde{\eta}) - Q_n(\hat{\eta})$ which asymptotically follows a chi-squared distribution with S degrees of freedom under the null hypothesis, where $\tilde{\eta}$ denotes the estimator under H_0 and $\hat{\eta}$ be the estimator under H_1.

2.3 Simulation Study

Aim

The aim of the simulation study is to evaluate the performance of the concurrent model. Since MR mainly focuses on hypothesis testing, the ideal model should well protect the type I error rate at the $\alpha = 0.05$ significance level and obtain good empirical power performance. In addition, investigating the properties to influence the type I error or power behavior is also of interest.

The investigated properties include sample size, confounding effects, within sample correlation. Besides, the QIF approach is applied to deal with the concurrent model and QIF requires working correlation assumption. It is also of interest to compare the concurrent model behavior if incorrect working correlation structure is assumed.

Data-generating Mechanisms

Two different sample size settings were considered: $n = 200$ and $n = 400$. We considered 20 repeated measurements for each subject and the time points $t_1, \cdots, t_{20}$ were chosen to be equidistant between 0.1 and 1. In our study, we assumed the effects of genetic variants on exposure and the effect of exposure on outcome both were time-varying.

In the first-stage regression $X(t_j) = \alpha(t_j)G + \epsilon_1(t_j)$, 15 SNPs were generated in total, among them 5 SNPs were simulated as valid instrumental variables with time-varying effects on exposure and the rest SNPs had zero coefficients. For each SNP variable G, the SNP allele frequency (p) was generated from a uniform (0.1, 0.4), then SNP values was sampled from {0, 1, 2} with probability p^2, $2p(1-p)$ and $(1-p)^2$ to obtain homozygous, heterozygous, and other homozygous genotypes, respectively. We defined the true varying coefficients for the intercept and the five SNPs as follows:

$$\begin{aligned}
&\alpha_0(t) = 0.1\cos(2\pi t) + 0.2, && \alpha_3(t) = 0.5\sin(\pi t) + 0.6,\\
&\alpha_1(t) = 2t, && \alpha_4(t) = 0.5\cos(\pi t/2) + 0.6,\\
&\alpha_2(t) = (1-t)^3 + 0.2, && \alpha_5(t) = 0.3\sin(\pi t/3) + 0.5,\\
&\alpha_6(t) = \cdots = \alpha_{15}(t) = 0.
\end{aligned}$$

where $\alpha_0(t)$ was the intercept function. The simulated X values were then applied in the second-stage regression to generate outcome Y.

In the second-stage regression $Y(t_j) = \beta_0(t_j) + \beta(t_j)X(t_j) + \epsilon_2(t_j)$, we set $\beta_0(t) = 0.2t + 0.2$ and $\beta_1(t) = 0$ or $\beta_1(t) = 0.015 + 0.01t$ to investigate the type I error and power, respectively.

To include confounding effect, error terms $\epsilon_1(t)$ and $\epsilon_2(t)$ were generated simultaneously

by assuming a variance-covariance matrix $\mathbf{\Sigma} = cov(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Two scenarios were considered: the first scenario aimed to investigate the effect of confounding factors, where diagonal matrix Σ_{11} and Σ_{22} were fixed and different Σ_{12} and Σ_{21} settings were generated; the second scenario aimed to investigate the effect of within sample correlation, where off-diagonal matrix Σ_{12} and Σ_{21} were fixed and different Σ_{11} and Σ_{22} settings were generated. Under each scenario, two types of variance-covariance structure were considered in total: exchangeable structure and auto-regressive with order 1, i.e., AR(1). The details of the variance-covariance matrix simulation were sated as follows:

1.
   - The entry in Σ_{11} and Σ_{22} was set to be $0.1 \times (0.5)^{|i-j|}$ for $i, j = 1, \cdots, 20$.
   - To generate Σ_{12} and Σ_{21}, the cross-correlation was defined as ρ and two types of correlation structure were considered:
     - If working structure is exchangeable, the off diagonal elements were generated to be $0.1 \times \rho$ and the diagonal part was $0.1 \times (\rho + 0.1)$.
     - If working structure is AR(1), the off-diagonal element was generated to be $0.1 \times \rho^{|i-j|}$, while the diagonal entry was set as $0.1 \times (\rho + 0.1)$.

2.
   - The off-diagonal elements of Σ_{12} and Σ_{21} were generated to be $0.1 \times (0.3)^{|i-j|}$, and the diagonal entry was set as 0.04.
   - To generateΣ_{11} and Σ_{22}, the within sample correlation was defined as δ and two types of working structure were considered:
     - If correlation is exchangeable, the off diagonal elements were generated to be $0.1 \times \delta$ and the diagonal part was $0.1 \times (\delta + 0.1)$.
     - If correlation is AR(1), the entry in Σ_{11} and Σ_{22} was set to be $0.1 \times \delta^{|i-j|}$.

In all the simulations, we assumed the AR(1) working correlation when analyzed the data. Under each setting, the simulation was repeated 1000 times.

Targets

The concurrent model was evaluated for testing the null hypothesis $\beta(t) = 0$. The testing performance was measured by the type I error rate and power.

Analysis Method

In the first stage regression, the QIF method was applied for instrumental variable selection and exposure estimation. In the second stage regression, the QIF testing was applied to test the existence of time-varying coefficient $\beta(t)$.

Performance Measures

The corresponding simulation results were shown in Table 2.1 and Table 2.2. From Table 2.1, the type I error rates were well-controlled at the $\alpha = 0.05$ significance level under different simulation settings. Incorrectly specifying working correlation did not influence the type I error rate. For the empirical power, although using the wrong working correlation for analysis decreased the power, the difference was not significant. The empirical power increased with the increase of the sample size and slightly decreased with the increase of the effects of confounding factors. When the effects of confounding factors was relatively small ($\rho = 0.1$), using correct working structure and analyzing with incorrect working correlation had quite similar empirical power performance.

Table 2.1 Effect of confounding on the type I error and power.

ρ	n	Type I error		Power	
		EXC.	AR-1	EXC.	AR-1
0.1	200	0.052	0.049	0.632	0.675
	400	0.049	0.049	0.930	0.946
0.3	200	0.053	0.052	0.512	0.643
	400	0.050	0.047	0.821	0.936
0.5	200	0.049	0.056	0.507	0.641
	400	0.053	0.046	0.711	0.921

Similar to Table 2.1, type I error rate could still be well protected at the $\alpha = 0.05$ significance

level under all different simulation scenarios in Table 2.2. Changing the within sample correlation and misspecifying the working correlation had no effects on Type I error rate. For the empirical power simulation, misspecifying the working structure still lowered the power and this influence was more significant compared to the situations in table 2.2. With the increase of the sample size, the empirical power also increased. Similar to the results in table 2.1, the lager the effects of within sample correlation was, the smaller empirical power would be. However, the influence of within sample correlation on empirical power was more significant compared to the influence of confounding factors.

Table 2.2 Effect of within sample correlation on the type I error and power.

δ	N	Type I error		Power	
		EXC.	AR-1	EXC.	AR-1
0.3	200	0.048	0.044	0.437	0.824
	400	0.050	0.051	0.754	0.994
0.5	200	0.055	0.054	0.413	0.659
	400	0.051	0.051	0.669	0.939
0.7	200	0.050	0.049	0.371	0.473
	400	0.053	0.052	0.565	0.778

2.4 Case Study: Albert Twin Data

2.4.1 Albert Twin data

The method was applied to the Albert twin data set to investigate the causal relationship between the hormone level and emotional eating behavior in teen girls. The Albert twin data set came from the Twin Study of Hormones and Behavior across the Menstrual Cycle project (TSHMBC) [49] within the Michigan State University Twin Registry (MSUTR; see [14, 48]for MSUTR description). This project is still in progress. The aim of the project is to investigate systematic changes in ovarian hormones (e.g., estrogen and progesterone) and emotional eating behavior across the menstrual cycle in identical and fraternal female twins between the ages of 15-30 years. Emotional eating was measured with the Dutch Eating Behavior Questionnaire (DEBQ) and negative affect was measured with the Negative Affect scale from the Positive and Negative Affect Schedule (PANAS).

The DEBQ assesses the tendency to eat in response to negative emotions while PANAS is used to measure negative emotional states like sadness and anxiety.

All participants were required to meet the following inclusion criteria: 1) menstruation every 22–32 days for the past 6 months; 2) no hormonal contraceptive use within the past 3 months; 3) no psychotropic or steroid medications within the past 4 weeks; 4) no pregnancy or lactation within the past 6 months; and 5) no history of genetic or medical conditions known to influence hormone functioning or appetite/weight [51]. The data dictionary was reported in Appendix A.2.

Table 2.3 Subject characteristics at baseline n=225.

Variables	Summary statistics
Age	Mean (sd): 17.5 (1.7) Median (quantiles): 17.0 (16.4,18.0) Range: 15-26
BMI	Mean (sd): 23.6 (5.5) Median (quantiles): 22.0 (20.2, 24.9) Range: 15-46
Status	MZ: n=124 (55%) DZ: n=101 (45%)
Estradiol	Mean (sd): 2.4 (1.4) Median (quantiles): 2.1 (1.6, 2.8) Range: 0-14
Progesterone	Mean (sd): 80.7 (58.0) Median (quantiles): 64.3 (40.7, 100.0) Range: 10-324
PANAS	Mean (sd): 15.0 (4.7) Median (quantiles): 13.8 (11.8, 17.0) Range: 10-39
DEBQE	Mean (sd): 1.4 (0.5) Median (quantiles): 1.2 (1.0, 1.5) Range: 0-4

Measurements for each participant were collected for 45 consecutive days within one menstrual cycle, which were then grouped into eight menstrual cycle phases, that is, ovulatory phase (1), transition ovulatory to midluteal (2), midluteal phase (3), transition midluteal to premenstrual (4), premenstrual phase including the first day of menstrual cycle (5), remaining days of menstrual cycle, part of follicular phase (6), follicular phase (7) and transition follicular to ovulatory phase (8). They

were grouped into these phases based on profiles of changes in progesterone across the cycle (by ConsensusFIN2). Since each phase contained more than 2 consecutive days measurements, we took two averaged measurements within each phase, which ended up with a total 16 data points for each individual for further analysis [80].

We started with 167,509 SNPs. After removing SNPs with a genotyping call rate of less than 90% or a minor allele frequency of less than 5%, there were 166,063 SNPs remained for further analysis. No subject was removed in this step. Since MZ twins shared all of their genetic variants, the missing SNPs for MZ twins were replaced with the SNPs in another twin pair. Imputation of the rest missing SNPs was based on the Wright equilibrium and had the following form:

$$P(G_{ij}) = \begin{cases} P(G=0) = (1-p_j)^2 + p_j(1-p_j)F_i \\ P(G=1) = 2p_j(1-p_j) - 2p_j(1-p_j) \\ P(G=2) = p_j^2 + p_j(1-p_j)F_i \end{cases} \tag{2.9}$$

where p_j was the frequency of the major allele for an SNP j, and F_i was the level of homozygosity of an individual i, estimated as a proportion of the amount of homozygous loci relative to the total of loci. Missing SNPs were then sampled from $\{0, 1, 2\}$ by $P(G_{ij})$.

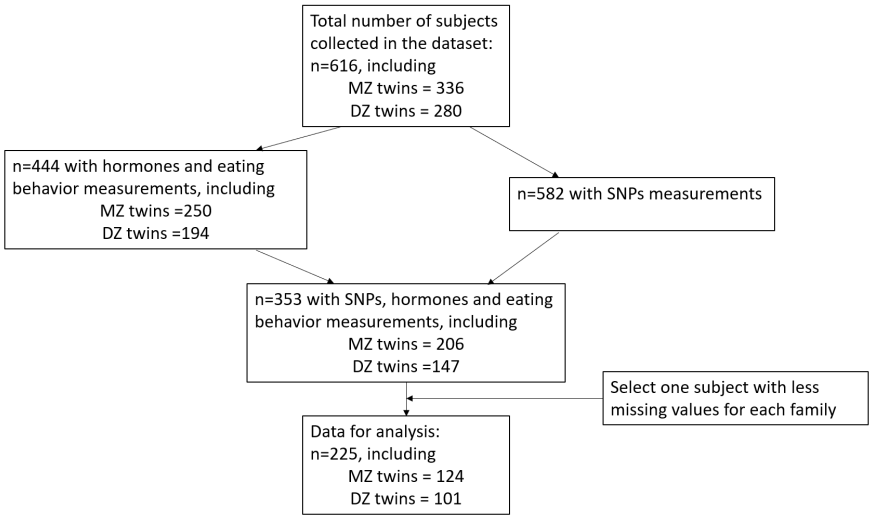


Figure 2.2 Flow diagram of subject selection.

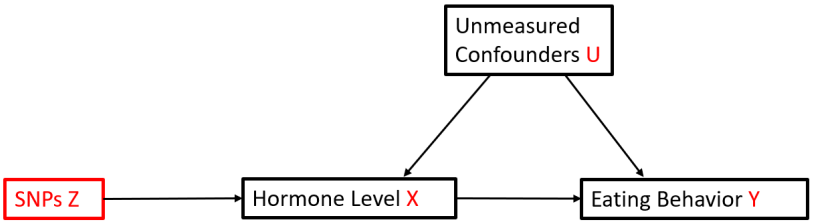


Figure 2.3 Diagram of the causal relationship for four models (combinations of hormone levels and eating behavior).

Albert twin data set had 616 subjects in total, among them 336 subjects were monozygotic(MZ) twins and 280 subjects were dizygotic(DZ) twins. DZ twins might have different genetic variants information, and MZ twins share the same genetic variants information. The measurements included for analysis were genetic variants, hormone levels and emotional eating behavior observations. There were two files. File 1 has longitudinal hormone level and eating behavior measurements with 444 individuals. File 2 contains SNP information with 582 individuals. Out of 444 individuals in file 1, 353 individuals are contained in file 2. Out of 582 individuals in file 2, 353 individuals are contained in file 1. After merging the two files with common IDs, there are 353 left (containing SNPs, hormones and eating behavior measurements). These 353 subjects belong to 225 unique families. Some families contain two twins, and some only contain one. For each family, we only picked one subject for further analysis to meet the sample independence assumption. If a family only had one subject, then that subject was picked. If family had two subjects, we compared their time period (by ConsensusFIN2), and subject with less missing values based on the two measurements (i.e., hormone and eating behavior) was chosen. Finally, 225 subjects from different families were chosen, including 124 monozygotic subjects and 101 dizygotic subjects. Figure 2.2 described the details of subjects selection steps.

In this project, the exposures were the two hormone levels (estradiol and progesterone) and

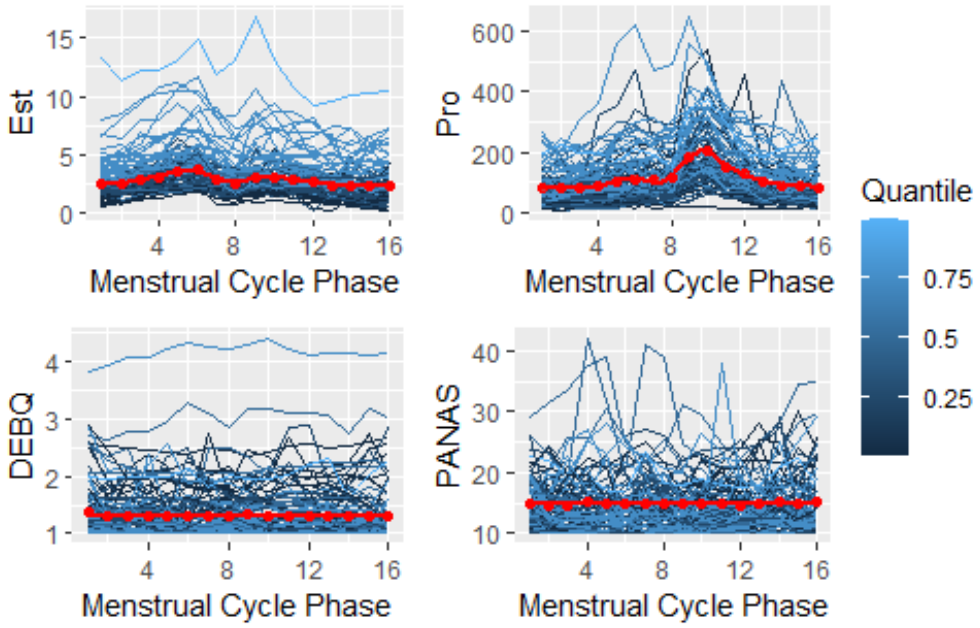


Figure 2.4 The observed hormones levels and eating behaviour measurements for the first 100 subjects.

the outcome was the two emotional eating behaviors (DEBQ and PANAS), both measured at 16 “time" points. The genetic variants were measured with $\{0, 1, 2\}$ to represent the number of minor allele. They were time-invariant and were treated as instrumental variables to control the effects of confounding factors when investigating the causal relationship between exposure and outcome. The goal was to evaluate if there exists causal relationship between hormone levels and eating behaviors and if so, what were the effect mechanisms. The characteristics of subjects used for analysis in the Albert twin data set was reported in Table 2.3. Borrowed the idea of Figure 2.1, Figure 2.3 was used to describe the diagram of the casual relationship that would be investigated.

Figure 2.4 showed the longitudinal Est, Pro, DEBQ and PANAS data for the first 100 subjects in the Albert twin data set with the red line representing the average measurements. The individual estradiol level was almost flat during the whole menstrual cycle phases. For progesterone level, most subjects reached their maximum value when the menstrual cycle phase is 10. The individual progesterone level was approximately flat before phase 7. For the individual emotional eating measurements DEBQ and PANAS, the individual fluctuation patterns were very different. While some subjects had relatively stable trajectories over phases, others had substantial changes,

(a) Distribution of Age

(b) Distribution of BMI at baseline

Figure 2.5 Distribution of Age and BMI at baseline.

hence realizing that the time-varying information might not be sufficiently captured by a single observation.

