Title
Bayesian Approaches for Instrumental Variable Analysis with Censored Time-to-Event Outcome

Author
Lu, Xuyang

Publication Date
2014

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Bayesian Approaches for Instrumental Variable Analysis with Censored Time-to-Event Outcome

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Xuyang Lu

2014

Abstract of the Dissertation

Bayesian Approaches for Instrumental Variable Analysis with Censored Time-to-Event Outcome

by

Xuyang Lu

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2014

Professor Gang Li, Chair

The method of instrumental variable (IV) analysis has been widely used in economics, epidemiology, and other fields to estimate the causal effects of intermediate covariates on outcomes, in the presence of unobserved confounders and/or measurement errors in covariates. Consistent estimation of the effect has been developed when the outcome is continuous, while methods for binary outcome produce inconsistent estimation. In this dissertation, we examine two IV methods in the literature for binary outcome and show the bias in parameter estimate by a simulation study. The identifiability problem of IV analysis with binary outcome is discussed. Moreover, IV methods for time-to-event outcome with censored data remain underdeveloped. We propose two Bayesian approaches for IV analysis with censored time-to-event outcome by using a two-stage linear model: One is a parametric Bayesian model with normal and non-normal elliptically contoured error distributions, and the other is a semiparametric Bayesian model with Dirichlet process mixtures for the random errors, in order to relax the parametric assumptions and address heterogeneous clustering problems. Markov Chain Monte Carlo sampling methods are developed for both paramet-

ric and semiparametric Bayesian models to estimate the endogenous parameter. Performance of our methods is examined by simulation studies. Both methods largely reduce bias in estimation and greatly improve coverage probability of the endogenous parameter, compared to the regular method where the unobserved confounders and/or measurement errors are ignored. We illustrate our methods on the Women's Health Initiative Observational Study and the Atherosclerosis Risk in Communities Study.

The dissertation of Xuyang Lu is approved.

Cho-Lea Tso

Donatello Telesca

Guido Eibl

Ron Brookmeyer

Gang Li, Committee Chair

University of California, Los Angeles

2014

This dissertation is dedicated to my grandmother, Liang Xiaorong, whom I miss dearly everyday.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deep gratitude to my advisor, Professor Gang Li, for his constructive guidance, great encouragement and generous support. I sincerely appreciate his constant kindness and patience. He is my role model on both professional and personal level. This dissertation would not have been possible without him.

I am also grateful to Professor Ronald Brookmeyer, Professor Donatello Telesca, Professor Cho-Lea Tso and Professor Guido Eibl for kindly serving on my doctoral committee. Professor Brookmeyer provided many insightful and valuable suggestions to improve this dissertation. Professor Telesca's expertise in Bayesian analysis has been extremely beneficial to our work, and I greatly appreciate his serious attention in discussing our research. Professor Tso offers a valuable perspective on my work in terms of its application, and I am grateful for her support in my job search. Professor Eibl has been a wonderful collaborator, and I am very thankful for his participation in my final defense committee at the last moment.

I also want to sincerely thank Professor Simin Liu for motivating this research and for providing the WHI-OS data.

Last but not least, I want to thank my parents Wanqin and Yuxun, my sister Yanbin, and my wife Huidong Liu for their love and support during this process.

Vita

2006	B.S. (Mathematics), Zhejiang University, Hangzhou, China.
2008	M.A. (Statistics), Bowling Green State University, Bowling Green, OH, USA.

CHAPTER 1

Introduction

1.1 Introduction to Instrumental Variable Analysis

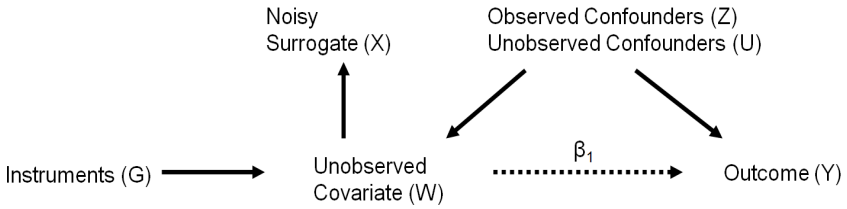
In statistical analysis, estimating the causal effects of covariates on outcomes are often of interest. For example, in epidemiological studies, the causal effect of a modifiable phenotype or exposure on a disease outcome is usually more important than the mere association between the two variables. However, a randomized control trial (RCT), which offers the best ability to make a causal inference, is not always possible due to ethical or other reasons, whilst inferring causation from an observational study is often difficult. Moreover, the problem of unobserved confounders is very common in observational studies: the presence of unknown or unmeasured confounders could lead to a spurious association between the covariate and the outcome. Furthermore, the covariates could be subject to measurement errors, which could bias the estimated association towards the null.

Instrumental variable (IV) analysis is a statistical tool that has been extensively used by economists, epidemiologists, and others to provide an alternative approach to the following three objectives: (1) To estimate the causal effect of covariates on outcomes; (2) To address the bias caused by unobserved confounders; (3) To account for measurement errors in the covariates. It involves using an extra variable, the *instrumental variable* or *instrument*, that satisfies certain criteria.

As shown in Figure 1.1, the primary variables in an IV analysis are: outcome

variable Y; true value of intermediate covariate W, which could be unobserved due to measurement errors; observed surrogate X of the covariate W; observed confounders Z; unobserved confounders U; and instruments G. The primary aim is to estimate the causal effect of W on Y, indicated by the dotted line in Figure 1.1. The instrument vector G is selected based on three criteria: (1) G is independent of U and measurement errors in W; (2) G is associated with W; (3) G is independent of Y given W, i.e. G is associated with Y only through W.

Figure 1.1: Directed acyclic graph of instrumental variable analysis



IV analysis has a long history dated back to early 20th century. An extensive literature has been developed in the field of economics for IV analysis with continuous outcomes (for example, Wright, 1928; Haavelmo, 1943, 1944; Theil, 1958; Goldberger, 1972; Bowden and Turkington, 1984; Heckman and Robb, 1985; Heckman and Hotz, 1989; Morgan, 1991). The covariate X is called an *endogenous variable* especially in economics, meaning that it is correlated to the error term in the model regressing outcome Y on X. IV analysis became popular in the field of epidemiology in the past two decades, especially with the special case of *Mendelian Randomization* (MR), where genetic markers are used as the instrument G (e.g. Davey Smith and Ebrahim, 2004; Thomas and Conti, 2004; Didelez and Sheehan, 2007; Lawlor et al., 2008; Wehby et al., 2008). The term was first used by Gray and Wheatley (1991), though its meaning has evolved over time (Davey Smith and Ebrahim, 2003). It refers to the similarities between random assortment of genes following the Mendel's law of independent assortment and

random treatment allocation in a randomized control trial (Nitsch et al., 2006). In MR analysis, Y is usually a disease related outcome, X is usually a modifiable exposure or phenotype, G is a vector of genetic instruments. Therefore, MR is also known as IV analysis using genetic instruments (Wehby et al., 2008). Genetic markers arise as a natural choice of instrument due to the fact that they are generally independent of typical confounders such as behavioral and environmental factors. The selection of genetic instruments is mostly based on biological and clinical information, since IV assumptions (1) and (3) are difficult to be validated directly. It is worth noting that for IV assumption (2), G does not have to be causal in the association between G and W to be an IV: The association could be due to a mediator variable that affects both G and W (e.g. when there is linkage disequilibrium between G and another genotype) (Ogbuanu et al., 2009).

The classic IV analysis estimates the endogenous parameter indirectly, by estimating the association between G and Y and the association between G and X. This is usually done by using a two-stage least squares (TSLS) estimation in a simultaneous equation model (Wright, 1928; Theil, 1958, among others). Variance estimation of IV estimates were derived based on the TSLS procedure, including the Murphy-Topel estimator (Murphy and Topel, 1985) and the Huber/White/Sandwich estimator (Huber, 1967; White, 1980). Specific forms of the two variance estimators in IV analysis were given by Hardin (2002) and Hardin and Carroll (2003). Details of variance estimation are presented in section 2.1.3. An alternative way to draw inference on variance is to use posterior variance and credible interval from a Bayesian model. Bayesian IV methods were developed for continuous outcomes, based on a more general form of the IV model and the assumption that the error terms follow a bivariate normal distribution (Kleibergen and Van Dijk, 1998; Hoogerheide et al., 2007). Nice reviews of Bayesian approaches for normal linear IV models, as well as their advantages and disadvan-

tages compared to the classical frequentist approaches, are given by Kleibergen and Zivot (2003) and Lancaster (2004). Conley et al. (2008) further developed a semiparametric Bayesian approach using Dirichlet process prior and showed that it is more efficient when the error terms are non-normal.

Causal inference, deconfounding and measurement error correction are the three main objectives of IV analysis, which have been well-established for continuous outcome. Angrist and Imbens (1995) and Angrist et al. (1996) showed that the TSLS can be used to estimate the average causal effect of the covariate, and that the IV estimand can be embedded within the Rubin Causal Model, which is a well-established framework of causal inference with observational data (Rubin, 1974, 1978; Holland, 1986). Didelez and Sheehan (2007) presented a formal framework for causal inference based on the MR approach. More comprehensive illustrations on causal inference based on IV analysis can be found in Pearl (2000), Heckman (2008) and the references therein. Didelez and Sheehan (2007) also proved the consistency of endogenous parameter estimation in the presence of unobserved confounders, based on a classic IV model where outcome Y and covariate W are both continuous, and a threshold IV model for situation when covariate W and instrument G are both binary (Bowden and Turkington, 1984). Details of consistency for the two models are presented in sections 2.1.1 and 2.1.4, respectively. Moreover, IV analysis can be used to consistently estimate the endogenous parameter when the true value of the covariate W is not observed and a noisy surrogate X is observed instead. An extensive literature on IV analysis focusing on the aspect of measurement error correction can be found in Durbin (1954), Fuller (1987), Carroll and Stefanski (1994), Buzas and Stefanski (1996), Goetghebeur and Vansteelandt (2005), Carroll et al. (2006), Gustafson (2007), and the references therein.