The average estradiol had bimodal distribution with the first peak at 6 (midluteal phase) and the second peak at 10 (premenstrual phase including the first day of menstrual cycle). Significant increase was observed between 2 and 6 followed by a rapid decrease between 6 and 8. After the second peak, average estradiol values decreased continuously.

The average maximum progesterone level reached at phase 10, which was the premenstrual phase and the distribution was a unimodal distribution with one clear peak. The average progesterone increased slowly between phase 1 and phase 8, and then displayed a rapid growth after phase 8. After reaching the peak at phase 10, it continuously went down. For the majority of the phases, the average progesterone had measurements less than 100.

We could not observe obvious trends for average DEBQ and PANAS measurements, and both curves oscillate over time. For DEBQ, it had the maximum value at the beginning of the menstrual cycle phases and then dramatically decreased. Three peaks existed, which were at phase 1, 3 and 9. After phase 12, the curve became flat. The trend was more complex for PANAS. It also had significant drop between phase 1 and phase 2. Moreover, it also had evident growth ranging from phase 2 to phase 4, and from phase 12 to phase 16. The maximum value of the average PANAS appeared at the last phase. The average PANAS value kept going up after phase 12.

The distribution of covariates (age and BMI) were plotted in Figure 2.5. Both variables had

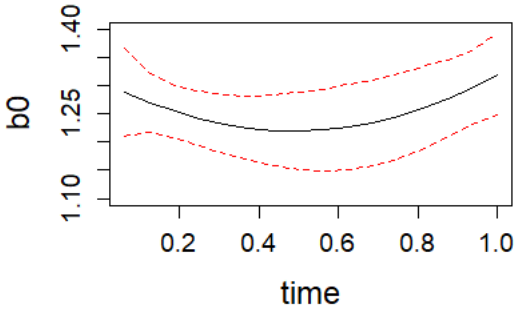
unimodal distribution and skewed to the right. The majority of the subjects were between 16-18 with BMI measurements between 15-30.

2.4.2 Concurrent Model Application

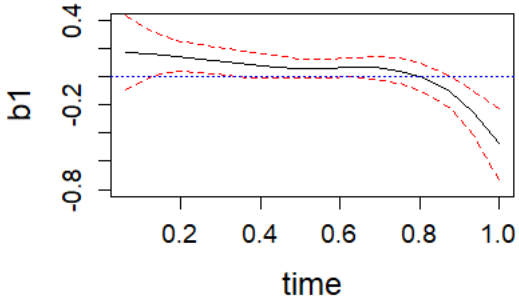
We applied the concurrent model to the Albert twin data set to evaluate if there exists causal relationship between hormone levels and eating behaviors. Time was divided by the maximum of time, which was 16, to make sure time interval was between 0 and 1. We first conducted marginal QIF testing for each SNPs and picked the top 100 SNPs which were then used for variable selection in the first step regression. The spline order and knots were set as 3 and 1, respectively. Then pQIF with grouped SCAD penalty was applied to estimate hormone value. The predicted hormone was used in the second step as input variable. In the second step, we fixed the order as 3, and let knots went from 1 to 5, then used BIC to pick optimal knots. P-value was computed by QIF after choosing the order and knots.

The DEBQ assessed the tendency to eat in response to negative emotions while PANAS was used to measure negative emotional states like sadness and anxiety. In this study, we wanted to examine how hormone (estrogen and progesterone) changes affect emotional eating measured by DEBQ and PANAS. The p-values of estrogen on both DEBQ and PANAS were 0.1569 and 0.1078, respectively, showing no significant causal effect. The p-value of progesterone on DEBQ was 0.00008, and on PANAS was 0.1944, indicating a causal relationship between progesterone and DEBQ. No causal relationship was founded between progesterone and PANAS.

We plotted the point-wise estimator of the coefficients together with their corresponding point-wise confidence interval. The plots of the coefficients for DEBQ and PANAS were shown in Figure 2.6 and Figure 2.7, respectively. The intercept $\beta_0(t)$ was always positive no matter the response variable was DEBQ or PANAS. Moreover, $\beta_0(t)$ increased when time increased for PANAS and first decreased then increased for DEBQ. The confidence interval of $\beta_0(t)$ was wider than that of $\beta_1(t)$. For both DEBQ and PANAS, the effects of PRO decreased over time, indicating a negative causal relationship between PRO and DEBQ over time (i.e., menstrual cycle).

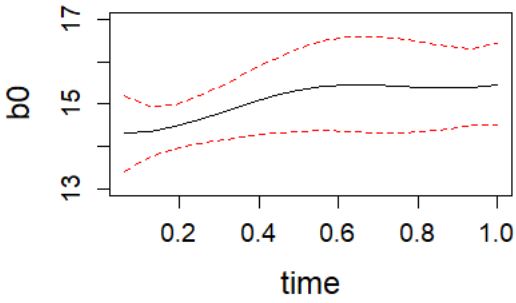


(a) Point-wise estimator of $\beta_0(t)$

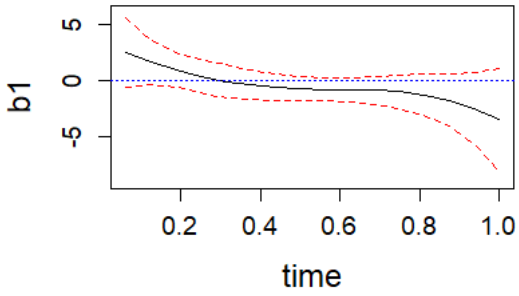


(b) Point-wise estimator of $\beta_1(t)$

Figure 2.6 Coefficients estimation for the relationship between PRO and DEBQ.



(a) Point-wise estimator of $\beta_0(t)$



(b) Point-wise estimator of $\beta_1(t)$

Figure 2.7 Coefficients estimation for the relationship between PRO and PANAS.

2.5 Conclusion and Discussion

Mendelian Randomization uses genetic variants as instruments in observational studies. There are several methods available for instrumental variable estimation using one time measurement. Building upon the two-stage method, we proposed a new method considering longitudinal information and time-varying effects for both instruments and exposures. The proposed concurrent model assumed current response is only affected by current exposure. We applied the idea of QIF regressions in a two-stage instrumental variable regression. In the first step, we used penalized QIF for instrumental variables selection and obtained the fitted exposure over time. The estimated exposure

was applied in the second-step QIF testing to test the causal relationship between the exposure and response. We demonstrated our proposed method in simulation studies. The simulation results suggested that our method could correctly control the type I error rate at the $\alpha = 0.05$ level and detected causal effect when it changed over time. In an application to a Albert twin data set, the analysis results showed progesterone level had causal relationship with eating behavior measured by DEBQ but did not have causal effect on PANAS.

For the instrumental variables selection, we used QIF method in the two-stage regression. However, in traditional Mendelian Randomization, the most popular method is cis-MR, and LD-pruning is the most common approach for selecting genetic variants for inclusion into a cis-MR study. In our study, since we assumed the time-varying coefficient of genetic variants on exposure, the threshold of LD-pruning was not appropriate in this situation. This was the reason why we used QIF approach. Further investigation is needed to check the validity of the QIF approach for instruments selection.

It should be noted that specific assumptions need to be assumed for this particular nonparametric concurrent model and we describe these assumptions as follows:

(a) Past outcomes do not directly affect current outcome.

(b) Past outcomes do not directly affect current exposure.

(c) Past exposures do not directly affect current outcome.

If one of these assumptions is violated, our proposed method might be invalid. This motivated the work in the next two chapters.

Moreover, the reliability of a MR investigation depends on the validity of the genetic variants as instrumental variables. Mendelian randomization studies are known to be affected by both weak instrument bias and the pleiotropic bias that arises when some genetic variants are invalid instrument variables. Sensitivity analysis is usually conducted to test the existence of weak instruments or pleiotropy. In the simulation, we assumed all valid instruments had strong effects and satisfied all

assumptions. We did not investigate situations in the presence of weak instrument variables and pleiotropy. More simulations need to be done for future studies.

In addition, the concurrent model we constructed did not include other covariates effects due to the simplicity consideration. How to adjust for other covariates effects is also of great importance in the model construction. We can simply incorporate the covariates into the two-stage regressions. However, there exist many situations in reality: the effects of covariates might be time-invariant, time-varying or both. The interaction effects could also be included in the model. Further investigations for the more complex cases are our future work.

CHAPTER 3

MENDELIAN RANDOMIZATION WITH TIME LAG EFFECT

In this chapter, we considered Mendelian Randomization analysis with delayed effects. Previously, we proposed a concurrent model which assumed that the current response is only affected by the current exposure. However, there may exist a delay effect for the exposure to have a causal effect on a disease outcome, due to complicated biological processes. For example, the phenomenon of delayed effect is often observed in the emerging and important field of immuno-oncology. In this chapter, we proposed a time lag model to investigate the delayed effects. We assumed that both the current exposure and the past values of exposure contributed together to the current outcome. In order to select the duration of delay included in the model, an algorithm was developed for the variable selection purpose. The point-wise testing and simultaneous testing were developed to test the existence of causal effects. The method was illustrated in the simulation studies and real data analysis.

3.1 Introduction

MR analysis is a method to analyze the causal effect of an environmental exposure variable on an outcome variable from observational studies by using genetic variants as instrumental variables. There exist many well-developed methods for Mendelian Randomization using instrument variable estimators, such as ratio method, two-stage methods, likelihood based methods, and semi-parametric methods. However, all above mentioned methods use cross-functional data, while in reality, many exposures of interest are time-varying, for example, BMI. Inferring causal effects from longitudinal repeated measures data has high relevance to a number of areas of research, including economics, social sciences and epidemiology. Current MR studies only use a single measurement of a time-varying exposure variable given that longitudinal measurements have been collected in many cohort studies. One measurement cannot adequately capture information of a time-varying exposure variable.

In the previous chapter, we proposed a concurrent model which assumed current outcome was only affected by current exposure and the effects of genetic variants on exposure and the effects of exposure on outcome both were time-varying. However, the effect of a specific exposure event sometimes is not limited to the period when it is observed, but it might delay in time. This introduces the problem of modelling the relationship between an exposure occurrence and a sequence of future outcomes, specifying the distribution of the effects at different times after the event (defined lags). In the previous chapter, the proposed concurrent model failed to allow adequately for time lags. In reality, it is necessary to take account of time lags for causal models because it takes time for a cause to exert effect. Organisms sometimes do not respond instantaneously to a change in the system. For example, proteins considered as predictors may have long half-lives.

A lot of published literature has been focused on taking advantages of time lagged variables for causal inference analysis. The general problem posed by the use of lagged variables as regressors using directed acyclic graphs was discussed by Pearl[60]. Reed[66] studied the use of lagged explanatory variables for causal inference in economics and focused on simultaneity and proposed the use of lagged explanatory variables as instruments for endogenous explanatory variables. Bellemare et al.[4] derived analytical results for the biases of lag identification in a common parametric setting: an ordinary least squares (OLS) regression and described the trade-offs between ignoring endogeneity and lagging explanatory variable. Du et al.[26] developed a probabilistic decomposed slab-and-spike (DSS) model to learn the causal relations as well as the lag among different time series simultaneously from data.

In this chapter, we assumed not only current but also recent past levels of the predictor process might play a role in predicting a response. Ultimately, this step required the definition of the additional lag dimension of an exposure–response relationship, describing the time structure of the effect. We proposed a time lag model to investigate delayed effects and we also constructed an algorithm to select the time lag $\triangle t$ which could be determined by data. The rest of chapter is organized as following. In section 3.2, we introduced the instrument variables model and the two-step estimation method. Algorithm to select window width and lag was also proposed in section3.2.

Section 3.3 included two hypothesis testing procedures: point-wise testing and simultaneous testing to test the existence of time-varying casual effect. Simulation studies and real data analysis were given in section 3.4 and section 3.5, respectively, followed by conclusion and discussion in section 3.6.

3.2 Time Lag Model

Suppose there are n subjects and for each individual, i, the exposure and the outcome are measured at multiple time points $\{t_j, j = 1, 2, \cdots, T\}$. Let $Y_i(t_j)$ and $X_i(t_j)$ be the time-varying outcome and exposure of subject i recorded at time t_j respectively. G_i denotes the vector of multiple SNPs of subject i and is time invariant. The data collected are

$$\{Y_i(t_j), X_i(t_j), G_i\}, \quad \text{for } i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, T.$$

The concurrent model assumed current outcome was only affected by current exposure. Although genetic variants are time-invariant, since we assume the effects of genetic variants changed over time, exposure might change over time as well. In this model, we assume the effects of genetic variants on exposure and the effects of exposure on outcome both are time-varying. The model could be formulated as follows:

$$X(t_j) = \alpha(t_j)G + \epsilon_1(t_j), \tag{3.1}$$

$$Y(t_j) = \beta_0(t_j) + \beta_1(t_j)X(t_j) + \epsilon_2(t_j), \tag{3.2}$$

where $\alpha(t)$ and $\beta(t)$ are coefficient functions; and $\epsilon_1(t)$ and $\epsilon_2(t)$ are model error functions with mean zero.

Model (3.2) assumed that response $Y(t_j)$ at current time t_j were only affect by current predictor value $X(t_i)$. In addition to current values, past predictor values might play a role in predicting a response. Assuming not only current but also recent past exposures affect the current response at time t, then model (3.2) becomes

$$Y(t_j) = \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j)X(t_{j-q-(r-1)}) + \epsilon_2(t_j) \tag{3.3}$$

Here p denotes the total number of past time points that is considered to affect the response at the current time. q is a time lag of size, and $(j-q)$ is the first time point that is included in the model. Starting from time $(j-q)$ forward, we continuously insert p time points in total. When $q=0$ and $p=1$, model (3.3) degenerates to the concurrent model (3.2).

When unobservable latent variables affect both $\mathbf{X}$ and Y, $\mathbf{G}$ can be considered as instrumental variables and the effects of $\mathbf{X}$ on Y can be estimated using $\mathbf{G}$. In the first stage, a penalized variable selection algorithm is applied to each gene expression, the same as we described in the concurrent model. Then $\mathbf{X}$ is replaced by the fitted values $\hat{\mathbf{X}}$ in the second stage and the model of interest becomes

$$\begin{aligned} X(t_j) &= \alpha(t_j)G + \epsilon_1(t_j), \\ Y(t_j) &= \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j)\hat{X}(t_{j-q-(r-1)}) + \epsilon_2(t_j). \end{aligned} \tag{3.4}$$

3.2.1 Estimation of the time-varying SNP effect

Following the idea of chapter 2, we assume varying-coefficients for SNPs and exposure. Here, we use the idea of variable selection and add penalties to select significant genetic variants. We use basis functions to approximate the varying coefficient. The choice of basis functions is flexible, and popular choices include Fourier basis, polynomial basis, or splines. Under the assumption that only a few genetic variants are valid instrumental variables with time varying coefficients, we use some basis functions to approximate varying coefficients $\boldsymbol{\alpha}(t)$.

To solve the first equation in (3.4), we do similar operations as we did in chapter 2. As we discussed in chapter 2, we used penalized quadratic inference function with group-wised SCAD penalty to select causal genetic variants. Suppose we have q genetic variants in total, among which only a small number of genetic variants have causal effects on exposure X. Since we assume time-varying effects of SNPS, the first moment assumption has the following form:

$$E(X_{it}) = \mu_{it} \text{ and } g(\mu_{it}) = \alpha(t)G_i \tag{3.5}$$

where $g(\cdot)$ is a known link function and α represents a q-dimensional regression coefficients vector.

Similar to Chapter 2, we use basis functions to approximate the varying coefficients $\alpha(t)$. Suppose $B_{lv}(t)$ is a set of basis functions of the functional space to which $\alpha_l(\cdot)$ belongs, for each $l = 0, \cdots, q$. For simplicity, we assume each $\alpha_l(t)$ has the basis functions $B(t) = (B_1(t), \cdots, B_V(t))$ with the same order M and knots K, where $V = M + K + 1$. Then $\alpha_l(t)$ could be approximated by a linear combination of the basis functions, i.e.

$$\alpha_l(t) \approx \sum_{v=0}^{V} \gamma_{lv} B_v(t), \quad \text{for } l = 0, \cdots, q,$$

where γ_{lv}'s are spline constants and V is associated with the number of basis functions for the coefficient. The second equation in (3.5) then has the following form:

$$g(\mu_{it}) = \alpha(t) G_i \approx \sum_{l=0}^{q} \sum_{v=0}^{V} \{G_{il} B_{lv}(t)\} \gamma_{lv}. \tag{3.6}$$

The quasi-likelihood equation for longitudinal data is

$$\sum_{i=1}^{n} \dot{\mu}_i' V_i^{-1} (x_i - \mu_i) = 0,$$

where $\mu_i = (\mu_{i1}, \cdots, \mu_{iT_i})$, $x_i = (x_{i1}, \cdots, x_{iT_i})$, $\dot{\mu}_i = \partial \mu_i / \partial \gamma$, and $V_i = var(x_i)$ and is often unknown in practice. To incorporate the within-cluster correlation, we apply the QIF to estimate γ and exert group-wise penalization to ensure that the spline coefficient vector of the same non-parametric component is treated as an entire group in model selection. V_i could be decomposed by $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal marginal variance matrix and R is a common working correlation. Instead of specifying the working correlation, series of basis matrices are utilized to estimate working correlation R given as follows:

$$R^{-1} \approx a_0 I + a_1 M_1 + \cdots a_m M_m,$$

where I is the identity matrix and M_i are symmetric matrices.