For binary outcomes, such as the presence or absence of some disease, Y is usually modeled by a logistic or probit regression model. Despite its common usage in recent medical studies (e.g. Brunner et al., 2008; Ding et al., 2009; Elliott et al., 2009; Kamstrup et al., 2009; Kivimäki et al., 2011; Thanassoulis et al., 2013), the classic IV estimation is no longer consistent when there is an non-null effect of covariate W on the binary outcome Y (Didelez and Sheehan, 2007). Palmer et al. (2008) investigated the extent of bias in classic IV analysis for binary outcomes by simulations, and showed that the unobserved confounders tend to bias the IV estimates of endogenous parameter towards null, i.e. the effect of W on Y tends to be underestimated. McKeigue et al. (2010) proposed a simplified model that ignores measurement errors in W (i.e assuming $X = W$), ignores the random error term in W, and assumes that all variation in W given instrument G can be explained by a univariate unobserved confounder. This assumption is strong and unlikely to be realistic, and violation of the assumption could bias the endogenous parameter estimate. We conduct a simulation study based on logistic regression in section 2.2.2 to show the bias from this simplified model and the classic IV model. When the random error in W is non-ignorable, the simplified model is similar to a logistic regression with binary outcome and measurement error in the covariate. Kuchenhoff (1995) proves the theoretical identifiability of this model, but also points out that it is not numerically identifiable. We further discuss the identifiability problem of IV analysis with binary outcome in section 2.2.3. We examine the likelihood function of the IV model using an arbitrary data and graphically show that the unique maximum likelihood estimate (MLE) of the endogenous parameter can not be numerically detected even if it existed. This suggests that the endogenous parameter can not be consistently estimated by methods based on likelihood, including MLE and Bayesian methods without informative priors.

It is often of interest in observational studies to investigate the effect of some intermediate covariate on time to the occurrence of events such as diseases and death, while the same issues of causal inference, unobserved confounders and/or measurement errors need to be considered. The problem can be further complicated by censoring in the time-to-event outcome. For example, in a case-control observational study, the outcome of time from baseline to some disease is right-censored for the controls, who are disease-free at the end of the study. By using instruments satisfying corresponding assumptions, the IV approach can be accommodated into survival analysis and provide a possible solution to this problem. The built-in Stata command "ivtobit" (StataCorp, 2011) and user-written Stata command "cmp" (Roodman, 2011) can be used to fit linear IV model with a tobit model for the censored time-to-event data. They use a maximum likelihood approach by assuming a bivariate normal error distribution. Bijwaard Bijwaard (2008) proposed an IV Linear Rank estimator for right-censored time-to-event data, based on a Generalized Accelerated Failure Time model. In addition, generalized method of moments (GMM) estimators (Hansen, 1982; Hall, 2005; Yin et al., 2011, among others) could potentially be used to draw parametric maximum likelihood inference for IV analysis with survival outcome. Nevertheless, IV methods for survival analysis remain underdeveloped, and yet to our knowledge no Bayesian IV model for survival outcome has been formally developed.

In this dissertation, we develop two Bayesian approaches for IV analysis with censored time-to-event outcome, based on a two-stage linear model. A general form of the IV model is used so that all model parameters are identifiable. This is done by jointly modelling the random error term in the time-to-event outcome and the random error term in the continuous covariate. The first approach is a parametric Bayesian model with a bivariate normal or other parametric bivariate distribution such as an elliptically contoured distribution. (see section 3.2.4 for a

detailed review of literature). The second approach is a semiparametric Bayesian model using Dirichlet process mixtures (DPM) for the random errors, in order to relax the parametric assumptions and address heterogeneous clustering problems. Instead of using a pre-specified number of mixture components, the DPM model allows the number of mixture components to be determined by both the prior and the data (see section 4.1 for a detailed review of literature). By using a Bayesian approach, prior information (such as estimates and confidence intervals of the parameters from former studies) could be incorporated by using informative priors. Markov Chain Monte Carlo (MCMC) sampling methods are developed for computation and estimation. Performance of the two approaches is evaluated by simulation studies, under frequentist criteria of bias, standard deviation and coverage probability.

We show that both proposed methods largely reduce bias in estimation and greatly improves coverage probability of the endogenous variable parameter, compared to the simple method where the unobserved confounders and measurement errors are ignored. Moreover, the parametric Bayesian approach with bivariate normal errors appears to be fairly robust against violation of the parametric assumption. The semiparametric Bayesian approach with DPM performs as well as the parametric Bayesian approach when the errors are bivariate normal, but has a higher precision of parameter estimates than the parametric Bayesian approach when the errors are non-normal. We illustrate our method using two real data sets: the Women's Health Initiative Observational Study (WHI-OS) and the Atherosclerosis Risk in Communities (ARIC) Study . In the WHI-OS, we investigate the effect of high-sensitivity C-reactive protein (hsCRP) on time to diagnosis of diabetes, using multiple selected single-nucleotide polymorphisms (SNPs) as genetic instruments. In the ARIC study, we examine the effect of systolic blood pressure (SBP) on time to diagnosis of coronary heart disease (CHD), using the

IV model to correct for measurement errors in SBP. MCMC convergence diagnostics are conducted for the real data analyses. Although we focus on their applications in epidemiology for simplicity of illustration, these two approaches are generally applicable to any IV analysis with censored time-to-event outcome.

1.2 Research Outline

The rest of the dissertation is organized as follows. Chapter 2 consists of two sections: Section 2.1 introduces the existing work for IV analysis with continuous outcome. Section 2.2 presents two methods in the literature for IV analysis with binary outcome and shows the bias in parameter estimation of the two methods by a simulation study. The identifiability problem for IV method with binary outcome is discussed. Chapter 3 describes our parametric Bayesian model for IV analysis with censored time-to-event outcome, along with the MCMC procedure. It includes a simulation study to examine the performance of the model, and application of the model to two real data examples to illustrate the method. Chapter 4 describes our semiparametric Bayesian model with Dirichlet Process Mixtures. A simulation study comparing the performance of the two models is conducted. We illustrate the method by its application to the two real data examples. Chapter 5 contains some remarks and possible future extensions.

CHAPTER 2

Instrumental Variable Analysis with Continuous Outcome and Binary Outcome

2.1 Continuous Outcome

In this section, we present the existing work for IV analysis with continuous outcome. Consistency property of the classic IV estimates for the endogenous variable parameter is justified. The two-stage least squares (TSLS) method is introduced. Two variance estimators of the IV estimates are presented. A threshold model is introduced for situations where covariate W and instrument G are both binary: The same classic IV approach can be applied, and corresponding consistency is justified.

2.1.1 Consistency

For each subject i, let Y_i be the continuous outcome variable, W_i be an unobserved continuous covariate of interest that is subject to measurement errors, X_i be the observed surrogate of W_i, Z_i be a vector of observed confounders, U_i be a vector of unobserved confounders, and G_i be a vector of instruments, $i = 1, \ldots, n$, where n is the total number of subjects. The variables follow the structure shown in Figure 1.1. For simplicity of illustration, in this section we assume instrument G_i is univariate, and we ignore the observed confounders Z_i, which can be easily

incorporated in the classic IV estimation for continuous outcomes.

Classical IV analysis assumes linear relationships between the instrument G, the covariate W and the outcome Y. With the structure in Figure 1.1, the variables can be modeled by the following three linear equations when Y, W and X are continues:

$$W_i = \alpha_0 + \alpha_1 G_i + \alpha_2{}' U_i + \varepsilon_{1i} \tag{2.1}$$

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2{}' U_i + \varepsilon_{2i} \tag{2.2}$$

$$X_i = W_i + \varepsilon_{3i} \tag{2.3}$$

$i = 1, \ldots n$, where ε_{1i}, ε_{2i}, and ε_{3i} are independent random errors with mean zero and finite variances τ_1^2, τ_2^2, and τ_3^2 respectively. Note that ε_{3i} is the measurement error in the intermediate covariate. The unobserved confounder vector U_i is standardized to have mean 0 and covariance matrix Σ_U, where Σ_U is the correlation matrix of U_i. The variables ε_{1i}, ε_{2i}, ε_{3i}, U_i and G_i are assumed to be independent. β_1 is the endogenous parameter and it is the parameter of primary interest.

Since W_i is not observed, we can replace W_i in equations (2.1) and (2.2) by $W_i = X_i - \varepsilon_{3i}$ from equation (2.3) and have the following two-stage linear model:

$$X_i = \alpha_0 + \alpha_1 G_i + \alpha_2{}' U_i + \varepsilon_{1i} + \varepsilon_{3i} \tag{2.4}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2{}' U_i + \varepsilon_{2i} - \beta_1 \varepsilon_{3i} \tag{2.5}$$

The standard way to consistently estimate β_1 is through a linear regression of Y on W and U together. This becomes problematic when W and/or U are unobserved due to measurement errors and/or unobserved confounders. Moreover, a linear regression of Y on X alone will result in biased estimate of β_1, since X is correlated with U and ε_3. Instead, we consider (1) a linear regression of Y on G, and (2) a linear regression of X on G. By substituting W_i in equation (2.2)

with equation (2.1):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 W_i + \beta_2{}' U_i + \varepsilon_{2i} \\ &= \beta_0 + \beta_1(\alpha_0 + \alpha_1 G_i + \alpha_2{}' U_i + \varepsilon_{1i}) + \beta_2{}' U_i + \varepsilon_{2i} \\ &= (\beta_0 + \beta_1\alpha_0) + \alpha_1\beta_1 G_i + (\beta_1\alpha_2{}' + \beta_2{}')U_i + (\beta_1\varepsilon_{1i} + \varepsilon_{2i}) \\ &= \eta_0 + \eta_1 G_i + (\beta_1\alpha_2{}' + \beta_2{}')U_i + \beta_1\varepsilon_{1i} + \varepsilon_{2i} \end{aligned} \tag{2.6}$$

where $\eta_0 = \beta_0 + \beta_1\alpha_0$ and $\eta_1 = \alpha_1\beta_1$. Based on equation (2.6), since G is independent of U, ε_1 and ε_2, the regression coefficient of G in a linear regression of Y on G alone by an ordinary least squares (OLS), denoted as $\hat{\eta}_1$, is an unbiased estimate of $\eta_1 = \alpha_1\beta_1$, due to the well-known unbiasedness property of OLS estimation. Similarly, based on equation (2.4), since G is independent of U, ε_1 and ε_3, the regression coefficient of G in a linear regression of X on G alone by an ordinary least squares (OLS), denoted as $\hat{\alpha}_1$, is an unbiased estimate of α_1. Hence, β_1 can be consistently estimated by the ratio of the two unbiased estimates:

$$\hat{\beta}_1 = \frac{\hat{\eta}_1}{\hat{\alpha}_1} \tag{2.7}$$

The following is a more detailed proof of (1) the bias resulted from regressing Y on X alone, and (2) the unbiasedness of $\hat{\eta}_1$ and $\hat{\beta}_1$:

Based on equation (2.5), the expectation of Y given X is:

$$\begin{aligned} E(Y|X=x) &= E_U E_{\varepsilon_3} E(Y|X=x, U, \varepsilon_3) \\ &= \beta_0 + \beta_1 x + \beta_2{}' E_{\varepsilon_3} E(U|X=x, \varepsilon_3) - \beta_1 E_U E(\varepsilon_3|X=x, U) \\ &= \beta_0 + \beta_1 x + \beta_2{}' E(U|X=x) - \beta_1 E(\varepsilon_3|X=x) \end{aligned} \tag{2.8}$$

Let $\beta = (\beta_0, \beta_1)'$ and Z be the design matrix. From a linear regression of Y on

X alone by OLS, expectation of estimate of β estimate is:

$$\begin{aligned} E(\hat{\beta}) &= (Z'Z)^{-1}Z'E(Y|X) \\ &= (Z'Z)^{-1}Z'(Z\beta + \beta_2{}'E(U|X) - \beta_1 E(\varepsilon_3|X)) \\ &= \beta + (Z'Z)^{-1}Z'\beta_2{}'(E(U|X) - \beta_1 E(\varepsilon_3|X)) \end{aligned} \tag{2.9}$$

Even though $E(U) = E(\varepsilon_3) = 0$, $E(U|X)$ and $E(\varepsilon_3|X)$ are usually nonzero. Therefore, the second part of (2.9) is the bias of $\hat{\beta}$.