These basis matrices are further used to define the extended score vector as follows:

$$\bar{g}_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} g_i(\gamma) = \begin{pmatrix} \sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1} (x_i - \mu_i) \\ \sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1/2} M_1 A_i^{-1/2} (x_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1/2} M_m A_i^{-1/2} (x_i - \mu_i) \end{pmatrix}.$$

The quadratic inference with respect to γ is then defined as

$$Q_n(\gamma) = n\bar{g}_n'\bar{C}_n^{-1}\bar{g}_n. \tag{3.7}$$

where $\bar{C}_n = n^{-1}\sum_{i=1}^{n} g_i g_i'$ is the sample covariance matrix.

To select important SNPs, we adopt group-wised SCAD penalty to equation (3.7) to guarantee that spline coefficient vector of the same nonparametric component is treated as an entire group in model selection. The group-wide penalized quadratic inference function is defined as follow:

$$Q_n^p(\gamma) = Q_n(\gamma) + n\sum_{l=1}^{q} p_\lambda(||\gamma_l||_H) \tag{3.8}$$

where $||\gamma_l||_H = (\gamma_l^T H \gamma_l)^{1/2}$, $H = (h_{ij})_{V\times V}$, $h_{ij} = \int_0^1 B_i(u)B_j^T(u)du$ and p_λ is the SCAD penalty function, the derivative of which is defined as:

$$p'_{\lambda_n}(\theta) = \lambda_n\{I(\theta \le \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n)\}$$

where $a > 2$, $\theta > 2$ and $p'_{\lambda_n}(0) = 0$.

Minimizing the penalized objective function of (3.8), we could get the penalized estimator $\hat{\gamma}$ by

$$\hat{\gamma} = \arg\min Q_n^p(\gamma). \tag{3.9}$$

3.2.2 Estimation and testing of the time-varying exposure effect

After obtaining the estimate of spline coefficients $\hat{\gamma}$ by minimizing the penalized QIF in (3.8), an estimator for $\alpha_l(t)$ can be given by $\hat{\alpha}_l(t) = \sum_{v=0}^{V}\hat{\gamma}_{lv}B_v(t)$. The fitted value $\hat{\boldsymbol{X}}(t) = \hat{\boldsymbol{\alpha}}(t)\boldsymbol{G}$ is then used to substitute in the second step for the time-varying exposure effect estimation.

We use the idea of two-step estimation method proposed by Şentürk and Müller [69] for the varying-coefficient model estimation in the second stage regression that considers delayed time lag effect. In the first step, the calculation is focused on particular time points. This gives the point-wise estimation which are then used in the second step to smooth the estimators. In the second step, the raw estimators are smoothed over all time points to improve the efficiency of the estimators.

We collect estimated predictors from first-stage regression and observed response into matrix form. Let $\hat{X}_{qpj} = (\hat{X}_{1,q,p,j}, \cdots, \hat{X}_{n_j,q,p,j})^T$ and $Y_j = (y_{1j}, \cdots, y_{n_j j})$, where $\hat{X}_{i,q,p,j} = \{1, \hat{x}_i(t_{j-q}), \cdots, \hat{x}_i(t_{j-q-p+1})\}^T$ and $\hat{x}_i(t_{j-q})$ is the predicted exposure value for subject i estimated at time t_{j-q}, p denotes the total number of time-points included in the model, i.e. the window width into the past, of the predictor process that is considered to affect the response at the current time and q is a time lag included to predict future values of response. Here n_j denotes the number of subjects observed at time t_j and $(t_{j-q}, \cdots, t_{j-q-p+1})$, and C_j denotes the set of corresponding subject indices. Then for each time t_j, the estimator $\beta(t_j)$ has the following form:

$$b_{pq}(t_j) = (b_{0j}, b_{1j}, \cdots, b_{pj})^T = (\hat{X}_{qpj-p}^T M_{j-p,j} \hat{X}_{qpj})^{-1} \hat{X}_{qpj-p}^T M_{j-p,j} Y_j \tag{3.10}$$

for $j = q + 2p, \cdots, T$. $M_{j-p,j}$ is an $n_{j-p} \times n_j$ matrix for which (a, b)th entry equals 1 if the ath entry of Y_{j-p} and the bth entry of Y_j comes from the same subject, and equals 0 otherwise.

Equation (3.10) gives the point-wise coefficient estimator. In the second step, the raw estimators from first step are smoothed over time. For the rth coefficient $\hat{\beta}_{rqp}(t)$, the following equation is used to smooth the function:

$$\hat{\beta}_{rqp}(t) = \sum_{j=1}^{T} w(t_j, t) b_{rj} \tag{3.11}$$

where $w(t_j, j)$ is smoothing weights and could be constructed by various smoothing techniques, such as local polynomial smoothing, spline smoothing or kernel smoothing. Fan and Zhang [31] considered local polynomial setting. They defined $D_j = (1, t_j - t, \cdots, (t_j - t)^p)^T$, $j = 1, 2, \cdots, T$ and $K_h(t) = K(t/h)/h$ be a kernel function with a bandwidth h. Then smoothing weights for the qth derivative of an underlying function $w_{q,p+1}(t_j, j) = q! e_{q+1,p+1}^T (D^T W D)^{-1} D_j W_j$, where $D = (D_1, D_2, \cdots, D_T)^T$ and $W = (W_1, \cdots, W_T)$ with $W_j = K_h(t_j - t)$.

3.2.3 Time lag selection

Selecting appropriate lag for the model is important, since too many lags inflate the standard errors of coefficient estimates and thus imply an increase in the mean-square forecast errors while omitting lags that should be included in the model may generate auto-correlated errors and result

in an estimation bias. Lag length is frequently selected using an explicit statistical criterion such as minimizing the Bayes information criterion (BIC) or the Akaike information criterion (AIC). In our study, we apply the idea backward stepwise deletion technique proposed by Fan et al. [30] to determine window width p and lag q simultaneously. An initial group of predictors $\{x(t_{j-q}), \cdots, x(t_{j-q-p+1})\}$ for predicting the response at time t_j are included at the beginning. Starting with the smallest and largest time lags $x(t_{j-q}), x(t_{j-q-p+1})$, performance of two groups of predictors without $x(t_{j-q})$ and $x(t_{j-q-p+1})$ respectively are compared to identify the least significant predictor among the two candidates, and that least significant predictor would be deleted from the group. Group $\{x(t_{j-q-1}), \cdots, x(t_{j-q-p+1})\}$ and group $\{x(t_{j-q}), \cdots, x(t_{j-q-p+2})\}$ are considered as reduced model and compared with initial full model respectively using F-statistics, which are calculated at time t_j by the following equation:

$$F_{rqp} = \frac{\{RSS_{qpj}(R) - RSS_{qpj}(F)\}/1}{RSS_{qpj}(F)/(n_{j,j-p} - p)},$$

where RSS stands for the residual sum of squares of the fitted model at time t_j, and is defined as follows:

$$RSS_{qpj} = \sum_{i=1}^{n_{j,j-p}} \{y(t_{ij}) - b_{0j} - \sum_{r=1}^{p} b_{rj} x_i(t_{j-q-(r-1)})\}^2$$

Here, p is the number of predictors considered. The group with smaller F-statistic is then selected as reduced model. Suppose $\{x(t_{j-q-1}), \cdots, x(t_{j-q-p+1})\}$ have smaller F-statistic, and AIC is applied to determine whether $x(t_{j-q})$ is finally deleted from initial set of considered predictors. AIC is defined as $AIC = \log\{RSS/(n_{j,j-p} - p)\} + 2p/n_{j,j-p}$. If the AIC of the reduced model is smaller than that of the full model, we then finally delete $x(t_{j-q})$ to get a new group of predictors and treat= $\{x(t_{j-q-1}), \cdots, x(t_{j-q-p+1})\}$ as initial group. This backward stepwise deletion is repeated until we could not delete any further predictors. However, one problem of this problem is that selection performance relies on the choice of initial groups of predictors. To solve this shortcoming, we try different initial settings, and for each setting, we could select corresponding p and q values. The most often selected p and q values are treated as window width and lag of the final model.

3.3 Model Testing

In this section, we propose two hypothesis testing procedures to test the existence of time-varying coefficient $\beta(t)$. The pointwise testing is introduced in section 3.3.1 and the simultaneous testing is discussed in section 3.3.2.

3.3.1 Pointwise Testing

In this section, we consider the pointwise testing for the exposure coefficient $\beta(t)$. In section 3.2.2, equation (3.10) gives the pointwise coefficient estimator. From Şentürk and Müller [69], this estimator follows the asymptotic Gaussian distribution under some conditions and is stated in the following theorem:

Theorem 3.3.1 (**Asymptotic property of Pointwise Estimator**). *Assuming conditions $A1 - A3$ hold, we have,*

$$\sqrt{n_{j-p,j}}\{b(t_j) - \beta(t_j) \to N(0_{p+1}, \chi_j^{-1}\Sigma_j\chi_j^{-1})\}$$

in distribution as $n_{j-p,j} \to \infty$ for all time-points t_j such that $j = q + 2p, \cdots, T$. Here χ_j and Σ_j are definded as follows:

$$\chi_j = E(n_{j-p,j}^{-1} X_{qpj-p}^T M_{j-p,j} X_{qpj})$$

$$(\Sigma_j)_{s,s'} = \begin{cases} E\{x(t_{j-q-p-s+2})x(t_{j-q-p-s'+2})\}\eta_{qpj}, & 2 \le s, s' \le p+1, \\ E\{x(t_{j-q-p-s'+2})\}\eta_{qpj}, & s = 1, 2 \le s' \le p+1, \\ E\{x(t_{j-q-p-s+2})\}\eta_{qpj}, & s' = 1, 2 \le s \le p+1, \\ \eta_{qpj}, & s = s' = 1, \end{cases}$$

where $\eta_{qpj} = \delta_j + \sigma_y^2$, and σ_y^2 is the variance of the zero-mean additive measurement error of response Y_j and δ_j is the variance function of the zero-mean stochastic process $\epsilon_2(t)$.

To test the null hypothesis that the exposure has no effect on the outcome at time t_j, i.e., $H_0 : b(t_j) = 0$, a Wald test is applied for the pointwise testing.

Following Şentürk and Müller [69], $cov(b_{rj}, b_{rj'})$ can be calculated by the standard least squares theory as,

$$cov(b_{rj}, b_{rj'}) = \begin{cases} \delta(t_j, t_{j'}) c_{r,p}^T (X_{qpj-p}^T M_{j-p,j} X_{qpj})^{-1} X_{qpj-p}^T M_{j-p,j} \\ \otimes M_{j,j'} M_{j',j'-p} X_{qpj'-p} (X_{qpj'-p}^T M_{j'-p,j'} X_{qpj'})^{-1} c_{rp}, \quad j \neq j', \\ (\delta_j + \sigma_y^2) c_{r,p}^T (X_{qpj-p}^T M_{j-p,j} X_{qpj})^{-1} X_{qpj-p}^T \\ \otimes M_{j-p,j} M_{j,j-p} X_{qpj-p} (X_{qpj-p}^T M_{j-p,j} X_{qpj})^{-1} c_{rp}, \qquad j = j', \end{cases} \tag{3.12}$$

where c_{rp} denotes a p-dimensional unit vector with 1 at its rth entry.

Once we obtain estimators for $\delta(t_j, t_{j'})$ and $\delta_j + \sigma_y^2$, the estimator of $cov(b_{rj}, b_{rj'})$ could be calculated by equation (3.12). We define $P_{qpj} = X_{qpj}(X_{qpj-p}^T M_{j-p,j} X_{qpj})^{-1} X_{qpj-p}^T M_{j-p,j}$, and $\hat{e}_{qpj} = (I_{n_j} - P_{qpj}) Y_j$ be the residuals at time t_j. If we assume that $tr\{(I_j - P_{qpj}) M_{j,j'} (I_{j'} - P_{qpj'})\} \neq 0$ and $n_j > p$, then $\delta(t_j, t_{j'})$ and $\delta_j + \sigma_y^2$ could be estimated by the following equations:

$$\hat{\delta}(t_j, t_{j'}) = tr(\hat{e}_{qpj} \hat{e}_{qpj'}^T) / tr\{(I_j - P_{qpj}) M_{j,j'} (I_{j'} - P_{qpj'})^T\} \tag{3.13}$$

$$\hat{\triangle}_j = \hat{e}_{qpj}^T \hat{e}_{qpj} / (n_j - p) \tag{3.14}$$

where $\triangle_j = \delta_j + \sigma_y^2$.

Plugging in respective estimators of $cov(b_{rj}, b_{rj'})$, we are able to compute the variance of estimated coefficient using the following function:

$$var(\hat{\beta}_{rqp}(t)) = \sum_{j=1}^{T} \sum_{j'=1}^{T} w_{rqp}(t_j, t) w_{rqp}(t_{j'}, t) cov(b_{rqpj}, b_{rqpj'}). \tag{3.15}$$

Then the 95% pointwise confidence interval could be constructed by

$$\hat{\beta}_{rqp}(t) \pm 2 var(\hat{\beta}_{rqp}(t))^{1/2} \tag{3.16}$$

Here we assume that the smoothers we employ use fixed smoothing windows and bias term is ignored in constructing confidence intervals.

3.3.2 Simultaneous Test

In section 3.3.1, we consider test the coefficient at each fixed time point. In this section, we focus on the overall testing problem and consider the general simultaneous hypothesis testing stated as follows:

$$H_0 : \beta(t) = 0 \text{ for all } t, \text{ vs. } H_a : \beta(t) \neq 0.$$

Before introducing the hypothesis testing procedure, we first show the estimated coefficient function followed asymptotic Gaussian process. The property is stated in theorem 3.3.2.

Theorem 3.3.2 (**Asymptotic property of Smoothed Estimator**). *Assuming conditions* $A1 - A6$ *hold, we have,*

$$\hat{\beta}_r(t) \sim GP(E(\hat{\beta}_r(t)), \gamma_\beta(t_i, t_k))$$

where

$$E(\hat{\beta}_r(t)) = \sum_{j=1}^{T} w_r(t_j, t)\beta_r(t_j)$$

and

$$\gamma_\beta(t_i, t_k) = cov(\hat{\beta}_r(t_i), \hat{\beta}_r(t_k)) = \sum_{j=1}^{T}\sum_{j'=1}^{T} w_r(t_j, t_i)w_r(t_{j'}, t_k)cov(b_r(t_j), b_r(t_{j'}))$$

We here propose the following global test statistic for the general hypothesis testing problem:

$$T_n = \sum_r \int_0^T \hat{\beta}_r{}^2(t)dt \tag{3.17}$$

To derive the asymptotic random expression of T_n, we assume that $\gamma_\beta(t_i, t_k)$ has finite trace, that is, $tr(\gamma_\beta) = \int \gamma_\beta(t, t)dt < \infty$. We then do eigenvalue decomposition for the γ_β. Let $\lambda_1, \lambda_2, \cdots$ be the eigenvalues in decreasing order, and $\phi_1(t), \phi_2(t), \cdots$ be the associated orthonormal eigenfunctions of $\gamma_\beta(t_i, t_k)$. Let M denotes the number of positive eigenvalues. Then $\lambda_m > 0$ for $m \leq M$ and $\lambda_m = 0$ for all $m > M$. The covaraince function $\gamma_\beta(t_i, t_k)$ has the following eigen-decomposition:

$$\gamma_\beta(t_i, t_k) = \sum_{m=1}^{M} \lambda_m \phi_m(t_i)\phi_m(t_k).$$

Then the asymptotic distribution of the test statistic could be calculated using the eigenvalue information and is stated in theorem 3.3.3.

Theorem 3.3.3 (**Asymptotic distribution of test statistic**). *Under conditions* $A1 - A6$*, we have*

$$T_n \overset{d}{=} \sum_{m=1}^{M} \lambda_m A_m + o_p(1), \quad A_m \sim \chi_1^2(u_m^2)$$

Under H_0*,* $u_m = 0$*.*

Simulation approximation is used to approximate the null distribution of T_n by the χ^2-type mixture $S = \sum_{m=1}^{\hat{M}} \hat{\lambda}_m A_m$, where $A_m \sim \chi_1^2$, $\hat{\lambda}_m$ are the eigenvalues of $\hat{\gamma}_\beta(t_i, t_k)$ and $\hat{M}$ is the corresponding number of positive eigenvalues of $\hat{\gamma}_\beta(t_i, t_k)$. The sampling distribution of S is computed based on a sample of S obtained via repeatedly generating $(A_1, A_2, \cdots, A_{\hat{M}})$.

3.4 Simulation Study

In this section, three different simulations were conducted. The goal of the simulation was to assess the effectiveness of the proposed procedure for dealing with the time lag Mendelian randomization model and to evaluate the performance of the proposed stepwise deletion algorithm for the choice of window widths and lags.

3.4.1 Selection Performance for p and q

Aim

The aim of this simulation study was to show the validity of the proposed backward stepwise time lag selection under different simulation setting.

Data-generating Mechanisms

Five different scenarios were considered in total, which were stated as follows:

1. $p = 1, q = 0$. $Y(t_j)$ was determined by $X(t_j)$.

2. $p = 1, q = 1$. $Y(t_j)$ was determined by $X(t_{j-1})$.

3. $p = 1, q = 2$. $Y(t_j)$ was determined by $X(t_{j-2})$.

4. $p = 2, q = 0$. $Y(t_j)$ was determined by $X(t_j)$ and $X(t_{j-1})$.

5. $p = 2, q = 1$. $Y(t_j)$ was determined by $X(t_{j-1})$ and $X(t_{j-2})$.

Each subject had 20 repeated measurements and the time points $t_1, \cdots, t_{20}$ were chosen to be equidistant between 0.1 and 1. In our study, we assumed the effects of genetic variants on exposure and the effect of exposure on outcome both were time-varying.

In the first-stage regression $X(t_j) = \alpha(t)G + \epsilon_1(t_j)$, 15 SNPs were generated in total, among them 5 SNPs were simulated as valid instrumental variables with time-varying effects on exposure and the rest SNPs had zero coefficients. For each SNP variable G, the SNP allele frequency (p) was generated from a uniform (0.1, 0.4), then SNP values was sampled from {0, 1, 2} with probability p^2, $2p(1-p)$ and $(1-p)^2$ to obtain homozygous, heterozygous, and other homozygous genotypes, respectively. We defined the true varying coefficients for the intercept and the five SNPs as follows:

$$\alpha_0(t) = 0.1\cos(2\pi t) + 0.2, \qquad \alpha_3(t) = 0.5\sin(\pi t) + 0.6,$$
$$\alpha_1(t) = 2t, \qquad \alpha_4(t) = 0.5\cos(\pi t/2) + 0.6,$$
$$\alpha_2(t) = (1-t)^3 + 0.2, \qquad \alpha_5(t) = 0.3\sin(\pi t/3) + 0.5,$$
$$\alpha_6(t) = \cdots = \alpha_{15}(t) = 0.$$

where $\alpha_0(t)$ was the intercept function. The simulated X values were then applied in the second-stage regression to generate outcome Y.

In the second-stage regression $Y(t_j) = \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j)X(t_{j-q-(r-1)}) + \epsilon_2(t_j)$, two different values of p were considered. When $p = 1$, $\beta_0(t) = 0.2t + 0.2$ and $\beta_1(t) = 0.015 + 0.01t$. When $p = 2$, we let $\beta_0(t) = 0.2t + 0.2$, $\beta_1(t) = 0.5 + sin(\pi t)$ and $\beta_2(t) = 0.3 + 0.5cos(\pi(t - 0.5))$ respectively.

To include confounding effects, error terms $\epsilon_1(t)$ and $\epsilon_2(t)$ were generated together by assuming a variance-covariance matrix $\mathbf{\Sigma} = cov(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. The entry in Σ_{11} and Σ_{22} was set to be $0.1 \times (0.5)^{|i-j|}$ for $i, j = 1, \cdots, 20$, and for Σ_{12} and Σ_{21}, the off-diagonal element was generated

to be $0.1 \times (0.1)^{|i-j|}$, while the diagonal entry was set as 0.02. Under each setting, the simulation was repeated 1000 times.

Targets

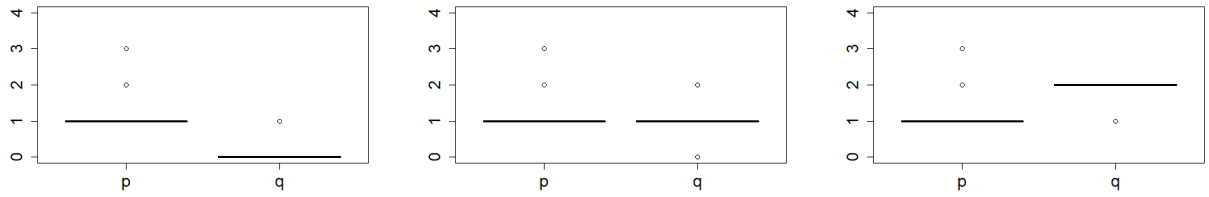
The task of the simulation was to evaluate model selection performance by the proposed time lag selection algorithm described in section 3.2.3. Under each scenario, the selected p and q values were plotted with the boxplot to measure the selection performance.

Analysis Method

In the first stage regression, the QIF with group-wise SCAD penalty was applied for instrumental variable selection and exposure estimation. In the second stage regression, the time lag selection algorithm was applied to choose optimal p and q values used for model construction.

Performance Measures

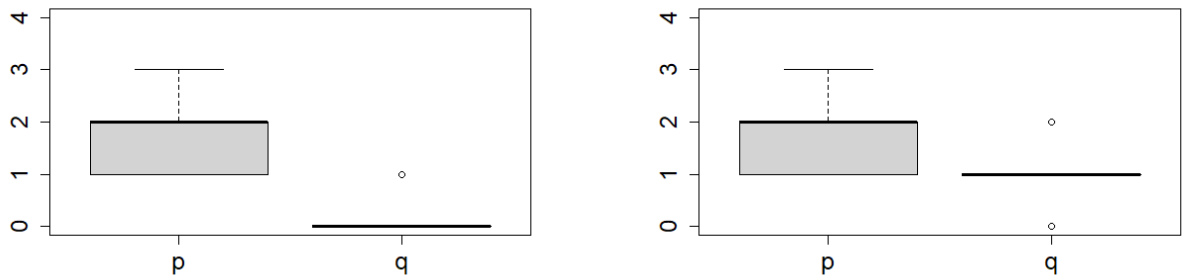
Since the effectiveness of variable selection depended on the initial set of predictors, we tried different initial groups setting for each simulation, which might result in different p and q selected values. Thus, for each simulation, we might get one or more than one pairs of (p, q) combination, and then we selected the most often pair as our final window width and lag. The results were shown in Figure 3.1 and Figure 3.2. When $p = 1$, p and q could be correctly selected under different q settings (see Figure 3.1). If we looked at all possible values selected for p and q, there existed other possible p and q combinations. Since we only focused on the most frequently chosen values, we could correctly pick window width and lag.



(a) $p = 1, q = 0$ (b) $p = 1, q = 1$ (c) $p = 1, q = 2$

Figure 3.1 Boxplot of the time lag selection under different true values of p and q.

When $p = 2$, we tried two different scenarios: $q = 0$ and $q = 1$. As displayed in Figure 3.2, q could be correctly selected under both scenarios. However, when true p value was 2, the percentage of true p being chosen was smaller than the percentage of true p being chosen when p was 1. Although p can be correctly selected in most cases, there were simulation runs that p was selected as 1 or 3. Overall, the simulation results showed that our proposed stepwise variable selection methods could reasonably select the true p and q values.



(a) $p = 2, q = 0$ (b) $p = 2, q = 1$

Figure 3.2 Boxplot of the time lag selection under different true values of p and q.

3.4.2 Performance of the Coefficient Estimation

Aim

The aim of this simulation study was to evaluate the performance of the estimation precision for the time lag model.

Data-generating Mechanisms

Two different sample size settings were considered: $n = 200$ and $n = 400$. In the first-stage regression $X(t_j) = \alpha(t)G + \epsilon_1(t_j)$, simulation setup was the same as the setup introduced in Section 3.4.1 and the details was omitted here.

In the second-stage regression $Y(t_j) = \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j)X(t_{j-q-(r-1)}) + \epsilon_2(t_j)$, we considered two different settings: $p = 1$ and $p = 2$. When $p = 1$, we let $q = 0$, which means current $X(t_j)$ has a causal effect on current $Y(t_j)$ value and Y was simulated by $Y(t_j) = \beta_0(t_j) + \beta_1(t_j)X(t_j) + \epsilon_2(t_j)$. To define the coefficient functions $\boldsymbol{\beta}(\cdot) = (\beta_0(\cdot), \beta_1(\cdot))^T$, we set $\beta_0(t) = 0.2t + 0.2$ and $\beta_1(t) = 0.5 + sin(\pi t)$. When $p = 2$, we also let $q = 0$. In this case, we assumed that both $X(t_j)$ and $X(t_{j-1})$ affect $Y(t_j)$. Y was then simulated by $Y(t_j) = \beta_0(t_j) + \beta_1(t_j)X(t) + \beta_2(t_j)X(t_{j-1}) + \epsilon_2(t_j)$, where $\boldsymbol{\beta}(\cdot) = (\beta_0(\cdot), \beta_1(\cdot), \beta_2(\cdot))^T$ and we let $\beta_0(t) = 0.2t + 0.2$, $\beta_1(t) = 0.5 + sin(\pi t)$ and $\beta_2(t) = 0.3 + 0.5cos(\pi(t - 0.5))$, respectively.

To include confounding effects, error terms $\epsilon_1(t)$ and $\epsilon_2(t)$ were generated simultaneously by assuming a variance-covariance matrix $\boldsymbol{\Sigma} = cov(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. The setup was the same as the setup introduced in Section 3.4.1 and the detail was omitted here.

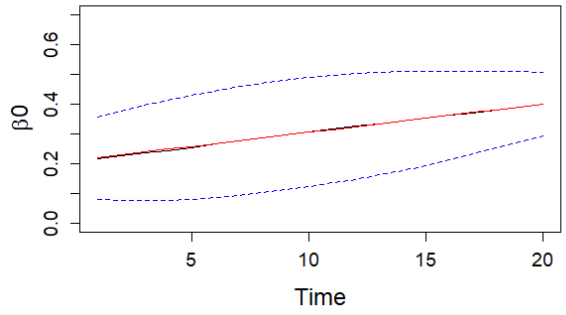
Again, 1000 simulation runs were conducted to obtain the point-wise estimator and the corresponding 95% confidence interval.

Estimands

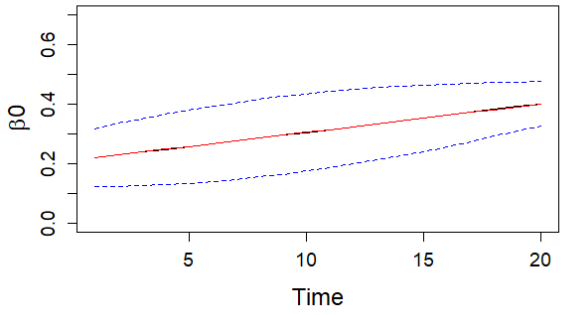
The estimands of the simulation was the time-varying coefficient $\beta(t)$ in the second stage regression. The point-wise estimator together with the corresponding 95% confidence interval were plotted to evaluate the estimation performance.

Analysis Method

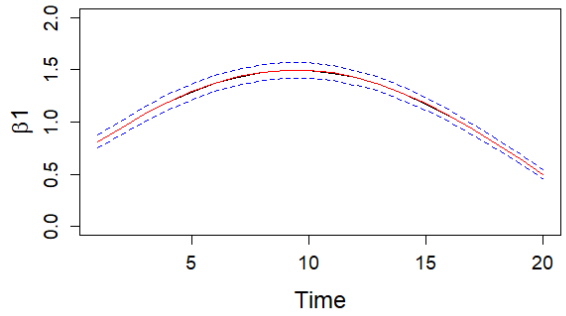
In the first stage regression, the QIF with group-wise SCAD penalty was applied for instrumental variable selection and exposure estimation. In the second stage regression, the two-step estimation



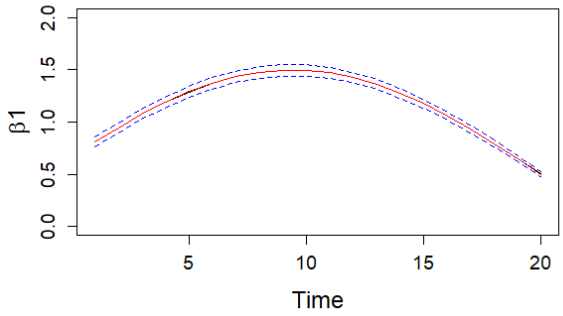
(a) β_0 estimation when $n = 200$



(b) β_0 estimation when $n = 400$



(c) β_1 estimation when $n = 200$



(d) β_1 estimation when $n = 400$

Figure 3.3 Coefficient estimation when $p = 1$. The solid red curve represents the true effect function. The solid black curve and dashed blue curves in each figure represent the estimated effect function and the 95% confidence interval, respectively.

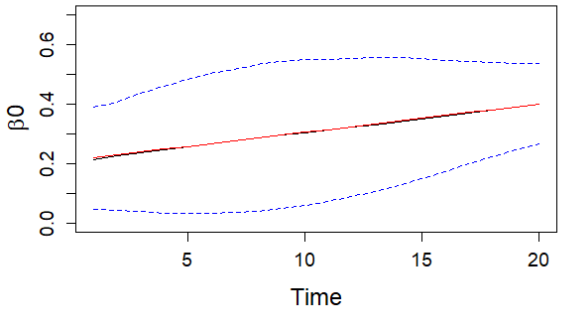
method was applied for point-wise estimator estimation and confidence interval construction.

Performance Measures

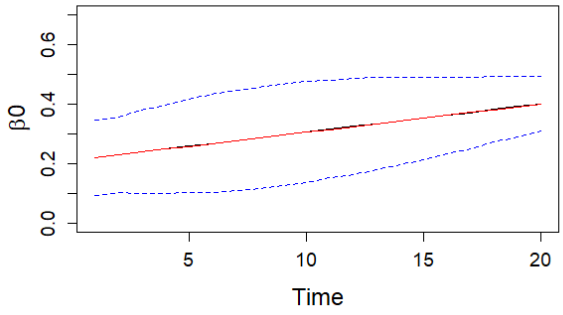
Simulation results were shown in Figure 3.3 for p=1 and in Figure 3.4 for p=2. In Figure 3.3, the plots showed our time lag model could correctly estimate the coefficients, since the estimated black line and the true red line almost exactly coincided for both intercept β_0 and coefficient β_1 when $p = 1$. This happened for both n=200 and n=400 situations. For the estimated point-wise confidence intervals, they always contained true coefficients under all the situations in Figure 3.3. It was obvious to see the confidence interval became narrower as the sample size increased.

When $p = 2$, the lag model could also precisely estimate the coefficients. The approximated

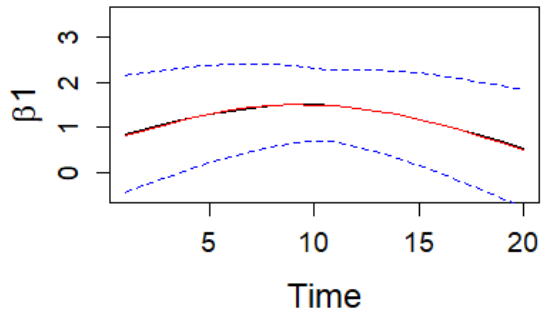
coefficients matched the true coefficients in all subplots in Figure 3.4. The confidence intervals also always involved the true coefficients for all time points and the confidence interval became tighter when the sample size increased from 200 to 400. When p=2, the confidence interval for the intercept β_0 behaved quite similar to the intercept for p=1. However, for the other coefficients, different performance could be observed. The confidence interval had approximately similar width at each time points when p=1. When p=2, the confidence interval was narrower in the middle but became wider at both the beginning and the end.



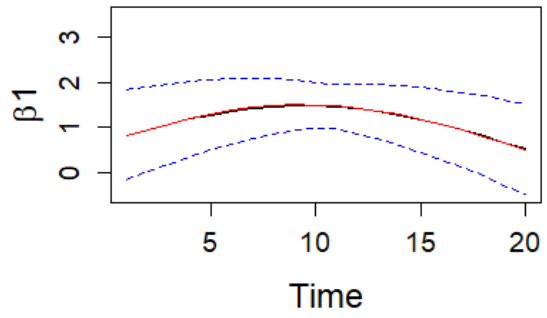
(a) β_0 estimation when $n = 200$



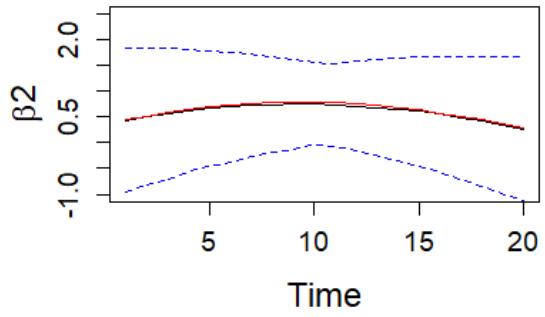
(b) β_0 estimation when $n = 400$



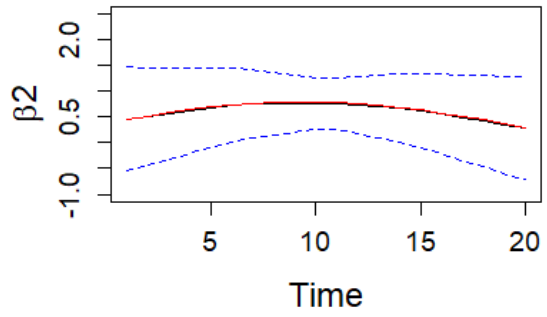
(c) β_1 estimation when $n = 200$



(d) β_1 estimation when $n = 400$



(e) β_2 estimation when $n = 200$



(f) β_2 estimation when $n = 400$

Figure 3.4 Coefficient estimation when $p = 2$. The solid red curve represents the true effect function. The solid black curve and dashed blue curves in each figure represent the estimated effect function and the 95% confidence interval, respectively.

3.4.3 Performance of Simultaneous Testing

Aim

The aim of this simulation study was to evaluate the simultaneous testing performance of the proposed time lag model. The ideal model should well protect the type I error rate at the $\alpha = 0.05$ significance level and obtain good empirical power performance. In addition, investigating the properties to influence the type I error or power behavior was also of interest. The investigated properties include sample size and time lag (p and q values).

Data-generating Mechanisms

Similar to the variable selection simulation, We considered the following four situations for both $n = 200$ and $n = 400$ sample size in total:

1. $p = 1, q = 0$. $Y(t_j)$ was determined by $X(t_j)$.
2. $p = 1, q = 1$. $Y(t_j)$ was determined by $X(t_{j-1})$.
3. $p = 2, q = 0$. $Y(t_j)$ was determined by $X(t_j)$ and $X(t_{j-1})$.
4. $p = 2, q = 1$. $Y(t_j)$ was determined by $X(t_{j-1})$ and $X(t_{j-2})$.