Instead, based on equation (2.6), the expectation of Y given G is:

$$\begin{aligned} E(Y|G=g) &= E_{(W,U)|G=g}E(Y|W,U,G=g) \\ &= E_{U|G=g}E_{W|U,G=g}E(Y|W,U) \qquad \text{since } Y \perp G|(W,U) \\ &= E_U E_{W|U,G=g}(\beta_0 + \beta_1 W + \beta_2{}'U) \\ &= E_U(\beta_0 + \beta_1(\alpha_0 + \alpha_1 g + \alpha_2{}'U) + \beta_2{}'U) \\ &= \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 g + (\beta_1\alpha_2{}' + \beta_2{}')E(U) \\ &= (\beta_0 + \beta_1\alpha_0) + \beta_1\alpha_1 g \\ &= \eta_0 + \eta_1 g \end{aligned} \tag{2.10}$$

Therefore, similar to the argument for equation (2.8), let $\eta = (\eta_0, \eta_1)'$, Z be the design matrix and $\hat{\eta}$ be the parameter estimate, we have $E(\hat{\eta}) = (Z'Z)^{-1}Z'E(Y|G) = \eta$. Therefore, $\hat{\eta}_1$ is an unbiased estimate of $\eta_1 = \beta_1\alpha_1$.

Furthermore, based on equation (2.4), the expectation of X given G is:

$$\begin{aligned} E(X|G=g) &= E_{(U,\varepsilon_3)|G=g}E(X|G=g,U,\varepsilon_3) \\ &= E_{(U,\varepsilon_3)}E(X|G=g,U,\varepsilon_3) \qquad \text{since } U \perp G,\ \varepsilon_3 \perp G \\ &= \alpha_0 + \alpha_1 g + \alpha_2{}'E(U) + E(\varepsilon_3) \\ &= \alpha_0 + \alpha_1 g \end{aligned} \tag{2.11}$$

Similarly, let $\alpha = (\alpha_0, \alpha_1)'$, Z be the design matrix and $\hat{\alpha}$ be the parameter estimate, we have $E(\hat{\alpha}) = (Z'Z)^{-1}Z'E(X|G) = \alpha$. Therefore, $\hat{\alpha}_1$ is an unbiased estimate of α_1.

Here we ignore observable confounders or other covariates of interest, as they can easily be incorporated into the models according to the assumed structure of the variables. In particular, if some confounder independent of instrument G is observed and added into the model, i.e. adjusting for the observed confounder in both OLS regressions, the variation in estimates $\hat{\alpha}_1$ and $\hat{\eta}_1$ will be reduced and therefore result in reduced variance for the IV estimate $\hat{\beta}_1$. Moreover, there can be no interactions with the unobserved confounder U and intermediate covariate W. Since U is unknown, this is obviously an untestable assumption.

2.1.2 Two-Stage Least Squares

The consistent estimate of β_1 from IV analysis is a ratio of two parameter estimates from two different models. It is difficult to derive a variance estimate for a ratio in this form. Another approach of classic IV analysis for continuous outcome is the Two-Stage Least Squares (TSLS) estimation. When G is univariate, the TSLS estimation of β_1 is identical to the ratio estimate described in section 2.1.1. The procedure is as follows:

1) Obtain unbiased estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ by a linear regression of X on G;

2) Find the predicted value of X, denoted as $\tilde{X}$, by $\tilde{X} = \hat{\alpha}_0 + \hat{\alpha}_1 G$;

3) Obtain the consistent estimate of β_1 by a linear regression of Y on $\tilde{X}$.

This can be summarized by a two-step regression model:

$$E(X_i) = \alpha_0 + \alpha_1 G_i \tag{2.12}$$

$$E(Y_i) = \beta_0 + \beta_1 E(X_i|G_i) \tag{2.13}$$

This following is a detailed proof to show that the TSLS estimate of β_1 is identical to the ratio estimate from (2.7). This can be shown by the explicit formula of the OLS estimator. Let $\mathbf{1}$ be a vector of length n with all elements equal 1, and $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)'$ be the coefficient estimate from the regression of Y on $\tilde{X}$. We are trying to show that $\tilde{\beta}_1 = \hat{\beta}_1 = \hat{\eta}_1/\hat{\alpha}_1$. Now let G, X, and Y be vectors of $\{G_1, ..., G_n\}$, $\{X_1, ..., X_n\}$, and $\{Y_1, ..., Y_n\}$, respectively. Based on the OLS estimation,

$$\begin{aligned}
\begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{pmatrix} &= \left[\begin{pmatrix} \mathbf{1}' \\ G' \end{pmatrix} \begin{pmatrix} \mathbf{1} & G \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{1}' \\ G' \end{pmatrix} X \\
&= \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'G \\ G'\mathbf{1} & G'G \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{1}'G \\ G'X \end{pmatrix} \\
&= \frac{1}{(\mathbf{1}'\mathbf{1}G'G - \mathbf{1}'GG'\mathbf{1})} \begin{pmatrix} G'G & -\mathbf{1}'G \\ -G'\mathbf{1} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{1}'X \\ G'X \end{pmatrix} \\
&= \frac{1}{(\mathbf{1}'\mathbf{1}G'G - \mathbf{1}'GG'\mathbf{1})} \begin{pmatrix} G'G\mathbf{1}'X - \mathbf{1}'GG'X \\ -G'\mathbf{1}\mathbf{1}'X + \mathbf{1}'\mathbf{1}G'X \end{pmatrix}
\end{aligned} \tag{2.14}$$

Similarly,

$$\begin{pmatrix} \hat{\eta}_0 \\ \hat{\eta}_1 \end{pmatrix} = \frac{1}{(\mathbf{1}'\mathbf{1}G'G - \mathbf{1}'GG'\mathbf{1})} \begin{pmatrix} G'G\mathbf{1}'Y - \mathbf{1}'GG'Y \\ -G'\mathbf{1}\mathbf{1}'Y + \mathbf{1}'\mathbf{1}G'Y \end{pmatrix}$$

Thus, by the ratio approach,

$$\hat{\eta}_1/\hat{\alpha}_1 = \frac{-G'\mathbf{1}\mathbf{1}'Y + \mathbf{1}'\mathbf{1}G'Y}{-G'\mathbf{1}\mathbf{1}'X + \mathbf{1}'\mathbf{1}G'X} \tag{2.15}$$

For the TSLS estimate,

$$\begin{aligned}
\begin{pmatrix}\tilde{\beta}_0\\ \tilde{\beta}_1\end{pmatrix} &= \left[\begin{pmatrix}\mathbf{1}\\ \hat{\alpha}_0\cdot\mathbf{1}'+\hat{\alpha}_1\cdot G'\end{pmatrix}\begin{pmatrix}\mathbf{1} & \hat{\alpha}_0\cdot\mathbf{1}+\hat{\alpha}_1\cdot G\end{pmatrix}\right]^{-1}\begin{pmatrix}\mathbf{1}'\\ \hat{\alpha}_0\cdot\mathbf{1}'+\hat{\alpha}_1\cdot G'\end{pmatrix}Y\\
&= \begin{pmatrix}\mathbf{1}'\mathbf{1} & \hat{\alpha}_0\mathbf{1}'\mathbf{1}+\hat{\alpha}_1\mathbf{1}'G\\ \hat{\alpha}_0\mathbf{1}'\mathbf{1}+\hat{\alpha}_1G'\mathbf{1} & \hat{\alpha}_0^2\mathbf{1}'\mathbf{1}+2\hat{\alpha}_0\hat{\alpha}_1\mathbf{1}'G+\hat{\alpha}_1^2G'G\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}'Y\\ \hat{\alpha}_0\mathbf{1}'Y+\hat{\alpha}_1G'Y\end{pmatrix}\\
&= \frac{1}{\hat{\alpha}_1^2(\mathbf{1}'\mathbf{1}G'G-\mathbf{1}'GG'\mathbf{1})}\begin{pmatrix}\hat{\alpha}_0^2\mathbf{1}'\mathbf{1}+2\hat{\alpha}_0\hat{\alpha}_1\mathbf{1}'G+\hat{\alpha}_1^2G'G & -\hat{\alpha}_0\mathbf{1}'\mathbf{1}-\hat{\alpha}_1\mathbf{1}'G\\ -\hat{\alpha}_0\mathbf{1}'\mathbf{1}-\hat{\alpha}_1G'\mathbf{1} & \mathbf{1}'\mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{1}'Y\\ \hat{\alpha}_0\mathbf{1}'Y+\hat{\alpha}_1G'Y\end{pmatrix}\\
&= \frac{1}{\hat{\alpha}_1^2(\mathbf{1}'\mathbf{1}G'G-\mathbf{1}'GG'\mathbf{1})}\begin{pmatrix}\hat{\alpha}_0\hat{\alpha}_1(\mathbf{1}'G\mathbf{1}'Y-\mathbf{1}'\mathbf{1}G'Y)+\hat{\alpha}_1^2(G'G\mathbf{1}'Y-\mathbf{1}'GG'Y)\\ \hat{\alpha}_1(\mathbf{1}'\mathbf{1}G'Y-G'\mathbf{1}1'Y)\end{pmatrix}
\end{aligned}\tag{2.16}$$

Hence, by (2.14), (2.15) and (2.16),

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\mathbf{1}'\mathbf{1}G'Y-G'\mathbf{1}\mathbf{1}'Y}{\hat{\alpha}_1(\mathbf{1}'\mathbf{1}G'G-\mathbf{1}'GG'\mathbf{1})}\\
&= \frac{\mathbf{1}'\mathbf{1}G'Y-G'\mathbf{1}\mathbf{1}'Y}{\mathbf{1}'\mathbf{1}G'G-\mathbf{1}'GG'\mathbf{1}}\cdot\frac{\mathbf{1}'\mathbf{1}G'G-\mathbf{1}'GG'\mathbf{1}}{\mathbf{1}'\mathbf{1}G'X-G'\mathbf{1}\mathbf{1}'X}\\
&= \frac{\mathbf{1}'\mathbf{1}G'Y-G'\mathbf{1}\mathbf{1}'Y}{\mathbf{1}'\mathbf{1}G'X-G'\mathbf{1}\mathbf{1}'X}\\
&= \hat{\eta}_1/\hat{\alpha}_1\\
&= \hat{\beta}_1
\end{aligned}$$

This completes the proof of equivalence of the two estimates. We assume all 2×2 matrices above are nonsingular. The same result can be easily derived using Pseudoinverse if the matrices are singular.

We can see the equivalence of the two estimates in a more general framework of maximum likelihood. Suppose the first IV estimate is based on maximizing the likelihood of a generalized linear model (GLM):

$$E(y_i)=h^{-1}(\eta_0+\eta_1 g_i)\tag{2.17}$$

and given OLS estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ from a linear model $E(x_i)=\alpha_0+\alpha_1 g_i$, where $h^{-1}(\cdot)$ is the link function. Let $\hat{\eta}_0$ and $\hat{\eta}_1$ be the corresponding Maximum

Likelihood Estimators (MLEs) from GLM (2.17). The second IV estimate is based on maximizing the likelihood of a GLM:

$$E(y_i) = h^{-1}(\beta_0 + \beta_1 \tilde{x}_i) \tag{2.18}$$

where $\tilde{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 g_i$. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the corresponding MLEs from GLM (2.18). We want to show that $\tilde{\beta}_1 = \hat{\eta}_1/\hat{\alpha}_1$.

Let $\mathcal{L}(\cdot)$ denotes the likelihood function.

$$\begin{aligned}
(\tilde{\beta}_0, \tilde{\beta}_1) &= \underset{(\beta_0,\beta_1)}{\operatorname{argmax}} \mathcal{L}(\beta_0, \beta_1 | y, \tilde{x}, \hat{\alpha}_0, \hat{\alpha}_1) \\
&= \underset{(\beta_0,\beta_1)}{\operatorname{argmax}} \mathcal{L}(\beta_0, \beta_1 | y, \hat{\alpha}_0 + \hat{\alpha}_1 g, \hat{\alpha}_0, \hat{\alpha}_1) \\
&= \underset{(\beta_0,\beta_1)}{\operatorname{argmax}} \mathcal{L}(\beta_0 + \hat{\alpha}_0\beta_1, \hat{\alpha}_1\beta_1 | y, g, \hat{\alpha}_0, \hat{\alpha}_1) \\
&= \underset{(\eta_0,\eta_1)}{\operatorname{argmax}} \mathcal{L}(\eta_0, \eta_1 | y, g, \hat{\alpha}_0, \hat{\alpha}_1) \cdot \left|\frac{d\beta}{d\eta}\right| \quad \text{where } \eta_0 = \beta_0 + \hat{\alpha}_0\beta_1, \eta_1 = \hat{\alpha}_1\beta_1 \\
&= (\hat{\eta}_0 - \frac{\hat{\alpha}_0}{\hat{\alpha}_1}\hat{\eta}_1, \frac{\hat{\eta}_1}{\hat{\alpha}_1})
\end{aligned} \tag{2.19}$$

Therefore, when the MLE is unique, we have $\tilde{\beta}_1 = \hat{\eta}_1/\hat{\alpha}_1$. Equivalence of the two estimates holds.

The two-step regression approach also provides a solution for calculating the endogenous parameter estimates by IV analysis when there are multiple instrumental variables, multiple intermediate covariates, and/or multiple continuous outcomes. Similar to equations (2.1) and (2.2), it can be modeled by two multivariate linear regression:

$$X_i = \alpha_0 + \alpha_1{}' G_i + \alpha_2{}' U_i + \varepsilon_{1i} + \varepsilon_{3i} \tag{2.20}$$

$$Y_i = \beta_0 + \beta_1{}' X_i + \beta_2{}' U_i + \varepsilon_{2i} \tag{2.21}$$

Again, the parameter matrix β_1 is of primary interest. Similar to the TSLS procedure described earlier, a consistent estimate of α_1 can be generated by a

multivariate linear model of X on G alone. Thus, a consistent estimate of β_1 can be generated by a multivariate linear model of Y on $\tilde{X}$, which is a matrix of the predicted value of X based on the first-step model (2.20).

2.1.3 Variance Estimation

Although the TSLS procedure in section 2.1.2 leads to a consistent estimate of β_1, the estimated covariance matrix for the second-step model (2.13) needs to be adjusted to take into account the variability in $\hat{\alpha}_0$ and $\hat{\alpha}_1$, since $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are estimates from the first-step model (2.12) other than their true values. Ignoring the fact that $\tilde{X}$ is estimated by using $\hat{\alpha}_0$ and $\hat{\alpha}_1$ will understate the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

There are two standard approaches for variance estimation with the TSLS procedure. The first estimator is proposed by Murphy and Topel (1985) based on a limited information maximum likelihood (LIML) two-step procedure. The LIML estimation fits a first-step model, which is then used to estimate regression coefficients for a second-step model of primary interest. The instrumental variable approach in section 2.1.2 is a special case of the LIML two-step procedure when the likelihood functions of models (2.12) and (2.13) are known. A general formula of the Murphy-Topel estimator for two-stage models is described by Hardin (2002) and Hole (2006).

We denote $\alpha = (\alpha_0, \alpha_1)'$ and $\beta = (\beta_0, \beta_1)'$. Based on the two-step regression model in the previous section, the Murphy-Topel estimate for β is given by

$$\hat{V}_2 + \hat{V}_2(\hat{C}\hat{V}_1\hat{C}^T - \hat{R}\hat{V}_1\hat{C}^T - \hat{C}\hat{V}_1\hat{R}^T)\hat{V}_2$$

where $\hat{V}_1$ and $\hat{V}_2$ are the estimated covariance matrices for regression models (2.12) and (2.13), respectively, where each is the model-based estimate not taking into

account the fact that the estimate of α is embedded in model (2.13). They can be calculated by

$$\begin{aligned}
\hat{V}_1 &= \left[-\sum_{i=1}^{n} \left(\frac{\partial^2 \ln f_{i1}}{\partial \alpha \partial \alpha^T} \right) \right]^{-1} \Bigg|_{\alpha = \hat{\alpha}} \\
\hat{V}_2 &= \left[-\sum_{i=1}^{n} \left(\frac{\partial^2 \ln f_{i2}}{\partial \beta \partial \beta^T} \right) \right]^{-1} \Bigg|_{\alpha = \hat{\alpha}, \beta = \hat{\beta}}
\end{aligned}$$

where f_{i1} and f_{i2} are observation i's contribution to the likelihood function of models (2.12) and (2.13), respectively. Note that likelihood function f_1 involves parameter α only.

Further,

$$\begin{aligned}
\hat{C} &= \sum_{i=1}^{n} \left(\frac{\partial \ln f_{i2}}{\partial \beta} \right) \left(\frac{\partial \ln f_{i2}}{\partial \alpha^T} \right) \Bigg|_{\alpha = \hat{\alpha}, \beta = \hat{\beta}} \\
\hat{R} &= \sum_{i=1}^{n} \left(\frac{\partial \ln f_{i2}}{\partial \beta} \right) \left(\frac{\partial \ln f_{i1}}{\partial \alpha^T} \right) \Bigg|_{\alpha = \hat{\alpha}, \beta = \hat{\beta}}
\end{aligned}$$

This implies that the Murphy-Topel estimate exceeds the naive variance estimate from the second-step model, $\hat{V}_2$, by a positive-definite matrix. Under the assumption that the first-step model produces consistent estimates of both first-step parameters and their asymptotic covariance matrix, the covariance matrix of the second-step parameter estimates can be consistently estimated by the Murphy-Topel estimator.

An alternative to the Murphy-Topel estimator is the Huber/White/Sandwich estimator (Huber, 1967; White, 1980). Hardin (2002) derives explicit formula of the sandwich variance estimator for the two-step procedure. Based on our two-step model, we assume the first-step model has an estimating equation $\Psi_1(\alpha)$, and the second-step model has an estimating equation $\Psi_2(\beta|\alpha)$. The overall

estimating equation can be partitioned as

$$[\Psi(\alpha,\beta)] = \begin{bmatrix} \Psi_1(\alpha) \\ \Psi_2(\beta|\alpha) \end{bmatrix} = [0]$$

The sandwich estimate of variance for the complete parameter vector $(\alpha', \beta')'$ is given by $V_S = A^{-1}BA^{-T}$, where

$$\begin{aligned} A &= \begin{bmatrix} \dfrac{\partial\Psi_1}{\partial\alpha^T} & \dfrac{\partial\Psi_1}{\partial\beta^T} \\ \dfrac{\partial\Psi_2}{\partial\alpha^T} & \dfrac{\partial\Psi_2}{\partial\beta^T} \end{bmatrix} \\ B &= \begin{bmatrix} \Psi_1\Psi_1^T & \Psi_1\Psi_2^T \\ \Psi_2\Psi_1^T & \Psi_2\Psi_2^T \end{bmatrix} \end{aligned}$$

Assuming that valid log-likelihood functions f_1 and f_2 exist for the two models, the estimating equations are derivatives of the model log-likelihoods,

$$\begin{aligned} \Psi_1(\alpha) &= \sum_{i=1}^{n} \frac{\partial \ln f_{i1}(\alpha)}{\partial\alpha} \\ \Psi_2(\beta|\alpha) &= \sum_{i=1}^{n} \frac{\partial \ln f_{i2}(\beta|\alpha)}{\partial\beta} \end{aligned}$$

Then the matrix elements of the sandwich estimator V_S are given by

$$\begin{aligned} V_S(\alpha) &= V_1 V_1^{*-1} V_1 = V_{S1} \\ Cov_S(\alpha,\beta) &= V_1 R^T V_2 - V_{S1} C^{*T} V_2 \\ V_S(\beta) &= V_2 V_2^{*-1} V_2 + V_2 \left(C^* V_1 V_1^{*-1} V_1 C^{*T} - R V_1 C^{*T} - C^* V_1 R^T \right) V_2 \\ &= V_{S2} + V_2 \left(C^* V_{S1} C^{*T} - R V_1 C^{*T} - C^* V_1 R^T \right) V_2 \end{aligned}$$

where the components are estimated by

$$\begin{aligned}
\hat{V}_1^* &= \left[\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \alpha}\right)\left(\frac{\partial \ln f_{i1}}{\partial \alpha^T}\right)\right]^{-1}\Bigg|_{\alpha=\hat{\alpha}} \\
\hat{V}_2^* &= \left[\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \beta}\right)\left(\frac{\partial \ln f_{i2}}{\partial \beta^T}\right)\right]^{-1}\Bigg|_{\alpha=\hat{\alpha},\,\beta=\hat{\beta}} \\
\hat{C}^* &= \sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{i2}}{\partial \beta \partial \alpha^T}\right)\Bigg|_{\alpha=\hat{\alpha},\,\beta=\hat{\beta}}
\end{aligned}$$

along with $\hat{V}_1$, $\hat{V}_2$ and $\hat{R}$, which are the same as in the Murphy-Topel estimator. The asterisks $(*)$ are used to distinguish similar matrix components in the Murphy-Topel estimator. $\hat{V}_{S1} = \hat{V}_1\hat{V}_1^{*-1}\hat{V}_1$ and $\hat{V}_{S2} = \hat{V}_2\hat{V}_2^{*-1}\hat{V}_2$ are sandwich estimators from the individual models (2.12) and (2.13). The sandwich estimator of primary interest is $\hat{V}_S(\beta)$. It is in a form that is similar to the Murphy-Topel estimator. The differences are the use of the sandwich estimators V_{S1} and V_{S2} from the individual models and the matrix of second derivatives C^* over the matrix of first derivative products C.