To simulate the data, we had the similar setting as we did in coefficient estimation simulation for the first stage regression simulation and confounding effects simulation. To simulate Y, we considered two different settings: $p = 1$ and $p = 2$. When $p = 1$, we defined the coefficient functions $\boldsymbol{\beta}(\cdot) = (\beta_0(\cdot), \beta_1(\cdot))^T$, and we set $\beta_0(t) = 0.2t + 0.2$ and $\beta_1(t) = 0$ or $\beta_1(t) = 0.5 + sin(\pi t)$ for type I error and power simulation respectively. When $p = 2$, $\boldsymbol{\beta}(\cdot) = (\beta_0(\cdot), \beta_1(\cdot), \beta_2(\cdot))^T$ and we defined $\beta_0(t) = 0.2t + 0.2$, $\beta_1(t) = 0$ and $\beta_2(t) = 0$ to test simultaneous type I error; while $\beta_0(t) = 0.2t + 0.2$, $\beta_1(t) = 0.5 + sin(\pi t)$ and $\beta_2(t) = 0.3 + 0.5cos(\pi(t - 0.5))$ to test simultaneous power. The only difference from the previous simulation is that we focus on pointwise estimation in the previous section, but our interest is simultaneous testing in this section.

Targets

The task of the simulation was to evaluate the time model for testing the null hypothesis $\beta(t) = 0$ for all t simultaneously. The testing performance was measured by the type I error rate and power.

Analysis Method

In the first stage regression, the QIF with group-wise SCAD penalty was applied for instrumental variable selection and exposure estimation. In the second stage regression, the two-step estimation method was applied to estimate smoothed coefficient $\beta(t)$ which was then used for simultaneous testing.

Performance Measures

The simultaneous testing results were given in Table 3.1. The Type I errors were well controlled at $\alpha = 0.05$ significance level under all simulation situations. The empirical power increased when the sample size went from $n = 200$ to $n = 400$. Under the same p, q had little impact on the testing power. This result showed our proposed time lag Mendelian Randomization model could not only protect type I error but also achieve good power performance.

Table 3.1 Simultaneous testing simulation result.

p	q	n	Type I error	power
1	0	200	0.052	0.855
		400	0.055	0.994
	1	200	0.045	0.861
		400	0.054	0.992
2	0	200	0.053	0.578
		400	0.052	0.996
	1	200	0.050	0.498
		400	0.053	0.998

3.5 Case Study: Albert Twin Data

We used the same data set, the Albert twin data set to investigate the delayed effects of hormones on teen girl's eating behavior, specifically, the effect in changes of estradiol and progesterone levels on emotional eating across the menstrual cycle. The details of data set was described in section 2.4.1 and was omitted here.

We first conducted marginal QIF testing for each SNPs and picked top 100 SNPs which were then used for variable selection in the first step regression. We defined the spline order to be 3 and knots to be 1 to calculate the spline basis function. Then pQIF with group SCAD penalty was applied to estimate the hormone values. The predicted hormone was used in the second step as input variable. In the second step, we first applied time lag selection algorithm to select optimal window width p and lag q. This algorithm gave us $p = 2$ and $q = 0$, which meant the current value Y_t was caused by the the current value X_t and the previous value X_{t-1}. We then used the two-step estimation to estimate the coefficient of progesterone on DEBQ and PANAS respectively. The simultaneous test was then conducted to test if progesterone was causally related to the emotional eating. When we used $p = 2$ and $q = 0$, the simultaneous test led to a p-value be equal to 0.02 for emotional eating DEBQ and 0.638 for PANAS. Thus, we focused on the analysis of progesterone level and emotional eating DEBQ across the menstrual cycle. We could conclude that the progesterone level had delayed causal effect on emotional eating DEBQ.

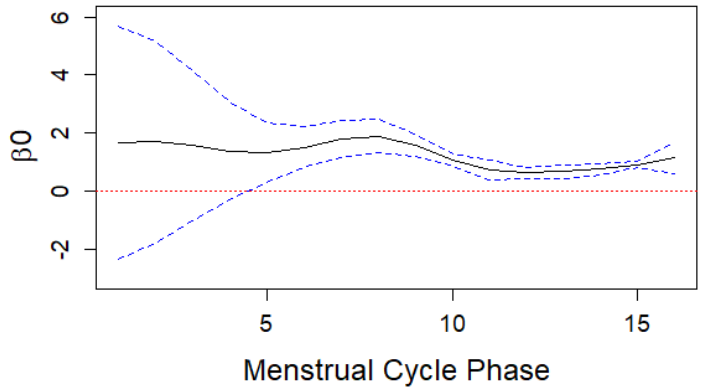


Figure 3.5 Relationship between pro and DEBQE.

We plotted the coefficients estimation together with corresponding point-wise confidence interval for β_0, β_1 and β_2 in Figure 3.5, Figure 3.6 and Figure 3.7, respectively. As shown in Figure 3.5, β_0 fluctuated slightly over time and all estimated β_0 values after smoothing were positive. The point-wise confidence interval was bigger at the beginning and shrank as the growth of phases. For β_1, it was negative during the majority of the phases, and showed an unapparent unimodal distribution. The estimated β_1 value climbed before point 7, reaching the peak at point 7 and kept decreasing slowly after the peak. The coefficient was approximately flat between point 5 and point 12, with significant growth only at the beginning. The point-wise confidence intervals at the start and at the final did not include 0, while 0 was contained in the confidence interval in the middle stages. For the coefficient β_2 of X_{t-1}, the approximated values were almost 0 between point 1 and point 3. The rest of the smoothed β_2 values were all around half negative and half positive during the remaining menstrual cycle phases. Similarly, the point-wise confidence interval was wider at the beginning. The main part of the confidence intervals involved 0.

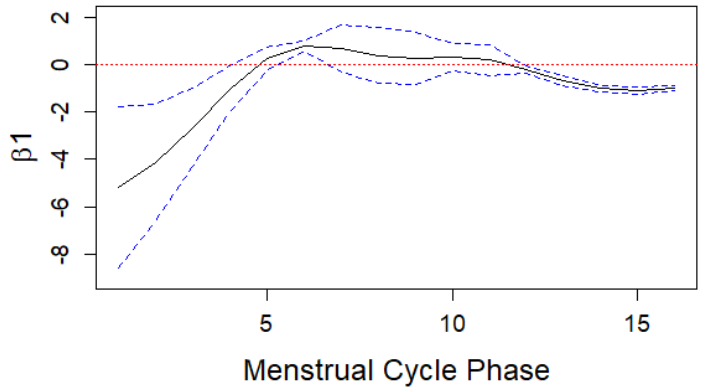


Figure 3.6 Relationship between Pro and DEBQE.

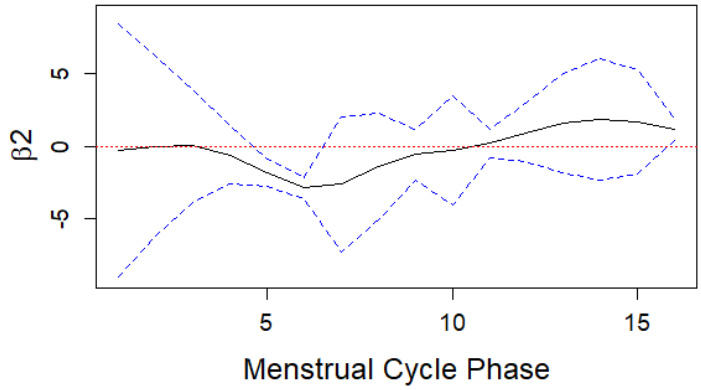


Figure 3.7 Relationship between Pro and DEBQE.

Besides simultaneous testing, we also conducted point-wise testing for the correlation between progesterone and DEBQ and plotted the corresponding point-wise $-log_{10}$(P value) in Figure 3.8. As shown in the bar plot, the red dotted line represented $\alpha = 0.05$ significance level, and we found significant p-values at phase 6, 7, 10 and 15. Although not all menstrual cycle phases had significant p-values for the point-wise testing, we could still achieve significant p-value when conducting simultaneous test.

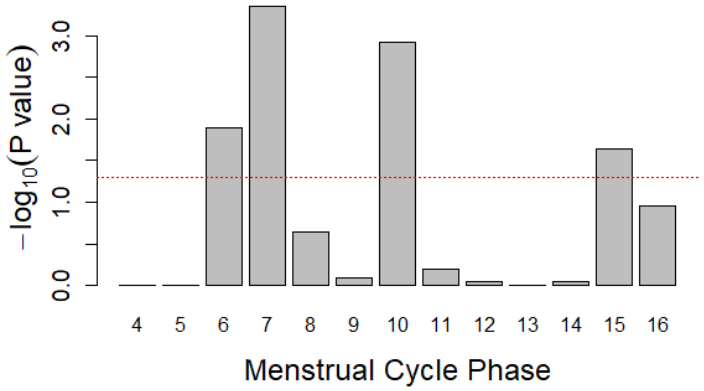


Figure 3.8 Pointwise p-value for DEBQE.

3.6 Conclusion and Discussion

One important problem in scientific fields is the identification of causal effects from observational study. For two series of longitudinal measurements X and Y, the most basic approach to inferring causal relationship is to use the correspondence measure between a lagged version of the potentially-causing X to the non-lagged potentially-caused Y. The notion of X-causing-Y can be inferred if a high degree of correspondence is founded between a k-lag of X and Y.

In this chapter, we proposed a time lag model and considered delayed effects situation. We assumed the current outcome Y was not only affected by current exposure but also might be affected by the previous value of exposure. We used two-stage instrumental variable regression to solve the proposed model. In the first step, we applied penalized QIF to estimate and select causal genetic variants which had time-varying effects on exposure. In the second step, our proposed model had similar formula to a finite distributed lag model which was a common model to analyze time series data in statistics and econometrics. The fitted exposure was then used to substitute in the second step regression. To select appropriate lag period included in the model, we proposed the algorithm based on the idea of backward stepwise deletion. The simulation results suggested our proposed algorithm could select the true lag period under different settings with reasonable accuracy. After selecting appropriate time lag, we considered two hypothesis testing problems: point-wise testing

and simultaneous testing to test the existence of the time-varying causal effect. From the real data analysis, we found one causal relationship between hormone progesterone and emotional eating DEBQ. And no casual relationship was found between progesterone level and PANAS. These findings are similar to the findings in chapter 2. However, only contemporary causal effect of progesterone level on DEBQ was observed in chapter 2. The using of time lag model contributed to the findings of not only current causal effect but also delayed causal effect of hormone progesterone on emotional eating DEBQ.

Although the proposed time model included the delayed effects of exposure variables, other covariates effects were not considered in the regression models due to the simplicity consideration. How to adjust for other covariates effects is also of great importance in the model construction. We can simply incorporate the covariates into the two-stage regressions. However, there exists many situations in reality: the effects of covariates might be time-invariant, time-varying or both. The interaction effects could also be included in the model.

For the second stage regression model, distributed lag model can involve using one or more lagged values of response variables as determinants of the current response value, such as the autoregressive lag model. Besides the autocorrelated response variables, there also exist possibility for model errors to be autocorrelated. However, we did not consider the autoregressive situations in time lag model. We actually assumed current response value was not affected by the past values of response but was only determined by exposure measurements. Further investigations for the more complex cases are our future work.

The correct MR results depend on three critical assumptions of the valid instruments variables, which are difficult to verify. Therefore, sensitivity analysis methods are necessary for evaluating results and making plausible conclusions. Weak instrument bias and the pleiotropic bias are two common challenges arise in MR due to invalid instrumental variables. Since we assumed time-varying coefficients for SNPs, QIF method was applied to select the IV variables instead of using common GWAS study with LD-pruning. In this study, we did not conduct sensitivity analysis to further investigate the pleiotropy and weak instruments situation. Further investigation is needed

to confirm the validity of the selected instrumental variables.

CHAPTER 4

MENDELIAN RANDOMIZATION FOR LONGITUDINAL DATA WITH CUMULATIVE EFFECT

The motivation of this chapter is to investigate the cumulative effect of exposure on an outcome at time t. For an outcome measured at time t, we would like to test if the cumulative effect of an exposure up to t exerts a causal effect on the outcome at time t. The concurrent model described in chapter 2 assumed the current response was only affected by the current exposure. In the real data analysis in Chapter 2, we found the contemporary casual relationship between hormone progesterone and emotional eating DEBQ, and the current DEBQ was affected by current progesterone level. One may further be interested in the cumulative effect of progesterone on DEBQ and would like to answer the question: "Does past value of hormone progesterone cumulatively contribute to the current emotional eating behaviour?" To address this problem, we consider a functional model which includes all information of the exposure up to time t in this chapter. We demonstrate our model in the simulation study with well protected type I error control.

4.1 Introduction

Functional data analysis (FDA) focuses on data that is on infinite-dimensional such as curves, shapes, images, or anything else varying over a continuum. In FDA, there are two typical types of data: dense functional data and sparse functional data. Dense functional data consists of a large number of regularly observed measurements for each subject. Sparse functional data contains irregularly-spaced measurements on some small number of time points over the time domain. Sparse functional data is common in many real-world applications such as longitudinal studies.

There exists substantial literature on modeling and estimation for sparse functional data. For the testing purpose, most of the functional testings require the individual curves from each subject being observed at the same dense regular grid. Pre-smoothing methods are applied for each curve if observing times are not the same for all the subjects. However, this technique based on individual

curves is not reliable for sparse functional data because of the limited number of observations for each subject. Wang [82] developed an asymptotic χ^2 test for detecting the differences among the mean functions of two independent stochastic process with homogeneous covaraince functions when only a few irregularly spaced measurements were given for each subject. The pseudo likelihood ratio test was proposed by Staicu et al. [74] and aimed to testing the structure of the mean function of complex functional processes and was applicable to sparsely sampled functional data. Pomann et al. [61] used marginal functional principal component analysis to decompose the curves and developed the nonparametric distribution test for testing the null hypothesis that two samples of curves observed at discrete grids and with noise had the same underlying distribution. Wang et al. [81] proposed unified empirical likelihood ratio tests to make pointwise and simultaneous inferences on functional concurrent linear models, treating sparse and dense functional data in a unified framework.

Another popular direction of sparse functional data analysis is to investigate dimensionality reduction. Staniswalis and Lee [75] used kernels to smooth the covariance surface from which the functional principal components were estimated, using quadrature to estimate the functional principal component scores when functions were sampled on sparse, irregular grids that varied across functions. Yao et al. [84] developed a version of functional principal components (FPCs) analysis, in which the FPC scores were framed as conditional expectations, to handle sparse and irregular longitudinal data for which the pooled time points were sufficiently dense. Di et al. [24] considered analysis of sparsely sampled multilevel functional data, where the basic observational unit was a function and data had a natural hierarchy of basic units. He proposed sparse multilevel FPCA (MFPCA) which generalized the MFPCA method he developed before from densely sampled functions to sparsely observed functions.

In this chapter, we consider the functional model for Mendelian Randomization analysis and assume the exposure has cumulative effects on the outcome. To solve this functional Mendelian Randomization model, we develop a two-stage instrumental variable regression. In the first step, the penalized QIF method is applied for causal instrumental variable selection and exposure prediction.

The estimated exposure is then applied in the second step. We treat the time series exposure as sparse functional data and construct FPC analysis using the PACE method proposed by Yao et al. [84]. The FPCs are then inserted in the regression model in the second step to investigate the causal relationship.

The rest of paper is organized as follows: Section 2 introduces the functional Mendelian Randomization models and the corresponding methods to deal with each regression. In section 3, we evaluate the performance of our model via simulation. We apply our approach to the Albert twin data set in section 4, followed by conclusion and discussion.

4.2 Functional Model

Suppose there are n subjects in total. For each individual, i, the exposure and the outcome are measured at multiple time points $\{t_j, j = 1, 2, \cdots, T\}$. Let $Y_i(t_j)$ and $X_i(t_j)$ be the time-varying outcome and exposure of subject i recorded at time t_j respectively. G_i denotes the vector of multiple SNPs of subject i and is time invariant. The data collected are denoted as,

$$\{Y_i(t_j), X_i(t_j), G_i\}, \quad \text{for } i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, T.$$

In the chapter 3, we considered a time lag model which assumed not only current predictor value $X(t_j)$ had an influence on response $Y(t_j)$ at current time t_j but also recent past exposures played a causal role on an outcome measure at time t_j. Although genetic variants are time-invariant, since we observe longitudinal exposure and outcome, we assume the effects of genetic variants on exposure change over time, and the effect of exposure on outcome variable might change over time as well. We assume the effects of genetic variants on exposure and the effects of exposure on outcome both are time-varying in the time lag model and the model can be formulated as follows:

$$X(t_j) = \alpha(t_j)G + \epsilon_1(t_j), \tag{4.1}$$

$$Y(t_j) = \beta_0(t_j) + \int_0^{t_j} \beta_1(t)X(t)dt + \epsilon_2(t_j) \tag{4.2}$$

The time lag model considered the cumulative effect of the previous p time points on an outcome. The total number of past points included in the model is either 2 or 3 in the previous

chapter. In this chapter, we consider beyond 2 time points of exposures having effects on outcome and treat the repeated measurements as sparse functional data. We have the same assumption on the genetic variants that the genetic effect on a time-varying exposure variable changes over time. For the time-varying exposure, we assume it has cumulative effects on the time-varying outcome and the dependent variable at time t_j can be determined by the independent variables measured up to time t_j.

One major complication that is emphasized is the possibility of inconsistent parameter estimation due to endogenous regressors. The estimated association does not mean causation under this situation. When unobservable latent variables affect both $\mathbf{X}$ and Y, $\mathbf{G}$ can be considered as instrumental variables to remove the effects of confounding factors and the effects of $\mathbf{X}$ on Y can be consistently estimated. In the first stage, a penalized variable selection algorithm is applied to select genetic variants and estimate time-varying exposure simultaneously. Then $\mathbf{X}$ is replaced by the fitted values $\hat{\mathbf{X}}$ in the second stage and $\hat{\mathbf{X}}$ removes the effects of confounding factors on exposure variable. $\hat{\mathbf{X}}$ is used in the second stage to infer the causal relationship between the cumulative effect of an exposure variable and an outcome at time t_j.

4.2.1 Estimation of the time-varying SNP effect

Similar to chapter 2 and chapter 3, we apply the idea of QIF in the first stage regression. We apply QIF method to test the significance of each individual genetic variant. To solve the first equation in (4.1), we do similar operations as we did in previous chapters. We use penalized quadratic inference function with group-wised SCAD penalty to select causal genetic variants. The detailed estimation procedure is omitted here.

4.2.2 Estimation and testing of the functional exposure effect

Even though longitudinal data have become more common in observational studies, they are often sparse and collected at irregular time points. Different number of observations may also be recorded for different subjects. To recover the curves and characterize the dominant modes of variation of

a sample of random trajectories, Yao et al. [84] proposed principal components analysis through conditional expectation (PACE) method to perform functional principal components analysis for the case of sparse and irregularly spaced longitudinal data by assuming that the longitudinal measurements are located randomly with a random number of repetitions for each subject and are sampled from an underlying curve with noise and the curves of all the subjects are independent with the same mean function and covariance function. We apply PACE method in the second-stage regression for the dimensionality reduction purpose.

Let mean function be $EX(t) = \mu(t)$ and covariance function be $G(s,t) = cov[X(s), X(t)]$ for the collection of curves that are assumed to be independent realizations of a smooth random function. The domain of $X(t)$ is in a closed and bounded time interval $\mathbb{T}$. Eigen decomposition can be performed to expand the covariance function as

$$G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t),$$

where λ_k's are nonnegative eigen-values with descending order and $\phi_k(t)$'s are corresponding eigen functions. The ith random curve can be expressed as

$$X_i(t) = \mu(t) + \sum_k \xi_k \phi_k(t)$$

in classical functional principal component (FPC) analysis. The kth FPC score of subject i, ξ_{ik} is regarded as uncorrelated random variables with mean 0 and variance $E\xi_{ik}^2 = \lambda_k$. When the density of the grid of measurements for each subject is sufficiently high, the FPC score is estimated by $\xi_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)dt$.

However, this integration does not provide reasonable approximations for sparse data and will lead to biased FPC scores if the measurements are contaminated with errors. To overcome the problem, the PACE method proposed by Yao et al.[84] introduces the best prediction of FPC score ξ_{ik} as the conditional expectation

$$\tilde{\xi}_{ik} = E(\xi_{ik}|X_i) = \lambda_k \phi_{ik}^T \Sigma_{X_i}^{-1}(X_i - \mu_i),$$

where $\phi_{ik} = (\phi_k(t_{i1}), \cdots, \phi_k(t_{iN_i}))^T$, $\Sigma_{X_i} = cov(X_i, X_i) + \sigma^2 I_{N_i}$, and $\mu_i = (\mu(t_{i1}), \cdots, \mu(t_{iN_i}))^T$. This conditional expectation is aimed at analyzing the model that incorporated uncorrelated additive

measurement errors with mean 0 and constant variance σ^2 and is defined as

$$x_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t_{ij}) + \epsilon_{ij},$$

where ξ_{ik} and ϵ are assumed to follow jointly Gaussian distribution. The mean function μ is estimated based on the pooled data from all individuals using a local linear smoother, while the covariance function $G(s,t)$ also borrows strength from the entire dataset and is estimated by fitting a local quadratic component along the direction perpendicular to the diagonal and a local linear component in the direction of the diagonal.

After computing the smooth surface estimator of $G(s,t)$, eigen decomposition is applied to calculate λ_k and ϕ_k, which can be plugged in the conditional expectation equation to estimate FPC score ξ_{ik}. If we assume that the infinite-dimensional processes under consideration are well approximated by the projection on the function space spanned by the first K eigenfunctions, then the recovered individual curve is predicted using the first K eigenfunctions and has the following form:

$$\hat{x}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t), \quad \text{for } t \in \mathbb{T}.$$

In the first step, we obtain an estimate of spline coefficients $\hat{\gamma}$ to get an estimator for $\alpha_l(t)$ given by $\hat{\alpha}_l(t) = \sum_{v=0}^{V} \hat{\gamma}_{lv}B_v(t)$. The fitted value $\hat{\boldsymbol{X}}(t) = \hat{\boldsymbol{\alpha}}(t)\boldsymbol{G}$ is then used to substitute in the second step for the time-varying exposure effect estimation. We then apply PACE method to perform functional principal component analysis for the sparse longitudinal data $\hat{\boldsymbol{X}}(t)$ from the first step. After the functional principal component analysis, we next regress response on the the principal component scores for the functional covariate estimation. Since the random curve $X_i(t)$ can be expressed as $\hat{X}_i(t_j) = \mu(t_j) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t_j)$, the functional slope $\beta_1(t)$ in model (4.2) can also be written in terms of $\phi_1, \phi_2, \cdots$ as

$$\beta_1(t) = \sum_{k=1}^{K} b_k\phi_k(t)$$

Regressing y_i on the principal component scores gives us the following model:

$$Y_i(t_j) = \beta_0 + \sum_{k=1}^{K} b_k\xi_{ik} + \epsilon_2(t_j) \tag{4.3}$$

in which the response is written as an linear combination of FPC score ξ_{ik}. Approximate pointwise standard errors can be constructed out of the covariance matrix of the b_k:

$$var(\hat{\beta}_1(t)) = (\phi_1(t) \cdots \phi_K(t))^T var(b)(\phi_1(t) \cdots \phi_K(t)).$$

In order to test the existence of causal effect $H_0 : \beta_1(t) = 0$, we can test the coefficients of principal component scores instead, i.e. $H_0 : b_k = 0$ for all k. The Wald test can be applied to analyze this hypothesis testing problem. Since regressing Y on the infinity predictors is impossible, one important thing is to select appropriate number of eigenfunctions. We introduce several criteria for the eigenfunctions selection in section 4.2.3.

4.2.3 Select the number of eigen-functions

Several criteria can be applied to select the number of eigenfunctions that provides a reasonable approximation to the infinite-dimensional process. We can use the cross-validation score based on the leave-one-curve-out prediction error which is defined as follows:

$$CV(K) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \{x_{ij} - \hat{x}_i^{(-i)}(t_{ij})\}^2$$

where $\hat{x}_i^{(-i)}$ is the predicted curve for the ith subject computed after removing the ith subject.

Another criteria we can adapt is Akaike information criterion (AIC). A pseudo-Gaussian log-likelihood, summing the contributions from all subjects, conditional on the estimated FPC scores $\hat{\xi}_{ik}$ is given by

$$\hat{L} = \sum_{i=1}^{n} \{-\frac{N_i}{2} \log(2\pi) - \frac{N_i}{2} \log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(x_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_{ik})^T (x_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_{ik})\}$$

AIC is then defined as $AIC = -\hat{L} + K$. Besides cross-validation score and AIC, the fraction of variance explained (FVE) or Bayesian information criterion (BIC) can also be applied to choose optimal number of eigenfunctions.

4.2.4 Functional F test

Since Mendelian Randomization aims to address causal questions about how modifiable exposures influence different outcomes, hypothesis testing is the primary research of interest instead of prediction. However, it is difficult to attempt to derive the theoretical null distribution for the test statistic because of the nature of functional statistics. As we discussed in section 4.2.2, a Wald test can be applied to test the existence of causal effect. In this section, we introduce an F statistic which is defined as follows:

$$F = \frac{Var(\hat{Y})}{\frac{1}{n}\sum(Y_i - \hat{Y}_i)^2},$$

where $\hat{Y}$ is the vector of predicted responses. This statistic is different from the classic F statistic in the manner in which it normalizes the numerator and denominator sums of squares. We can compute a different random permutation each time and use the permutation data to calculate the test statistic. By repeating it several hundred times, a null distribution from the observed data is constructed directly. If there is no relationship between the response and the exposure, it should make no difference if we randomly rearrange the way they are paired. The p-value for the test can then be calculated by counting the proportion of permutation F values that are larger than the F statistic for the observed pairing.

4.3 Simulation Study

Aim

The aim of this simulation study was to evaluate the functional testing performance of the proposed cumulative effect model. The ideal model should well protect the type I error rate at the $\alpha = 0.05$ significance level and obtain good empirical power performance. In addition, investigating the properties to influence the type I error or power behavior was also of interest. The investigated properties include sample size and number of time points included for analysis.

Data-generating Mechanisms

Two different sample size settings were considered in total: $n = 200$ and $n = 400$. Each subject had 20 repeated measurements and the time points $t_1, \cdots, t_{20}$ were chosen to be equidistant between 0.1 and 1. In our study, we assumed the effects of genetic variants on exposure and the effect of exposure on outcome both were time-varying.

In the first-stage regression $X(t_j) = \alpha(t)G + \epsilon_1(t_j)$, 15 SNPs were generated in total, among them 5 SNPs were simulated as valid instrumental variables with time-varying effects on exposure and the rest SNPs had zero coefficients. For each SNP variable G, the SNP allele frequency (p) was generated from a uniform (0.1, 0.4), then SNP values were sampled from $\{0, 1, 2\}$ with probability p^2, $2p(1-p)$ and $(1-p)^2$ to obtain homozygous, heterozygous, and other homozygous genotypes, respectively. We defined the true varying coefficients for the intercept and the five SNPs as follows:

$$\alpha_0(t) = 0.1\cos(2\pi t) + 0.2, \qquad \alpha_3(t) = 0.5\sin(\pi t) + 0.6,$$
$$\alpha_1(t) = 2t, \qquad \alpha_4(t) = 0.5\cos(\pi t/2) + 0.6,$$
$$\alpha_2(t) = (1-t)^3 + 0.2, \qquad \alpha_5(t) = 0.3\sin(\pi t/3) + 0.5,$$
$$\alpha_6(t) = \cdots = \alpha_{15}(t) = 0.$$

where $\alpha_0(t)$ was the intercept function. The simulated X values were then applied in the second-stage regression to generate outcome Y.

To simulate Y, we used the PACE method to perform functional principal components analysis on X. The PACE results give us the estimated mean function, the eigenfunctions together with the corresponding eigen values and FPC scores, which will be used to simulate response variable. The regression function $\beta(t)$ was generated using eigenfunctions $\beta(t) = \sum b_k \phi_k(t)$, where $\phi_k(t)$ are eigenfunctions and K is the number of eigenfunctions which can be selected using AIC, BIC, cross-validation score or FVE. In the simulation, we used FVE and the FVE threshold was set to be 0.95 which means 95% of the total variance was explained by chosen FPCs. The response variable was then simulated by $Y(t_j) = \beta_0 + \sum_{k=1}^{K} b_k \xi_k + \epsilon_2(t_j)$, where b_k was the coefficient for FPC score ξ_k and was generated between 0.2 and 0.01 if K was at least 5, and between 0.3 and 0.1 if K was

more than 1 but less than 5. When K was just one, we set b_k equaled 0.8. As for the intercept b_0, we simulated it from a standard normal distribution.

To include confounding effects, error terms $\epsilon_1(t)$ and $\epsilon_2(t)$ were generated simultaneously by assuming a variance-covariance matrix $\mathbf{\Sigma} = cov(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. The entry in Σ_{11} and Σ_{22} was set to be $(0.5)^{|i-j|}$ for $i, j = 1, \cdots, 20$, and for Σ_{12} and Σ_{21}, the off-diagonal element was generated to be $(0.1)^{|i-j|}$, while the diagonal entry was set as 0.2. Then we simulate $(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) \sim N_{40}(\mathbf{0}, \Sigma)$.

The same simulation was repeated for sample size $n = 200$ and $n = 400$. Under each setting, the simulation was repeated 1000 times.

Targets

The task of the simulation was to evaluate the cumulative effect model for testing the null hypothesis $\beta(t) = 0$ at each t. The testing performance was measured by the type I error rate and power.

Analysis Method

In the first stage regression, the QIF with group-wise SCAD penalty was applied for instrumental variable selection and exposure estimation. In the second stage regression, PACE method was applied on the fitted exposure value from step one for the dimensionality reduction purpose. The FPC score coefficient was estimated by regressing outcome on FPC scores. Testing the functional coefficient $\beta(t)$ was equivalent to test FPC score coefficients.

Performance Measures

The type I error and power measured at three different time points ($T = 5, 10, 15$) were reported in Table 4.1. From the table, we could conclude that Type I error were well-controlled under all cases. For the power simulation, the empirical power improved with increasing sample size. However, when we did not have enough time information ($T = 5$), the empirical power was low. The simulation results suggested 10 repeated measurements were large enough to obtain a

relatively good performance when sample size was 400. With the increasing number of subjects, the requirement for the number of time points could be appropriately reduced.

Table 4.1 List of Type I error and power under different sample size and time points.

n	T	Type I error	power
	5	0.054	0.428
200	10	0.055	0.739
	20	0.046	0.764
	5	0.055	0.714
400	10	0.055	0.964
	20	0.051	0.956

We also conducted simulations to test the effects of confounding factors on type I error and empirical power. The simulation settings and results are introduced in appendix.

4.4 Case Study: Albert Twin Data

We used the same data set, the Albert twin data set to investigate the cumulative effects of hormones on teen girl's eating behavior, specifically, the effect in changes of estradiol and progesterone levels on emotional eating across the menstrual cycle. The details of data set was described in section 2.4.1 and is omitted here.

We first conducted QIF testing method for each SNP and picked SNPs with p-values less than 10^{-4}. In GWAS, the threshold τ is often taken to be the GWAS significance threshold $\tau = 5 \times 10^{-8}$ in order to reduce the number of false-positive associations arising from the vast number of statistical tests performed. Since we assumed varying-coefficients for SNPs and we had longitudinal measurement for exposure, the QIF results did not provide p-values as small as GWAS. Thus, using more relaxed threshold might be beneficial in this study. The threshold 10^{-4} led to 8 SNPs being selected as IVs.

Using 8 genetic variants, we predicted the progesterone levels in the first step. Then the PACE method was applied to approximate the individual progesterone levels to recover the individual progesterone curves from the estimated longitudinal data. The smoothed mean progesterone function was showed in Figure 4.1, which had a sine or cosine distribution trend. The smoothed

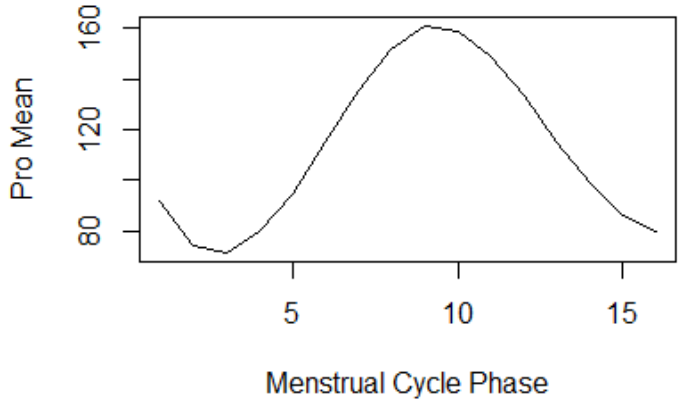


Figure 4.1 Smoothed progesterone mean function.

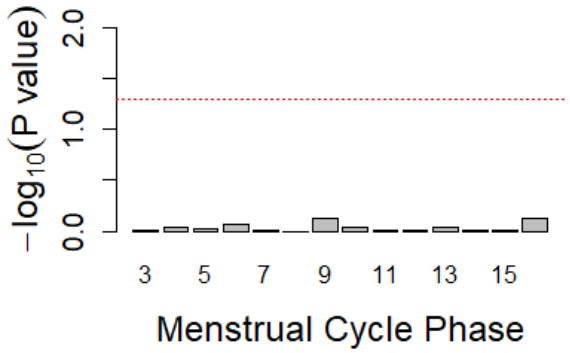


Figure 4.2 Plot of -log10(p-value) for DEBQ.

mean function decreased at the beginning, then increased after phase 3 and reached the peak at phase 9. The mean function continuously decreased after phase 9. We used the leading eigenfunctions to explain a total of 95% variation. For most phases, the leading eigenfunctions were chosen to be two with only phase 3, phase 15 and phase 16 having 3 leading eigenfunctions.

We used Wald test to test the null hypothesis that the cumulative time-varying progesterone had no effect on the emotional eating outcome. Figure 4.2 and Figure 4.3 showed the $-log_{10}$(p-value) for DEBQ and PANAS, respectively. For both plots, we could not find any significant p-values during all menstrual cycle phases. The results indicated that the cumulative time-varying

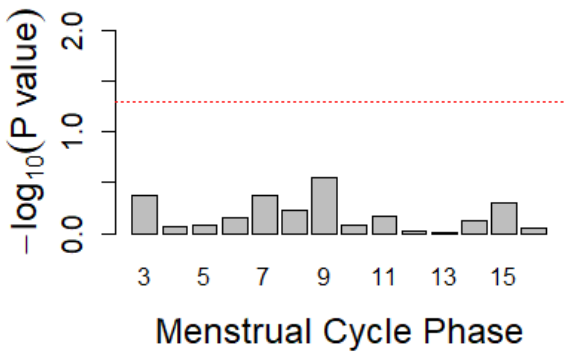


Figure 4.3 Plot of -log10(p-value) for PANAS.

progesterone had no effect on either DEBQ or PANAS. Note that the results of cumulative effect model are quite different from those obtained with the time lag model. When the outcome at time t is only affected by exposures at closed time points, including more time points may introduce more noise, hence dilute the testing signal. This might explain why we observed significant results in the time lag model but not in the cumulative effect model.