The sandwich estimator provides consistent estimates of the covariance matrix for the parameter estimates when the fitted parametric model is incorrect or not specified, as well as in the presence of heteroscedasticity. It is sometimes called the *robust covariance matrix estimator*, the *heteroskedasticity-consistent covariance matrix estimator*, or the *empirical covariance matrix estimator* due to its desirable model-robustness property. For the two-step model with equations (2.4) and (2.5), the sandwich estimator is robust to the underlying distributions of the unobserved confounders U, measurement errors ε_3, and the random errors ε_1 and ε_2. Kauermann and Carroll (2001) investigates the performance of the sandwich estimator and concludes that the sandwich estimator generally have a larger variance than model-based classical variance estimates, which is the price

that one pays to obtain consistency. This increased variability in the variance estimates also causes the problem of undercoverage of confidence intervals.

The relationship between the Murphy-Topel estimator and the sandwich estimator was described in Hardin (2002). The two estimators are asymptotically equal when the assumed model distributions are true. The sandwich estimator is computationally more difficult than the Murphy-Topel estimator due to the use of C^*, which requires computing second derivatives of the second model's log likelihood. However, the robustness property of the sandwich estimator is appealing especially when the parametric assumptions are invalid. Moreover, the full sandwich variance matrix for (α, β) can be calculated. This allows Wald tests of hypothesis of the parameters across the two models, which are not possible using the Murphy-Topel estimator.

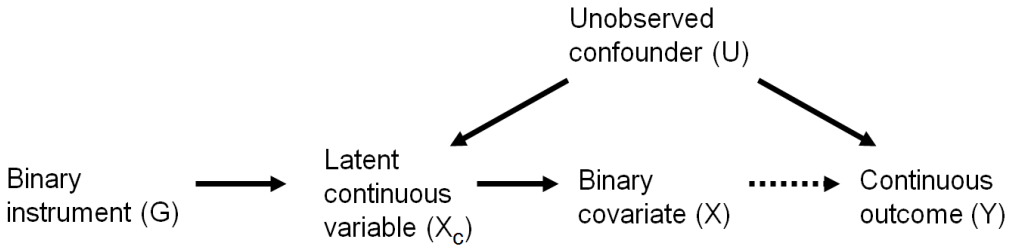
As indicated by the two variance estimators, variance of β_1 from the two-step models, as well as the inferences drawn from the estimates, will be extremely sensitive to the precision of the first-step estimates $\hat{\alpha}$. Besides sample size, the precision of $\hat{\alpha}$ is mainly determined by the correlation between instrument G and surrogate covariate X, which is an indicator of the instrumental strength of G.

Both the Murphy-Topel estimator and the sandwich estimator can be calculated by using Stata's **qvf** command for fitting generalized linear models with instrumental variables. A detailed illustration of the software command can be found in Hardin et al. (2003).

2.1.4 Binary Covariate and Binary Instrument

In this subsection, we assume that no measurement error in the intermediate covariate is involved, i.e. $X = W$. We also assume that the unobserved confounder U is univariate and standardized to have mean 0 and variance 1. When

Figure 2.1: Structure of instrumental variable analysis with binary covariate and binary instrument



both covariate X and instrument G are binary, the common method is to use a threshold model which assumes an underlying unobservable continuous variable X_c with linear conditional expectation:

$$E(X_c|G = g, U = u) = \alpha_0 + \alpha_1 g + \alpha_2 u$$

and define

$$X_i \;=\; \begin{cases} 1 & \text{if } X_c > 0 \\ 0 & \text{otherwise} \end{cases}$$

The dependence structure can be represented as in Figure 2.1, where the relationship between X_c and X is deterministic.

The conditional independencies are

$$Y \perp (G, X_c)|(U, X), \quad X \perp (G, U)|X_c \;\text{ and }\; G \perp U$$

However, since X_c is not observed, we have

$$Y \perp G|(U, X), \quad X \,\text{not}\perp G \;\text{ and }\; G \perp U$$

for the remaining variables. Therefore, the core conditions of instrumental variable apply to (G, U, X, Y) when X_c is ignored.

Without loss of generality, we assume $\alpha_2 > 0$. By an argument similar to that in section 2.1.1, we have

$$\begin{aligned} E(Y|G=g) &= E_U E_{X|U,G=g} E(Y|X,U) && \text{since } Y \perp G|(X,U),\ U \perp G \\ &= E_U E_{X|U,G=g}(\beta_0 + \beta_1 X + \beta_2 U) \\ &= E_U(\beta_0 + \beta_1 I(\alpha_0 + \alpha_1 g + \alpha_2 U > 0) + \beta_2 U) \\ &= \beta_0 + \beta_1 P_U\left(U > \frac{-\alpha_0 - \alpha_1 g}{\alpha_2}\right) \end{aligned} \tag{2.22}$$

where $I(\cdot)$ is the indicator function. We assume the binary instrument G takes value 0 or 1. Let

$$\mu_0 = \frac{-\alpha_0}{\alpha_2} \quad \text{and} \quad \mu_1 = \frac{-\alpha_0 - \alpha_1}{\alpha_2}$$

then model (2.22) can be written as

$$E(Y|G=g) = \beta_0 + \beta_1 P_U(U > \mu_0) + \beta_1(P_U(U > \mu_1) - P_U(U > \mu_0))g \tag{2.23}$$

Equation (2.23) is linear in G, thus the parameter of G:

$$\tau_{Y|G} = \beta_1(P_U(U > \mu_1) - P_U(U > \mu_0)) \tag{2.24}$$

can be consistently estimated by an OLS regression of Y on G.

Similarly,

$$\begin{aligned} E(X|G=g) &= E_U E(X|G=g,U) \\ &= E_U I(\alpha_0 + \alpha_1 g + \alpha_2 U > 0) \\ &= P_U(U > \mu_0) + (P_U(U > \mu_1) - P_U(U > \mu_0))g \end{aligned} \tag{2.25}$$

This is also linear in G, with parameter

$$\tau_{X|G} = P_U(U > \mu_1) - P_U(U > \mu_0) \tag{2.26}$$

consistently estimated by OLS regression of X on G. By (2.24) and (2.26), we can generate consistent estimate of β_1 by taking the ratio of the two coefficient estimates:

$$\hat{\beta}_1 = \frac{\hat{\tau}_{Y|G}}{\hat{\tau}_{X|G}} \tag{2.27}$$

This is identical to the IV approach when G, X and Y are all continuous.

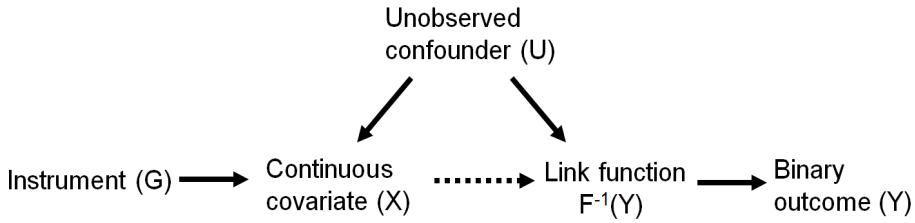
2.2 Binary Outcome

In this section, we first present two IV methods for binary outcome in the literature: one is the *ad hoc* method following the continuous outcome scenario, the other is based on a simplified model ignoring the random error term in the continuous intermediate covariate. We conduct a simulation study to show that both methods generate inconsistent estimates of the endogenous parameter when the true underlying model is subject to unobserved confounders and the random error term in the continuous covariate is not ignorable. We further examine the likelihood function and graphically show that the maximum likelihood estimate (MLE) of the endogenous parameter is not numerically identifiable. This indicates that the endogenous parameter is not practically estimable by methods solely based on the likelihood.

2.2.1 Inconsistent Estimation

The outcome variable in observational studies is often a binary indicator of disease status. For IV analysis with binary outcome, a logistic regression model or a probit regression model is usually used in replacement of equation(2.5). Figure 2.2 shows the structure of IV analysis with binary outcome, where $F^{-1}(\cdot)$ is a logit or probit link function. Here we focus on the situation with a logistic regression model, since a probit regression model has very similar properties of bias and unidentifiability. Also, here we ignore the potential measurement errors in the intermediate covariate and assume that X is an accurate measurement (i.e. $X = W$), and we assume unobserved confounder U is univariate. This is because a univariate unobserved confounder alone will cause the bias and identifiability problems.

Figure 2.2: Structure of instrumental variable analysis with binary outcome



The structure in Figure 2.2 with a logit link function can be modeled by:

$$X_i = \alpha_0 + \alpha_1 G_i + \alpha_2 U_i + \varepsilon_i \tag{2.28}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + \beta_2 U_i \tag{2.29}$$

where $p_i = \Pr(Y_i = 1)$, Y_i independently follows a Bernoulli distribution with probability p_i, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$. Again, the coefficient β_1 is of primary interest, as e^{β_1} is the disease odds ratio for each unit increment in X.

The classic IV analysis with binary outcome is based on an *ad hoc* method similar to the continuous outcome scenario. After substituting X_i in equation (2.29) with equation (2.28), we have

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \beta_1 X_i + \beta_2 U_i \\ &= \beta_0 + \beta_1(\alpha_0 + \alpha_1 G_i + \alpha_2 U_i + \varepsilon_i) + \beta_2 U_i \\ &= \eta_0 + \eta_1 G_i + v_i \end{aligned} \tag{2.30}$$

where $\eta_0 = \beta_0 + \beta_1\alpha_0$, $\eta_1 = \beta_1\alpha_1$, and $v_i = (\beta_1\alpha_2 + \beta_2)U_i + \beta_1\varepsilon$. The term v combines unobserved confounder U and random error ε. An unbiased OLS estimate of α_1, $\hat{\alpha}_1$, is derived by a linear regression of X on G, similar to section 2.1.1. An estimate of η_1, denoted as $\hat{\eta}_1$, is derived by a logistic regression of Y on G. The ratio $\hat{\eta}_1/\hat{\alpha}_1$ serves as an estimate of β_1. An equivalent method is similar

to the TSLS approach described in section 2.1.2, where β_1 is derived by a logistic regression of Y on the predicted value of X that is based on the linear regression of X on G ($\tilde{X}$, as in section 2.1.2).

Although this *ad hoc* method is the most commonly used method in the medical literature for IV analysis with binary outcomes (examples of recent publications include Brunner et al., 2008; Ding et al., 2009; Elliott et al., 2009; Kamstrup et al., 2009; Kivimäki et al., 2011; Thanassoulis et al., 2013), the estimate $\hat{\beta}_1$ is not a consistent estimate of β_1 when $\beta_1 \neq 0$. This is because $\hat{\eta}_1$, the MLE based on a logistic regression model ignoring v_i, is not a consistent estimate of $\eta_1 = \alpha_1\beta_1$, even though $E(v) = 0$ and $G \perp v$. Instead of maximizing the likelihood of the conditional model

$$\mathcal{L}_{cond} = \prod_{i=1}^{n} \frac{e^{(\eta_0+\eta_1 G_i+v_i)Y_i}}{1+e^{\eta_0+\eta_1 G_i+v_i}}$$

, the MLEs maximize the likelihood of the marginal model

$$\mathcal{L}_{marg} = \prod_{i=1}^{n} \frac{e^{(\eta_0+\eta_1 G_i)Y_i}}{1+e^{\eta_0+\eta_1 G_i}}$$

Without the linear relationship between X and Y, the MLEs for the two likelihood functions are generally different.

McKeigue et al. (2010) proposed another IV method for binary outcomes, using a simplified model that ignores the random error ε in X:

$$X_i = \alpha_0 + \alpha_1 G_i + \alpha_2 U_i \tag{2.31}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + \beta_2 U_i \tag{2.32}$$

where $E(U) = 0$ and $Var(U) = 1$. It makes a fairly strong assumption: all the variation in X that can not be explained by the instrument G is contributed by a univariate unobserved confounder U. Although McKeigue et al. proposed a

Bayesian approach by assuming U_i follows a normal distribution , a frequentist approach can be derived based on this simplified model:

1. Derive robust estimates of α_0 and α_1 by OLS estimate from a linear regression of X on G;
2. The residuals from step 1 are estimate of U_i: $\hat{U}_i = X_i - \hat{\alpha}_0 - \hat{\alpha}_1 G_i$;
3. An estimate of β_1 can be derived by a logistic regression of Y on X, while adjusting for $\hat{U}_i$ from step 2.

The estimate based on this simplified model relies on the strong and fairly unrealistic assumption that there is no random errors in X other than one confounder U. The estimate of β_1 will not be consistent if this assumption is invalid.

2.2.2 Bias Evaluation by a Simulation Study

As an illustration of the bias problem of the existing IV methods for binary outcome, we conduct a simulation study to evaluate the estimation bias caused by the *ad hoc* approach and the simplified model described in section 2.2.1.

We generate the data following the model with equations (2.28) and (2.29), where ε_i is independent and identically distributed (*i.i.d.*) as $\sim N(0,1)$, U_i *i.i.d.* $\sim N(0,1)$ and $\alpha_1 = 1$, i.e. unobserved confounder U and random error ε explain equal variation in covariate X. Instrument G is generated as *i.i.d.* Binomial$(2, 0.5)$, representing an additive genetic model of one locus with risk allele frequency equal 0.5. Thus, $E(G) = 1$ and $Var(G) = 1/2$. α_1 is set as 0.67, 1.00 and 1.31, corresponding to population R-square of 10%, 20% and 30% between X and G. We refer to these values as "weak", "moderate" and "strong", respectively. α_0 is arbitrarily set as 0.

For the second-stage equation (2.29), we use three values of β_1: 0, 0.5 and 1, giving odds ratios 1, 1.6 and 2.7. These values represent none, small and median effect, respectively, of covariate X on the disease outcome Y. β_2 is set as 1, indicating a moderate confounding effect. β_0 is set as $-\alpha_1\beta_1$, so the sample disease prevalence rate ($Pr(Y = 1)$) is approximately 0.5. Total sample size n takes values 100, 200, 500, 1000 and 1500. We applied both the *ad hoc* method and the frequentist approach for the simplified model to the simulated data. Each mean estimate is based on 10000 simulations. Results are summarized in table 2.1.

When covariate X has no effect on outcome Y, i.e. $\beta_1 = 0$, estimates from both approaches appear to be consistent, as the mean of the estimates reduce to close to 0 when the sample size n gets larger (bias$\leq$ 0.002 when $n \geq 1000$). For fixed sample size, the bias decreases as the instrument strength (population R-square between X and G) increases. However, the estimates from both approaches have substantial bias when β_1 is nonzero even when sample size is very large. Neither increase in instrument strength nor increase in sample size can effectively reduce the bias. Moreover, for the *ad hoc* approach, the bias/effect ratio increases as the effect β_1 gets larger, resulting in estimates that are more understated. Although the simplified model results in smaller biases than the *ad hoc* approach for our settings when β_1 is nonzero, we note that the results are sensitive to the strength of confounding effect and the proportions of variation in X contributed by U and ε.

Table 2.1: Simulation results for IV analysis with Binary Outcome

Performance of the of the *ad hoc* method and the simplified model under various sample sizes (n), true values of the intermediate covariate (β_1), and instrument strength ($R^2(X, G)$). Bias is calculated as the absolute value of the difference between the sample mean of the β_1 estimates and the true value of β_1. Each result is based on 10000 simulations.

		Estimation Bias of β_1					
		$\beta_1 = 0$		$\beta_1 = 0.5$		$\beta_1 = 1$	
$R^2(X, G)$	Sample size n	*ad hoc*	Simple	*ad hoc*	Simple	*ad hoc*	Simple
10%	100	-0.059	-0.064	-0.21	-0.107	-0.465	-0.105
	200	-0.023	-0.025	-0.168	-0.057	-0.441	-0.078
	500	-0.010	-0.011	-0.158	-0.046	-0.435	-0.080
	1000	-0.002	-0.002	-0.156	-0.044	-0.435	-0.082
	1500	-0.002	-0.002	-0.156	-0.044	-0.434	-0.080
20%	100	-0.025	-0.028	-0.16	-0.042	-0.424	-0.04
	200	-0.010	-0.011	-0.158	-0.044	-0.432	-0.067
	500	-0.003	-0.004	-0.155	-0.042	-0.431	-0.072
	1000	-0.001	-0.001	-0.155	-0.043	-0.432	-0.079
	1500	-0.001	-0.001	-0.154	-0.043	-0.432	-0.079
30%	100	-0.013	-0.015	-0.152	-0.032	-0.417	-0.032
	200	-0.003	-0.003	-0.154	-0.04	-0.427	-0.062
	500	-0.003	-0.003	-0.154	-0.041	-0.428	-0.073
	1000	-0.002	-0.002	-0.154	-0.042	-0.429	-0.076
	1500	-0.001	-0.001	-0.155	-0.043	-0.430	-0.078

2.2.3 Identifiability

As shown in the simulation study, the *ad hoc* method and the simplified model will generate inconsistent estimates of β_1 when the true underlying model has unobserved confounder U and random error ε in X. Therefore, it is desirable to derive a consistent estimate of β_1. However, by describing the pattern of the likelihood function from equations (2.28) and (2.29) graphically, we show that β_1 is not estimable by likelihood-based methods (including MLE and Bayesian methods) without additional information on σ^2, α_2 or the ratio of α_2^2/σ^2.

Following the two methods discussed in section 2.2.1, there are, in general, two ways to make use of the instrument G. The first one is to follow equation (2.30) and try to consistently estimate $\eta_1 = \beta_1\alpha_1$ by taking into account the random effect v_i. The second way is to substitute U_i in equation (2.29) by U_i in equation (2.28), similar to the frequentist approach for the simplified model. Both ways of estimation involve one random effect in the logistic regression model. We will use the second way to illustrate the identifiability issue, as the argument for the first way will be similar. In that case, equation (2.29) becomes:

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \beta_1 X_i + \beta_2 U_i \\ &= \beta_0 + \beta_1 X_i + \beta_2(X_i - \alpha_0 - \alpha_1 G_i - \varepsilon_i) \\ &= d_0 + d_1 X_i + d_2 G_i + d_3 \nu_i \end{aligned} \tag{2.33}$$

where $d_0 = \beta_0 - \alpha_0\beta_2$, $d_1 = \beta_1 + \beta_2$, $d_2 = -\alpha_1\beta_2$, $d_3 = -\beta_2\sigma$ and $\nu_i = \varepsilon_i/\sigma$, so $E(\nu_i) = 0$ and $Var(\nu_i) = 1$. A natural approach is to assume ν_i follows a parametric distribution, e.g. $\nu_i \sim N(0, 1)$, and find the MLE of (d_1, d_2, d_3, d_4) by maximizing the likelihood of model (2.33), then back-calculate the MLE of $(\beta_0, \beta_1, \beta_2, \sigma^2)$. Since the variable ν is a random error term that is not observed, the likelihood function involves marginalizing out ν.

Equation (2.33) can be viewed as a logistic regression where one covariate is subject to measurement error. Kuchenhoff (1995) proved that a logistic regression model with normal measurement error in one of the covariates is identifiable. By definition, a model parameter is identifiable if it is uniquely determined by the likelihood function. The whole model is called identifiable if all model parameters are identifiable. Following the proof in Kuchenhoff (1995), the parameters (d_1, d_2, d_3, d_4) is uniquely determined by:

$$\begin{aligned} d_1 &= \lim_{X\to\infty} \frac{\frac{\partial}{\partial X}Q(X,G)}{Q(X,G)} + \lim_{X\to-\infty} \frac{\frac{\partial}{\partial X}Q(X,G)}{Q(X,G)} \\ d_2 &= \lim_{G\to\infty} \frac{\frac{\partial}{\partial X}Q(X,G)}{Q(X,G)} + \lim_{G\to-\infty} \frac{\frac{\partial}{\partial X}Q(X,G)}{Q(X,G)} \\ d_0 &= -d_1 Q^{-1}(X = 1/2, G = 0) \\ d_3 &= -K^{-1}(Q(X = \frac{1-d_0}{d_1}, G = 0)) \end{aligned}$$

where $Q(X,G)$ is the likelihood function of model (2.33) when outcome $Y = 1$ and random effect ν is marginalized:

$$Q(X,G) = \int R(d_0 + d_1 X + d_2 G + d_3 \nu)\varphi(\nu)d\nu$$

and

$$\begin{aligned} R(t) &= \frac{\exp(t)}{1+\exp(t)} \\ K(t) &= \int G(1+t\nu)\varphi(\nu)d\nu \end{aligned}$$

where $\varphi(\cdot)$ is the density function of standard normal distribution. The details of the proof is a slight variant of the proof in the appendix of Kuchenhoff (1995).