4.5 Conclusion and Discussion

In this chapter, we further investigate the cumulative causal relationship that are beyond two time points. Instead of using longitudinal analysis approach, we treated the repeated measurements as sparse functional data and proposed a functional model for Mendelian Randomization analysis purpose. In the first step, we applied QIF method as we did in the concurrent model and time lag model. In the second step, the predicted exposure from step one was utilized for functional principal component analysis. Instead of regressing response on the predicted exposure directly, we regressed response on FPC scores.

The simulation studies suggested that our proposed method could well protect the type I error rate at the significance level $\alpha = 0.05$. For the empirical power simulation, the results showed that 5 time points was relatively small for us to obtain good performance, but 10 time points contained enough information to reach a good power. The real data analysis did not give us any new findings.

We did not find any new causal relationship between hormone progesterone and emotional eating behavior using the proposed functional Mendelian Randomization model for the Albert twin data set analysis.

Compared to the results we found in the previous chapter, where not only current progesterone level, but also one past progesterone value had cumulative causal effects on DEBQ, we did not identify any causal relationship using the functional model in this chapter. The reason of the difference might be that including too many insignificant time points in the model might introduce more noise when the outcome at time t was only affected by exposures at closed time points, hence diluted the testing signal and leading to insignificant p-values. Thus, even we found significant effects using two time point in the time lag model, we did not observe significant results after integrating more observations.

In this chapter, we analyzed the response at a fixed time point. Thus the functional model in the second stage regression used a scalar response at each time point. Instead of solving functional regression with scalar response, we could also considered all response information simultaneously and applied a functional response in the future. However, if the functional response is included in the model, then the model becomes

$$Y(t) = \beta_0(t) + \int_0^t \beta_1(s,t)X(s)d(s) + \epsilon(t),$$

where $\beta_1(s,t)$ is a vector of unknown two-dimensional functional coefficient. Since the primary goal of Mendelian Randomization is testing the existence of casual effect, the hypothesis testing problem becomes complicate under this situation. More investigations are needed in the future work.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The main goal of this dissertation is to develop novel Mendelian Randomization analysis methods to investigate causal relationship when longitudinal measurements are obtained in observational studies. We considered 3 different models in total. We first proposed a concurrent model in Chapter 2 which assumed contemporary causal relationship, i.e. current response was only affected by current exposure and the linear relationship held at every time point. The idea of quadratic inference function was applied to solve the concurrent model. Following the work of Chapter 2, we extended the current model to a time lag model in Chapter 3. The time lag model considered the cumulative delayed effects and assumed not only current exposure but also past values of exposure contributed together to the current response. One important part of time lag model was to select appropriate number of time points included in the model. To solve this question, we proposed the algorithm for the variable selection. When only current exposure is selected, the time lag model degenerates to the concurrent model. We also considered both point-wise testing and simultaneous testing for the time lag model. The total time points selected for the time lag model in Chapter 3 was no more than 3. In Chapter 4, we further investigated the cumulative effects and considered more time points than time lag model. Instead of using longitudinal methods to investigate Mendelian Randomization, we treated the time series observations as sparse functional data. The functional model was proposed to investigate the overall cumulative effect.

For the simulation perspective, since Mendelian Randomization focuses more on hypothesis testing problem than regression estimation, we did simulations to compare the type I error and empirical power performance. The simulation results suggested the three different models we considered could all well protect Type I error rate at the significance level $\alpha = 0.05$ and achieve good empirical power performance.

From the application perspective, our methods development was well motivated by the Albert twin data set to investigate the causal relationship between hormone measurement and emotional eating behaviour during menstrual cycle phase. Using our proposed models, we found the timely causal relationship between progesterone level and emotional eating behaviour measurement DEBQ when applying concurrent model. For the time lag model, the variable selection algorithm chose two time points and we concluded that not only current progesterone level but also progesterone measured at time $t-1$ had casual effects on DEBQ. However, when we wanted to further investigate the cumulative effect of progesterone, the functional model did not give any causality findings. So far, we could only conclude the emotional eating behavior DEBQ measured at time t is determined by progesterone level measured at time t and time $t-1$ together.

In conclusion, this dissertation considered three different models under the MR framework with longitudinal data. We proposed new models to deal with three different effect assumptions and illustrated our developed methods by simulation studies and real data analysis.

5.2 Future Work

As we introduced in the first chapter, the reliability of Mendelian Randomization relies on the validity of genetic variants as instrumental variables. In order to be valid instrumental variables, several assumptions need to be satisfied. To test whether those assumptions hold, traditional Mendelian Randomization analysis usually conduct the sensitivity analysis. Sensitivity analysis identifies weak instruments and pleiotropy which are two common challenges we might face when calculating the casual effect. In the dissertation, we assume genetic variants have time-varying effects on the exposure X, since genetic variants are time-invariant but we observe longitudinal exposure values. The sensitivity analysis can be conducted in the future investigation to evaluate the impact of weak instruments and pleiotropy effect.

In addition to sensitivity analysis, the approach of selecting instrumental variables also need further discussion. In the dissertation, we use the idea of quadratic inference function for valid instrumental variables selection, which is different from the common used variable selection

methods in Mendelian Randomization, such as cis-MR. LD-pruning is the most common approach for selecting genetic variants for inclusion into a cis-MR study. Since we use different approach, the threshold of LD-pruning is not suitable for our studies. Different approaches to deal with genetic variants with time-varying effects may be developed in the future.

In addition, the three models we constructed did not include other covariates effects due to the simplicity consideration. How to adjust for other covariates effects is also of great importance in the model construction. We can simply incorporate the covariates into the two-stage regressions. However, there exists many situations in reality: the effects of covariates might be time-invariant, time-varying or both. The interaction effects could also be included in the model. Further investigations for the more complex cases are our future work.

Currently, there is very limited literature about Mendelian Randomization analysis for longitudinal data due to the lack of appropriate data set. In the dissertation, we applied our proposed models to the Albert twin data set. It will be helpful to evaluate the method performance in other real data sets.

Moreover, genetic instrumental variables are traditionally considered to be sufficient if the corresponding F-statistic is greater than 10 in traditional Mendelian Randomization. This criteria is often applied to decide the weak instruments and strong instruments. However, there is no clear criteria for the longitudinal data. Studying the criteria to include instrumental variables in longitudinal data is also our future work.

APPENDICES

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Simulation studies to test the influence of instrumental variables strength

To test the effect of instrumental variables strength, we did additional simulations by changing the time-varying effects of genetic variants on exposure variable. The total number of subjects was set as 200. Each subject had 20 repeated measurements and the time points $t_1, \cdots, t_{20}$ were chosen to be equidistant between 0.1 and 1. Similarly, we still assumed the effects of genetic variants on exposure and the effect of exposure on outcome both were time-varying. We generated 5 SNPs in total in the simulation and assumed all the generated SNPs were valid instrumental variables with time-varying effects on exposure variable. In Chapter 2, we defined the true varying coefficients of SNPs having the following forms:

$$\alpha_0(t) = 0.1\cos(2\pi t) + 0.2, \qquad \alpha_3(t) = 0.5\sin(\pi t) + 0.6,$$
$$\alpha_1(t) = 2t, \qquad \alpha_4(t) = 0.5\cos(\pi t/2) + 0.6,$$
$$\alpha_2(t) = (1-t)^3 + 0.2, \qquad \alpha_5(t) = 0.3\sin(\pi t/3) + 0.5.$$

where $\alpha_0(t)$ was the intercept function and $\alpha_1(t)$-$\alpha_5(t)$ were SNPs coefficients. We simulated a total of 5 variables of SNPs G and for each SNP, we first randomly picked one value p from uniform (0.1, 0.4) as the frequency of the major allele for an SNP. Then we sampled 0, 1 and 2 with probability p^2, $2p(1-p)$ and $(1-p)^2$ to obtain homozygous, heterozygous, and other homozygous genotype respectively.

In order to include confounding effects, we generated $\boldsymbol{\epsilon_1}$ and $\boldsymbol{\epsilon_2}$ simultaneously by assuming a variance-covariance matrix. In this simulation, we only considered one type of variance-covariance structure auto-regressive order 1(AR-1). The covariance matrix $\boldsymbol{\Sigma} = cov(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ was specified as follows: The entry in Σ_{11} and Σ_{22} was set to be $0.1 \times (0.5)^{|i-j|}$ for $i, j = 1, \cdots, 20$.

For Σ_{12} and Σ_{21}, the off-diagonal element was generated to be $0.1 \times (0.1)^{|i-j|}$, while the diagonal entry was set as 0.02. Then we simulate $(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) \sim N_{40}(\mathbf{0}, \Sigma)$. $\boldsymbol{X}$ was then generated by $\boldsymbol{X} = \boldsymbol{\alpha(t)G} + \boldsymbol{\epsilon_1}$ and $\boldsymbol{Y}$ was simulated by $\boldsymbol{Y} = \boldsymbol{\beta(t)X(t)} + \boldsymbol{\epsilon_2}$. To define the coefficient functions $\boldsymbol{\beta}(\cdot) = (\beta_0(\cdot), \beta_1(\cdot))^T$, we set $\beta_0(t) = 0.2t + 0.2$ and $\beta_1(t) = 0$ or $\beta_1(t) = 0.015 + 0.01t$ to investigate type I error and power respectively. Since the main goal of the simulation is to investigate the influence of instrumental variables strength on the type I error and empirical power, we considered three different simulation settings in total:

1. $\{\alpha_0(t), 0.5\alpha_1(t), 0.5\alpha_2(t), 0.5\alpha_3(t), 0.5\alpha_4(t), 0.5\alpha_5(t)\}$,
2. $\{\alpha_0(t), \alpha_1(t), \alpha_2(t), \alpha_3(t), \alpha_4(t), \alpha_5(t)\}$,
3. $\{\alpha_0(t), 2\alpha_1(t), 2\alpha_2(t), 2\alpha_3(t), 2\alpha_4(t), 2\alpha_5(t)\}$.

The simulation results are reported in Table A.1.

Table A.1 Simulation Results under different IV strength.

IV strength	Type I error	power
$0.5\alpha(t)$	0.049	0.217
$\alpha(t)$	0.052	0.609
$2\alpha(t)$	0.047	0.997

From Table A.1, changing the strength of instrumental variables has little impact on the Type I error which can be well protected at the $\alpha = 0.05$ level under all three simulation settings. For the empirical power simulation, the strength of genetic variants effects have significant influence on the simulation results. It is obvious to see the stronger the genetic variants are, the larger the empirical power. Thus, it is important to select valid and strong genetic variants as instrumental variables for Mendelian Randomization analysis. In traditional Mendelian Randomization study, sensitivity analysis is usually conducted to test the validity of the selected instrumental variables. Since we try longitudinal data in the study, further discussion about how to conduct sensitivity analysis is needed and will be our future work.

A.2 Data description of Albert Twin Data Subject

Table A.2 Subject Characteristics.

Characteristics	Descriptions
ConsensusFIN2	Describe menstrual cycle phases, ranging from 1-8
est	Hormone estradiol level measurements
pro	Hormone progesterone level measurements
PANNA	Negative emotional eating effect measured with the Negative Affect scale from the Positive and Negative Affect Schedule (PANAS).
DEBQE	Emotional eating measured with the Dutch Eating Behavior Questionnaire (DEBQ)
bmi	Describe subjects' BMI observations
FamID	Describe family ID, twins have same FamID
TwinID	Describe number of twins in each family: 1 means the first twin in the family; 2 means the second twin in the family
zyg	Describe types of twins: zyg=1 monozygotic or identical (MZ) twins; zyg=2 dizygotic, fraternal or non-identical (DZ) twins
agetwin	Describe subjects' age, ranging between 15-26
StudyDay	Describe the day of measurement, ranging from 1-45
FID	Describe family ID, twins have same FamID
IID	Describe subject ID
SNP(167,509)	Genotype encoding 0,1,2 corresponds to the number of minor allele in the genotype

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Proof of Theorem 3.3.2

To prove Theorem 3.3.2, we need the following conditions:

A1 The design time points t_j, $j = 1, \cdots, T_i$ are independent and identically distributed random variables following a probability density function $f(t)$ and t is a continuous point of f in the interior of the support of f.

A2 The kernel function $K(\cdot)$ is a bounded symmetric probability density function with bounded support [-1,1].

A3 The variance of $x(t_{j-q-s})$, $x(t_{j-q-p-s'})$, $\{x(t_{j-q-s})x(t_{j-q-p-s'})\}$ and the expected values of $x(t_{j-q-p-s'})$, $\{x(t_{j-q-s})x(t_{j-q-p-s'})\}$ are finte for all $s, s' = 0, \cdots, p-1$.

A4 $E(\epsilon_i^2(t_j))$ and σ_y^2 are finite.

A5 $\{\epsilon_i(t_j)\}$ is a zero-mean strongly mixing sequence of random variables with covariance function $\delta(t, t') = cov\{\epsilon_i(t), \epsilon_i(t')\}$.

A6 $\bar{\rho}_1^* < 1$, where $\bar{\rho}_1^* = \sup \rho(\sigma(\epsilon_i(t_j), j \in J), \sigma(\epsilon_i(t_{j'}), j' \in J'))$, and J, J' are nonempty subsets such that $dist(J, J') \geq 1$.

Let $\xi_j = b_{rj} - \beta_{rj}$, $a_j = w(t_j, t) = e_{1,p'+1}^T (C^T W C)^{-1} C_j W_j$, where $C_j = (1, (t_j - t), \cdots, (t_j - t)^{p'})^T$ and $W_j = K_h(t_j - t)$, $C = (C_1, C_2, \cdots, C_T)^T$, $W = diag(W_1, \cdots, W_T)$. Let $\sigma_T^2 = var(\sum_{j=1}^T a_j \xi_j)$. ξ_j has the following form:

$$\xi_j = b_{rj} - \beta_{rj} = c_{r,p}^T (X_{j-p}^T M_{j-p,j} X_j)^{-1} X_{j-p}^T M_{j-p,j} \{\epsilon_j + e_{yj}\}$$

where $c_{r,p}$ denotes a p-dimensional unit vector with 1 at its rth entry.

From condition A5, $\{\xi_j\}$ is a strongly mixing sequence and it is easy to get $E(\xi_j) = 0$.

In order to show $\{\xi_j^2\}$ is a uniformly integrable family, it's enough to show for a finite collection $\mathcal{T} = \{1, \cdots, T\}$, $E(|\xi_j^2|) < \infty$ for each $j \in \mathcal{T}$.

$$E(|\xi_j^2|) = c_{r,p}^T (X_{j-p}^T M_{j-p,j} X_j)^{-1} X_{j-p}^T M_{j-p,j} M_{j,j-p} X_{j-p} (X_{j-p}^T M_{j-p,j} X_j)^{-1} c_{r,p} \{E(\epsilon_i^2(t_j)) + \sigma_y^2\}$$

is finite under conditions A3 and A4. Also, it is obvious to see $E(\xi_j^2)$ is always positive, thus $\inf_j E(\xi_j^2) > 0$.

Suppose conditions A1 and A2 hold. If $h \to 0$ and $Th \to \infty$ as $T \to \infty$, it's easy to show

$$|a_j| = |w(t_j, t)| \leq \sum_{j=1}^{T} |w(t_j, t)| \leq (Th)^{1/2} \{\sum_{j=1}^{T} w^2(t_j, t)\}^{1/2} = (Th)^{1/2} \{O(Th)^{-1}\}^{1/2} = O(1),$$

then $max_{1 \leq j \leq T} \frac{|a_j|}{\sigma_T} \to 0$, as $T \to \infty$.

Under Condition A6, applying Magda's result(On the Asymptotic Normality of Sequences of Weak Dependent Random Variables), we could get $\frac{1}{\sigma_T} \sum_{j=1}^{T} a_j \xi_j \xrightarrow{\mathcal{D}} N(0, 1)$, as $T \to \infty$. Thus, $\hat{\beta}_r(t) = \sum_{j=1}^{T} w_r(t_j, t) b_r(t_j)$ is asymptotic Gaussian process with mean function $E(\hat{\beta}_r(t))$ and covariance function $\gamma_\beta(t_i, t_k)$. $E(\hat{\beta}_r(t))$ and $\gamma_\beta(t_i, t_k)$ are defined as follows:

$$E(\hat{\beta}_r(t)) = \sum_{j=1}^{T} w_r(t_j, t) \beta_r(t_j)$$

and

$$\gamma_\beta(t_i, t_k) = cov(\hat{\beta}_r(t_i), \hat{\beta}_r(t_k)) = \sum_{j=1}^{T} \sum_{j'=1}^{T} w_r(t_j, t_i) w_r(t_{j'}, t_k) cov(b_r(t_j), b_r(t_{j'})).$$

APPENDIX C

APPENDIX FOR CHAPTER 4

C.1 Simulation studies to test the influence of within sample correlation and confounding factors

To test the effect of confounding factors, we did additional simulations as we did in Chapter 4 but changed the value of confounding effects. In the simulation studies, we generated $\boldsymbol{\epsilon_1}$ and $\boldsymbol{\epsilon_2}$ simultaneously by assuming a variance-covariance matrix to include the effect of confounding factors. The covariance matrix $\boldsymbol{\Sigma} = cov(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ was specified as follows: The entry in Σ_{11} and Σ_{22} was set to be $(0.5)^{|i-j|}$ for $i, j = 1, \cdots, 20$, and for Σ_{12} and Σ_{21}, the off-diagonal element was generated to be $(0.1)^{|i-j|}$, while the diagonal entry was set as 0.2. Then we simulated $(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}) \sim N_{40}(\mathbf{0}, \Sigma)$.

Two different settings were considered in total to change the variance-covariance matrix: in the first setting, we fixed the off-diagonal matrix Σ_{12} and Σ_{21} and changed the structure of diagonal matrix Σ_{11} and Σ_{22}; in the second setting, we fixed diagonal matrix Σ_{11} and Σ_{22} and changed the off-diagonal matrix Σ_{12} and Σ_{21}. In the first setting, Σ_{12} and Σ_{21} was generated to be $(0.1)^{|i-j|}$ and the diagonal entry was set as 0.2. For the entry in Σ_{11} and Σ_{22}, we let it to be $\delta^{|i-j|}$ for $i, j = 1, \cdots, 20$, where δ was chosen as 0.3 and 0.7 respectively. The results were reported in Table C.1. Type I error rate could still be controlled at the $\alpha = 0.05$ significance level under all cases in Table C.1. For the empirical power simulation, we did not observe significant difference when the time points were large enough ($T = 10$ or $T = 20$). However, when we only had five repeated measurements, the empirical power increased with the increase of correlation δ. In addition, the difference was more obvious when the sample size was 200 compared to the 400 sample size.

Table C.1 Effect of within sample correlation on Type I error and power for functional MR model.

δ	n	T	Type I error	power
0.3	200	5	0.055	0.393
		10	0.055	0.724
		20	0.054	0.752
	400	5	0.046	0.694
		10	0.055	0.949
		20	0.053	0.955
0.5	200	5	0.054	0.428
		10	0.055	0.739
		20	0.046	0.764
	400	5	0.055	0.714
		10	0.055	0.964
		20	0.051	0.956
0.7	200	5	0.052	0.465
		10	0.054	0.761
		20	0.053	0.763
	400	5	0.050	0.733
		10	0.055	0.954
		20	0.052	0.972

In the second setting, we fixed diagonal matrix Σ_{11} and Σ_{22} and changed the off-diagonal matrix Σ_{12} and Σ_{21}. The entry in Σ_{11} and Σ_{22} was set to be $(0.5)^{|i-j|}$ for $i, j = 1, \cdots, 20$, and for Σ_{12} and Σ_{21}, the off-diagonal element was generated to be $\rho^{|i-j|}$, while the diagonal entry was set as $(\rho + 0.1)$, where we considered three different ρ values: 0.1, 0.3 and 0.5. We showed the simulation results in Table C.2. Similar to the previous simulation, Type I error rate could be well controlled at the $\alpha = 0.05$ significance level under all cases in Table C.2. For the empirical power simulation, they all had similar performance under different settings. The difference was much smaller compared to the difference in Table C.1.

Table C.2 Effect of confounding on Type I error and power for functional MR model.

ρ	n	T	Type I error	power
0.1	200	5	0.054	0.428
		10	0.055	0.739
		20	0.046	0.764
	400	5	0.055	0.714
		10	0.055	0.964
		20	0.051	0.956
0.3	200	5	0.054	0.468
		10	0.055	0.749
		20	0.053	0.747
	400	5	0.051	0.702
		10	0.053	0.961
		20	0.054	0.969
0.5	200	5	0.053	0.436
		10	0.052	0.731
		20	0.055	0.742
	400	5	0.052	0.714
		10	0.054	0.954
		20	0.049	0.956

In summary, the Type I error is well protected under all simulation settings. For the empirical power, changing the diagonal matrix Σ_{11} and Σ_{22} have more significant influence on the power compared with changing the off-diagonal matrix Σ_{12} and Σ_{21}.

C.2 Simulation studies for functional F test

In this section, we used functional F test to test the functional coefficient $\beta_1(t)$. The functional F test was introduced in chapter 4. One advantage of functional F test is that we consider the permutation test and we no longer need to rely on the distributional assumption. The simulation settings were the same to the settings in section C.1. We still included two different simulations: changing the structure of diagonal matrix Σ_{11} and Σ_{22}, and changing the off-diagonal matrix Σ_{12} and Σ_{21}. In the first simulation, we generated Σ_{11} and Σ_{22} using AR-1 model with different correlation δ values. The δ value was assumed to be 0.3, 0.5, 0.7 respectively. In the second simulation, the off-diagonal matrix Σ_{12} and Σ_{21} were also generated using AR-1 model with different correlation ρ values. In this case, we let ρ to be 0.1, 0.3, 0.5 respectively. We reported the simulation results

under different settings in Table C.3 and Table C.4 respectively.

Table C.3 Effect of within sample correlation on Type I error and power using functional F test.

δ	n	T	Type I error	power
0.3	200	5	0.043	0.382
		10	0.045	0.786
		20	0.045	0.773
	400	5	0.046	0.689
		10	0.051	0.968
		20	0.049	0.982
0.5	200	5	0.049	0.397
		10	0.050	0.748
		20	0.050	0.772
	400	5	0.051	0.687
		10	0.045	0.925
		20	0.047	0.984
0.7	200	5	0.052	0.391
		10	0.053	0.735
		20	0.055	0.729
	400	5	0.049	0.713
		10	0.051	0.926
		20	0.049	0.974

From Table C.3, Type I error rate could be well controlled at the $\alpha = 0.05$ significance level under all situations in the first simulation. For the empirical power simulation, we still could not observe significant difference no matter how many data points were included for analysis and no matter how many subjects were used for the hypothesis testing. This result is a little different from the results in section C.1 when using Wald test. In section C.1, the empirical power increased with the increase of correlation ρ when we only had five repeated measurements, and the difference was more obvious when the sample size was 200 compared to the 400 sample size. However, the results from functional F test did not show us obvious difference not only for five repeated measurements but also for enough time series observations.

Table C.4 Effect of confounding on Type I error and power using functional F test.

ρ	n	T	Type I error	power
0.1	200	5	0.049	0.397
		10	0.050	0.748
		20	0.050	0.772
	400	5	0.051	0.687
		10	0.045	0.925
		20	0.047	0.984
0.3	200	5	0.052	0.404
		10	0.054	0.736
		20	0.052	0.732
	400	5	0.055	0.701
		10	0.045	0.980
		20	0.047	0.993
0.5	200	5	0.049	0.395
		10	0.052	0.721
		20	0.050	0.697
	400	5	0.051	0.704
		10	0.046	0.922
		20	0.047	0.958

Similar conclusions could be drawn from Table C.4. In this table, Type I error rate could still be well protected. For the empirical power simulation, we did not observe significant difference under all different settings, indicating that the inclusion of the IVs can lead to reasonable power for causal inference regardless of the underlying confounding level. In addition, the empirical power increased with the increase of sample size. As the results suggested, five data points were not good enough for the hypothesis testing. Including at least 10 time points could substantially improve the testing power.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527, 2000.

[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[3] Joshua D Angrist and Jorn-Steffen Pischke. Instrumental variables in action: sometimes you get what you need. *Mostly harmless econometrics: an empiricist's companion*, pages 113–220, 2009.

[4] Marc F Bellemare, Takaaki Masaki, and Thomas B Pepinsky. Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics*, 79(3):949–963, 2017.

[5] Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

[6] Kenneth A Bollen and Judea Pearl. Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer, 2013.

[7] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

[8] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

[9] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665, 2013.

[10] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[11] Stephen Burgess and Simon G Thompson. Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in medicine*, 30(11):1312–1323, 2011.

[12] Stephen Burgess and Simon G Thompson. *Mendelian randomization: methods for using*

genetic variants in causal estimation. CRC Press, 2015.

[13] Stephen Burgess, Simon G Thompson, and Crp Chd Genetics Collaboration. Avoiding bias from weak instruments in mendelian randomization studies. *International journal of epidemiology*, 40(3):755–764, 2011.

[14] S Alexandra Burt and Kelly L Klump. The michigan state university twin registry (msutr): an update. *Twin Research and Human Genetics*, 16(1):344–350, 2013.

[15] S Alexandra Burt and Kelly L Klump. The michigan state university twin registry (msutr): 15 years of twin and family research. *Twin Research and Human Genetics*, 22(6):741–745, 2019.

[16] Ying Cao, Suja S Rajan, and Peng Wei. Mendelian randomization analysis of a time-varying exposure for binary disease outcomes using functional data analysis methods. *Genetic epidemiology*, 40(8):744–755, 2016.

[17] Tianjiao Chu, Clark Glymour, and Greg Ridgeway. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(5), 2008.

[18] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978.

[19] Jeff B Cromwell and Michel Terraza. *Multivariate tests for time series models*. Number 100. Sage, 1994.

[20] John L Czajka, Sharon M Hirabayashi, Roderick JA Little, and Donald B Rubin. Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, 10(2):117–131, 1992.

[21] Russell Davidson, James G MacKinnon, et al. *Estimation and inference in econometrics*, volume 63. Oxford New York, 1993.

[22] Neil M Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in medicine*, 34(3):454–468, 2015.

[23] Phoebus J Dhrymes. *Econometrics: Statistical foundations and applications*. Springer Science & Business Media, 2012.

[24] Chongzhi Di, Ciprian M Crainiceanu, and Wolfgang S Jank. Multilevel sparse functional principal component analysis. *Stat*, 3(1):126–143, 2014.

[25] Vanessa Didelez, Sha Meng, and Nuala A Sheehan. Assumptions of iv methods for observational epidemiology. *Statistical Science*, 25(1):22–40, 2010.

[26] Sizhen Du, Guojie Song, Lei Han, and Haikun Hong. Temporal causal inference with time lag. *Neural Computation*, 30(1):271–291, 2017.

[27] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

[28] Michael Eichler. Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1997):20110613, 2013.

[29] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[30] Jianqing Fan, Qiwei Yao, and Zongwu Cai. Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 65(1):57–80, 2003.

[31] Jianqing Fan and J-T Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):303–322, 2000.

[32] Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *The annals of Statistics*, 27(5):1491–1518, 1999.

[33] E Michael Foster. Instrumental variables for logistic regression: an illustration. *Social Science Research*, 26(4):487–504, 1997.

[34] Edward W Frees et al. *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press, 2004.

[35] Irving John Good. A causal calculus (i). *The British journal for the philosophy of science*, 11(44):305–318, 1961.

[36] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[37] Clive WJ Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

[38] Chirok Han. Detecting invalid instruments using l1-gmm. *Economics Letters*, 101(3):285–287, 2008.

[39] F Hartwig, G Davey Smith, and J Bowden. summary data mendelian randomisation via the zero modal pleiotropy assumption. international journal of epidemiology, 46 (6), 1985-1998.[dyx102]. https://doi. org/10.1093/ije/dyx102. *International Journal of Epidemiology*,

1:14, 2017.

[40] Audinga-Dea Hazewinkel, Rebecca C Richmond, Kaitlin H Wade, and Padraig Dixon. Mendelian randomization analysis of the causal impact of body mass index and waist-hip ratio on rates of hospital admission. *Economics & Human Biology*, 44:101088, 2022.

[41] Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human molecular genetics*, 27(R2):R195–R208, 2018.

[42] Joseph W Hogan and Tony Lancaster. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13(1):17–48, 2004.

[43] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[44] Aapo Hyvärinen, Shohei Shimizu, and Patrik O Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proceedings of the 25th international conference on Machine learning*, pages 424–431, 2008.

[45] Kosuke Imai, Gary King, and Elizabeth A Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2):481–502, 2008.

[46] T Johnson. Efficient calculation for multi-snp genetic risk scores technical report. *The Comprehensive R Archive Network*, 2013.

[47] John Kendall. Designing a research project: randomised controlled trials and their principles. *Emergency medicine journal: EMJ*, 20(2):164, 2003.

[48] Kelly L Klump and S Alexandra Burt. The michigan state university twin registry (msutr): Genetic, environmental and neurobiological influences on behavior across development. *Twin Research and Human Genetics*, 9(6):971–977, 2006.

[49] Kelly L Klump, Pamela K Keel, Sarah E Racine, S Alexandra Burt, Michael Neale, Cheryl L Sisk, Steven Boker, and Jean Yueqin Hu. The interactive effects of estrogen and progesterone on changes in emotional eating across the menstrual cycle. *Journal of abnormal psychology*, 122(1):131, 2013.

[50] Kelly L Klump, Sarah E Racine, Britny Hildebrandt, S Alexandra Burt, Michael Neale, Cheryl L Sisk, Steven Boker, and Pamela K Keel. Influences of ovarian hormones on dysregulated eating: A comparison of associations in women with versus women without binge episodes. *Clinical Psychological Science*, 2(5):545–559, 2014.

[51] KL Klump, SE Racine, B Hildebrandt, SA Burt, M Neale, CL Sisk, S Boker, and PK Keel. Ovarian hormone influences on dysregulated eating: A comparison of associations between women with versus without binge episodes. *Clinical Psychological Science*, 2(5):545–559, 2014.

[52] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

[53] David J Lederer, Scott C Bell, Richard D Branson, James D Chalmers, Rachel Marshall, David M Maslove, David E Ost, Naresh M Punjabi, Michael Schatz, Alan R Smyth, et al. Control of confounding and reporting of results in causal inference studies. guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society*, 16(1):22–28, 2019.

[54] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[55] Herbert Matschinger, Dirk Heider, and Hans-Helmut König. A comparison of matching and weighting methods for causal inference based on routine health insurance data, or: what to do if an rct is impossible. *Das Gesundheitswesen*, 82(S 02):S139–S150, 2020.

[56] I Miko. Gregor mendel and the principles of inheritance. *Nature Education*, 1(1):134, 2008.

[57] Whitney K Newey. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of econometrics*, 36(3):231–250, 1987.

[58] Chao Ning, Huimin Kang, Lei Zhou, Dan Wang, Haifei Wang, Aiguo Wang, Jinluan Fu, Shengli Zhang, and Jianfeng Liu. Performance gains in genome-wide association studies for longitudinal traits via modeling time-varied effects. *Scientific reports*, 7(1):1–12, 2017.

[59] Tom M Palmer, Michael V Holmes, Brendan J Keating, and Nuala A Sheehan. Correcting the standard errors of 2-stage residual inclusion estimators for mendelian randomization studies. *American journal of epidemiology*, 186(9):1104–1114, 2017.

[60] Judea Pearl. *Causality*. Cambridge university press, 2009.

[61] Gina-Maria Pomann, Ana-Maria Staicu, and Sujit Ghosh. A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(3):395–414, 2016.

[62] Annie Qu and Runze Li. Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, 62(2):379–391, 2006.

[63] Annie Qu, Bruce G Lindsay, and Bing Li. Improving generalised estimating equations using

quadratic inference functions. *Biometrika*, 87(4):823–836, 2000.

[64] Annie Qu and Peter X-K Song. Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika*, 91(2):447–459, 2004.

[65] Ridho Rahmadi, Perry Groot, Marieke HC van Rijn, Jan AJG van den Brand, Marianne Heins, Hans Knoop, Tom Heskes, Alzheimer's Disease Neuroimaging Initiative, MASTER-PLAN Study Group, and OPTIMISTIC consortium. Causality on longitudinal data: Stable specification search in constrained structural equation modeling. *Statistical Methods in Medical Research*, 27(12):3814–3834, 2018.

[66] William Robert Reed. On the practice of lagging variables to avoid simultaneity. *Oxford Bulletin of Economics and Statistics*, 77(6):897–905, 2015.

[67] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[68] Peggy Sekula, M Fabiola Del Greco, Cristian Pattaro, and Anna Köttgen. Mendelian randomization as an approach to assess causality using observational data. *Journal of the American Society of Nephrology*, 27(11):3253–3265, 2016.

[69] Damla Şentürk and Hans-Georg Müller. Generalized varying coefficient models for longitudinal data. *Biometrika*, 95(3):653–666, 2008.

[70] Koichiro Shiba and Takuya Kawahara. Using propensity scores for causal inference: pitfalls and tips. *Journal of epidemiology*, page JE20210145, 2021.

[71] Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. arxiv 2021. *arXiv preprint arXiv:2105.02675*, 2021.

[72] Anton P Sidawy and Bruce A Perler. *Rutherford's vascular surgery and endovascular therapy, E-Book*. Elsevier Health Sciences, 2018.

[73] Peter X-K Song, Zhichang Jiang, Eunjoo Park, and Annie Qu. Quadratic inference functions in marginal models for longitudinal data. *Statistics in medicine*, 28(29):3683–3696, 2009.

[74] Ana-Maria Staicu, Yingxing Li, Ciprian M Crainiceanu, and David Ruppert. Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, 41(4):932–949, 2014.

[75] Joan G Staniswalis and J Jack Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418, 1998.

[76] Daniel I Swerdlow, Karoline B Kuchenbaecker, Sonia Shah, Reecha Sofat, Michael V Holmes,

Jon White, Jennifer S Mindell, Mika Kivimaki, Eric J Brunner, John C Whittaker, et al. Selecting instruments for mendelian randomization in the wake of genome-wide association studies. *International journal of epidemiology*, 45(5):1600–1616, 2016.

[77] John R Thompson, Cosetta Minelli, Jack Bowden, Fabiola M Del Greco, Dipender Gill, Elinor M Jones, Chin Yang Shapland, and Nuala A Sheehan. Mendelian randomization incorporating uncertainty about pleiotropy. *Statistics in Medicine*, 36(29):4627–4645, 2017.

[78] Tyler J VanderWeele, John W Jackson, and Shanshan Li. Causal inference and longitudinal data: a case study of religion and mental health. *Social psychiatry and psychiatric epidemiology*, 51(11):1457–1466, 2016.

[79] Abraham Wald. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300, 1940.

[80] Honglang Wang, Jingyi Zhang, Kelly L Klump, Sybil Alexandra Burt, and Yuehua Cui. Multivariate partial linear varying coefficients model for gene-environment interactions with multiple longitudinal traits. *Statistics in Medicine*, 2022.

[81] Honglang Wang, Ping-Shou Zhong, Yuehua Cui, and Yehua Li. Unified empirical likelihood ratio tests for functional concurrent linear models and the phase transition from sparse to dense functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):343–364, 2018.

[82] Qiyao Wang. Two-sample inference for sparse functional data. *Electronic Journal of Statistics*, 15(1):1395–1423, 2021.

[83] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[84] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005.