However, Kuchenhoff (1995) also pointed out that the model is practically non-identifiable without extra information. In order to show this, we generate a random sample of 20 observations by $X \sim N(0,1)$, $\nu \sim N(0,1)$, with $d_1 = 3$, $d_3 = 2$, and $d_0 = d_2 = 0$. To construct the likelihood, we assume d_0 and d_2 are

known and plot the likelihood versus d_1 and d_3. The likelihood function with observed outcome Y and covariate X is:

$$\mathcal{L} = \prod_{i=1}^{n} \int \left(\frac{e^{d_1 X_i + d_3 \nu}}{1 + e^{d_1 X_i + d_3 \nu}} \right)^{Y_i} \left(\frac{1}{1 + e^{d_1 X_i + d_3 \nu}} \right)^{1 - Y_i} \varphi(\nu) d\nu \qquad (2.34)$$

Integrations are calculated numerically and approximated by definite integrals from -6 to 6. This approximation is accurate, since the first two terms are bounded within $(0, 1)$ and $\varphi(\cdot)$ is close to 0 outside of $(-6, 6)$. We use *MATLAB* software (version 7.1) (MATLAB, 2005) to calculate the integration and generate the likelihood plots. The plot of likelihood versus d_1 and d_3 is shown in Figure 2.3. We can see a flat 'ridge' in the plot, spreading among different values of d_1 and d_3. The difference in largest values ('peaks') of the ridge is too small to detect, indicating that the MLE of the parameters is not numerically detectable even if it is unique. This is confirmed by the two plots of profile likelihoods (likelihood function of one parameter while maximizing with respect to the other one) in Figure 2.4. They implies that for different d_3's, there is always a value of d_1 to attain the 'ridge', showing that different combinations of (d_1, d_3) can achieve values that are very close to the maximum likelihood. Note that the spikes of the likelihood are due to calculation errors from the numerical integration and should be ignored. However, if we have additional information, that is, a fixed value or at least a strong prior, on σ^2 (or equivalently, α_2 or α_2^2/σ^2), the MLE will be practically identifiable. This might be available by using repeated measures, while it is reasonable to assume that the random errors are mainly caused by within-subject variation. Moreover, the same identifiability problem will arise if we model the binary outcome Y with a probit link. Therefore, we conclude that for the two-stage model (2.28) and (2.29), consistent estimate of β_1 can not be derived by likelihood-based method without additional information.

Figure 2.3: Likelihood of IV analysis with binary outcome

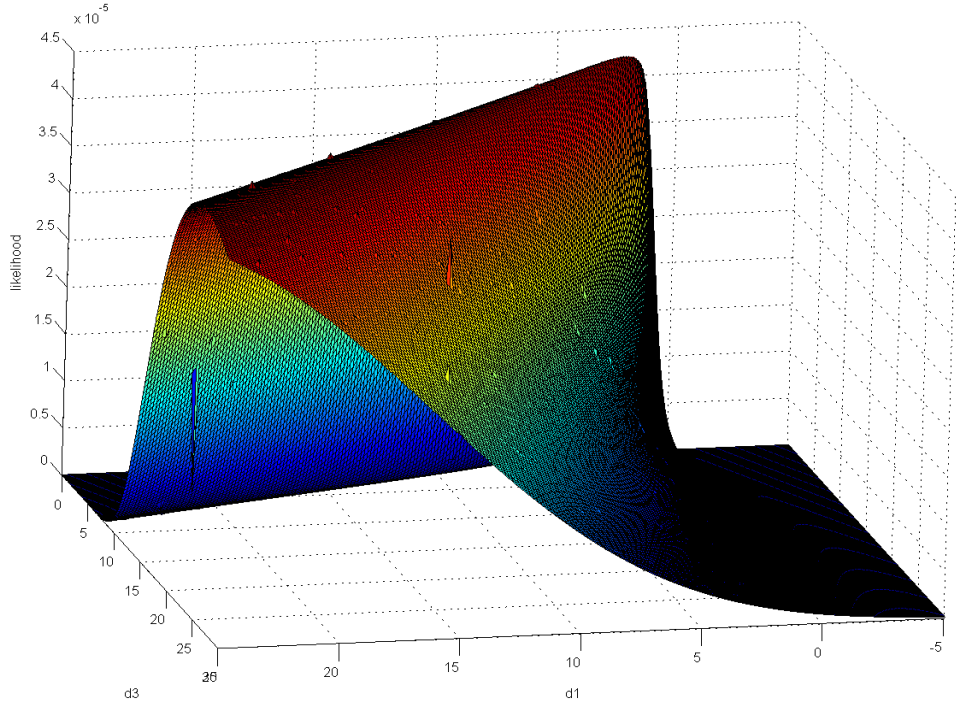
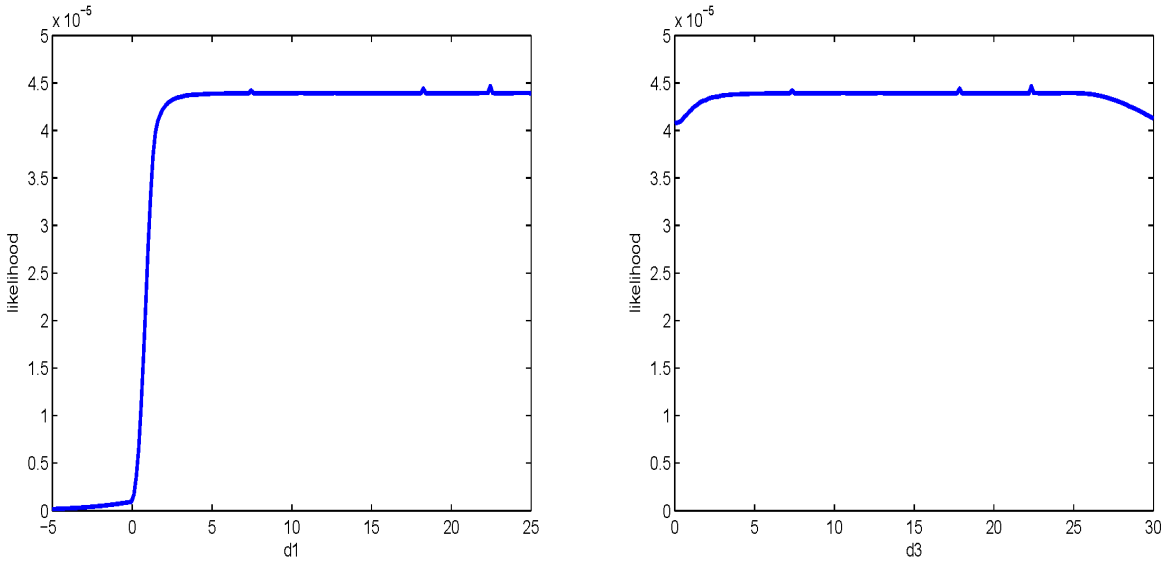


Figure 2.4: Profile Likelihood of IV analysis with binary outcome



CHAPTER 3

A Parametric Bayesian Approach for Instrumental Variable Analysis with Censored Time-to-Event Outcome

In this chapter, we first introduce some basic concepts of time-to-event data analysis, also known as survival analysis, and review the methods for time-to-event data based on linear models. We then introduce IV analysis with right-censored time-to-event outcome, and propose a parametric Bayesian model with a bivariate normal or non-normal elliptically contoured error distribution. We examine the performance of the model with normal error distribution through simulation studies. Two real data examples, the Women's Health Initiative Observational Study and the Atherosclerosis Risk in Communities Study, are used as illustration.

3.1 Preliminaries

3.1.1 Introduction to Time-to-Event Outcome

The identifiability problem of IV analysis with binary outcome in section 2.2.3 is fundamentally caused by the loss of information while the underlying continuous link function is dichotomized into binary outcome. Additional information can

sometimes be achieved in the form of time-to-event data, also known as survival data.

Survival analysis is a branch of statistics in which the time to and rate of occurrence of events (e.g. diseases and death) are of primary interest. *Censorship* is a defining feature of survival analysis. A subject is censored if his exact time to the event is not observed, but partial information on the event's occurrence is available. Right-censoring is the most common type of censoring, meaning that the subject's event time (also called a survival time, lifetime, or failure time) is longer than his observed on-study time. There are other types of censoring (e.g. left censoring and interval censoring), but we will focus on right-censored data in this section.

For each subject i, let Y_i be the time-to-event outcome variable, and C_i be a corresponding right-censoring time. Only the smaller of the two, $T_i = \min(Y_i, C_i)$, can be observed. If $Y_i < C_i$, the subject experiences the event and $Y_i = T_i$; otherwise the patient is right-censored and $Y_i > T_i$. Let $X_i = (X_{1i}, \ldots, X_{ki})'$ denote a vector of fixed-time explanatory covariates. Thus the observed data consist of (T_i, δ_i, X_i), where censoring indicator $\delta_i = \mathrm{I}[Y_i \leq C_i]$. In this dissertation, we assume the censoring is conditional random, meaning that Y_i is independent of C_i given X_i.

Important functions of the survival time Y include survival distribution function $S(y)$, hazard function $\lambda(y)$ and survival time density function $f(y)$. The survival distribution function represents the probability that time to event for a subject is beyond y:

$$S(y) = P(Y > y)$$

The hazard function describes failure rate at time y, which is the probability that

a subject fails at time y given that he has survived until y:

$$\lambda(y) = \lim_{\Delta y \to 0} \frac{P(y \leq Y \leq y + \Delta y \mid Y \geq y)}{\Delta y}$$

The survival time density function represents the instantaneous probability of failure at time y:

$$\begin{aligned} f(y) &= -\frac{d}{dy} S(y) \\ &= \lambda(y) S(y) \end{aligned}$$

In all the above functions, we assume y is continuous on $(0, \infty)$. Thus the likelihood function of observing (T_i, δ_i), $i = 1 \ldots n$ is:

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{n} f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda(T_i)^{\delta_i} S(T_i) \end{aligned} \tag{3.1}$$

3.1.2 Accelerated Failure Time Models

Accelerated Failure Time (AFT) models (Cox and Oakes, 1984) are a popular class of parametric models for survival data that are based on linear models. They provide alternatives to the commonly-used proportional hazards models (Cox, 1972). Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by a constant, an AFT model assumes that the effect of a covariate is to multiply the predicted event time by a constant. The AFT model has the following form:

$$\log(Y) = \beta_0 + X'\beta + \sigma W \tag{3.2}$$

where X is a vector of covariates, σ is a scale parameter for the random error W. Therefore the effect of covariate X on the survival function is:

$$S(y \mid X) = S_0(ye^{-X'\beta})$$

where $S_0(y)$ is the baseline survival function. The factor $e^{-X'\beta}$ is called an *acceleration factor* indicating how a change in covariate values changes the time scale from the baseline time scale. This means that the covariates have direct effects on the survival distribution without requiring any proportional hazards assumptions. AFT models can therefore be framed as linear models for the logarithm of the survival time.

3.1.2.1 Parametric AFT Models

For most AFT models, the distribution of the error term W is assumed to be of a known parametric form. This leads to a parametric distribution for survival time Y, since Y is a shifted and scaled transformation of W by model (3.2). Common parametric forms for W include: a standard extreme value distribution leading to a Weibull distribution for Y, a standard normal distribution leading to a log-normal distribution for Y, a standard logistic distribution leading to a log-logistic distribution for Y, etc. The fully parametric models have the advantages of increased power to detect covariate effects. The parameter estimates can be obtained through maximum likelihood estimation. The estimates have desirable properties such as asymptotically normal distributions with variances that can be estimated consistently from the data.

Parametric AFT models are comparatively easy to implement in standard software such as R and SAS. Diagnosis of parametric assumptions are mainly assessed by graphical methods such as hazard plot, Cox-Snell residual plot and quantile-quantile plot (Klein and Moeschberger, 2003).

3.1.2.2 Semiparametric AFT Model and Buckley-James Estimator

The parametric assumptions by fully parametric AFT models might not be valid in certain datasets. Instead, the semiparametric AFT model:

$$\log(Y) = X'\beta + \epsilon \tag{3.3}$$

allows the random error ϵ to have an unspecified distribution F_ϵ with mean μ and finite variance σ^2: $\epsilon \sim F_\epsilon$, $\mathrm{E}(\epsilon) = \mu$ and $\mathrm{Var}(\epsilon) = \sigma^2 < \infty$. This model only assumes homoscedasticity for the random errors.

One approach to estimate β in model (3.3) is the Buckley-James estimator (Buckley and James, 1979). Let $Z_i = \log(Y_i)$ and $Z_i^o = \log(T_i) = \log(Y_i) \wedge \log(C_i)$. The AFT model in (3.3) becomes:

$$Z_i = X_i'\beta + \epsilon_i \tag{3.4}$$

This is simply a standard linear regression model if there is no censoring. A new variable is defined as:

$$Z_i^* = \delta_i Z_i^o + (1 - \delta_i)\mathrm{E}[Z_i \mid Z_i \geq Z_i^o, X_i'] \tag{3.5}$$

with desirable property $\mathrm{E}(Z_i^* \mid X_i) = \mathrm{E}(\epsilon_i) + X_i'\beta = \mathrm{E}(Z_i \mid X_i)$. If Z_i^* were known, ordinary least squares could be used with the transformed responses Z_i^*. The Buckley-James estimating algorithm simultaneously updates the response estimate $\hat{Z}_i^*$ and parameter estimate $\hat{\beta}$ at each step and proceeds iteratively:

1. Select a reasonable initial estimator $\beta^{(0)}$, and let $\tilde{Z} = X\beta^{(0)}$.

2. Compute the residuals $\epsilon = Z^o - \tilde{Z}$ and the estimated transformed response

$$\begin{aligned} \hat{Z}_i^* &= \delta_i Z_i^o + (1 - \delta_i)\hat{E}(Z_i \mid Z_i \geq Z_i^o, X_i) \\ &= \delta_i Z_i^o + (1 - \delta_i)\left[\tilde{Z}_i - \{\hat{S}_\epsilon(\epsilon_i)\}^{-1}\int_{\epsilon_i}^{\infty} s d\hat{S}_\epsilon(s)\right], \; i = 1, \ldots, n, \end{aligned}$$

where $\hat{S}_\epsilon(\cdot)$ is the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function $1 - F_\epsilon$ using the censored residuals $\{\epsilon_i, \delta_i\}$.

3. Apply ordinary least squares to $\{\hat{Z}_i^*, X_i\}$. Update $\tilde{Z} = X\hat{\beta}$.

4. Stop if $\tilde{Z}$ converges. Otherwise, return to step 2.

The iterating procedure might eventually oscillate between two values of $\hat{\beta}$. In that case, the average of the two values will be used.

The Buckley-James estimator produces unbiased estimates of the covariate effects under relatively weak assumptions. Furthermore, it can be used to check the appropriateness of a parametric specification. The *bj* function in R package *Design* can be used to produce the Buckley-James estimates (Stare et al., 2001; Harrell, 2009; R Core Team, 2012). Ritov (1990) and Lai and Ying (1991) established the asymptotic properties of the estimator. However, the variance estimation of the Buckley-James estimator remains very difficult due to the presence of censored data in nonparametric density estimation.

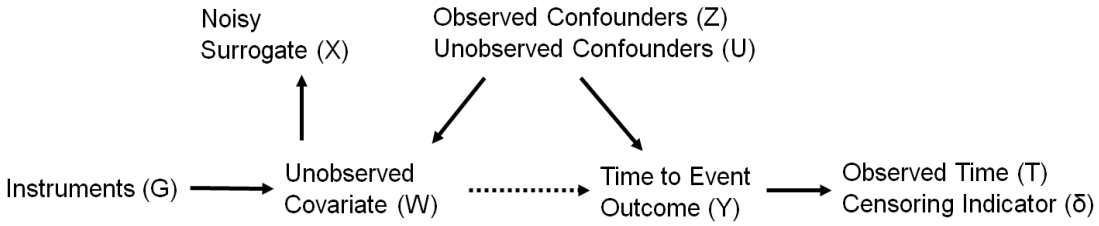
3.2 IV Analysis with Censored Time-to-Event Outcome

3.2.1 Introduction

In this section, we propose a parametric Bayesian model for IV analysis with censored time-to-event outcome. We examine its performance by simulation studies, and illustrate the method on two real data examples. Although our method can be easily extended to the case of arbitrary censoring (e.g. left censoring, interval censoring), we focus on right-censored time-to-event outcome for simplicity of illustration.

Similar to IV analysis with continuous outcome introduced in section 1.1, IV

Figure 3.1: Directed acyclic graph for instrumental variable model with right-censored time-to-event outcome in the presence of unobserved confounders and measurement errors in the intermediate covariate



analysis with censored time-to-event outcome follows a structure as indicated in Figure 3.1. For each subject i, let Y_i be the time-to-event outcome variable, C_i be the corresponding right-censoring time, $T_i = \min(Y_i, C_i)$ be the observed time, W_i be an unobserved continuous covariate of interest that is subject to measurement errors, X_i be the observed surrogate of W_i, Z_i be a vector of observed confounders, U_i be a vector of unobserved confounders, and G_i be a vector of instruments, $i = 1, \ldots, n$. The primary aim is to estimate the causal effect of W_i on Y_i based on the observed right-censored data consisting of n independent and identically distributed observations $(T_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots n$, where censoring indicator $\delta_i = I[Y_i \leq C_i]$.

With assumption of linear relationships among the variables Y_i, W_i, U_i, Z_i and G_i, the underlying structure in Figure 3.1 can be modeled by

$$W_i = \alpha_0 + \alpha_1{}'G_i + \alpha_2{}'Z_i + \alpha_3{}'U_i + \varepsilon_{1i} \tag{3.6}$$

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2{}'Z_i + \beta_3{}'U_i + \varepsilon_{2i} \tag{3.7}$$

$$X_i = W_i + \varepsilon_{3i} \tag{3.8}$$

$i = 1, \ldots n$, where ε_{1i}, ε_{2i}, and ε_{3i} are independent random errors with mean 0 and finite variances τ_1^2, τ_2^2, and τ_3^2 respectively, similar to the model in section 2.1.1. Note that ε_{3i} is the measurement error in the intermediate covariate. The unobserved confounder vector U_i is standardized to have mean 0 and covariance matrix Σ_U, where Σ_U is the correlation matrix of U_i. The variables ε_{1i}, ε_{2i}, ε_{3i}, U_i and G_i are assumed to be independent. Y_i is usually a monotone transformed survival time. For example, the second-stage equation (3.7) is an accelerated failure time model if Y_i is the log-transformed survival time. Again, the endogenous parameter β_1 is the parameter of primary interest.

3.2.2 A Parametric Bayesian Instrumental Variable Model

We consider the following two-stage linear model:

$$X_i = \alpha_0 + \alpha_1{}'G_i + \alpha_2{}'Z_i + \xi_{1i} \tag{3.9}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2{}'Z_i + \xi_{2i} \tag{3.10}$$

where the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution:

$$\begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \tag{3.11}$$

Our primary goal is to estimate and make inference on the endogenous parameter β_1.

This model can be used to adjust for unobserved confounders and measurement errors simultaneously. It is a reduced form of the model with equations (3.6),(3.7) and (3.8). Since W_i is not observed, we can replace W_i in equations (3.6) and (3.7) by $W_i = X_i - \varepsilon_{3i}$ from equation (3.8) and have the following

two-stage linear model:

$$X_i = \alpha_0 + \alpha_1{}'G_i + \alpha_2{}'Z_i + \alpha_3{}'U_i + \varepsilon_{1i} + \varepsilon_{3i} \quad (3.12)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2{}'Z_i + \beta_3{}'U_i + \varepsilon_{2i} - \beta_1\varepsilon_{3i} \quad (3.13)$$

Based on these two equations, parameters related to unobserved confounders U_i and random errors $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})$, which are $(\alpha_3, \beta_3, \tau_1^2, \tau_2^2, \tau_3^2)$, are not all identifiable. Since the parameter of primary interest is β_1, we can reduce the two equations to the two-stage model with equations (3.9) and (3.10), by combining the unobserved confounders and random errors:

$$\xi_{1i} = \alpha_3{}'U_i + \varepsilon_{1i} + \varepsilon_{3i} \text{ and } \xi_{2i} = \beta_3{}'U_i + \varepsilon_{2i} - \beta_1\varepsilon_{3i}$$

With normality assumptions on U_i, ε_{1i}, ε_{2i} and ε_{3i}, the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution as shown in equation (3.11), where $\sigma_1^2 = \alpha_3{}'\Sigma_U\alpha_3 + \tau_1^2 + \tau_3^2$, $\sigma_2^2 = \beta_3{}'\Sigma_U\beta_3 + \tau_2^2 + \beta_1^2\tau_3^2$, and $\rho\sigma_1\sigma_2 = \alpha_3{}'\Sigma_U\beta_3 - \beta_1\tau_3^2$.

Although we aim to estimate the endogenous parameter in a censored time-to-event context, the underlying model is the same as the two-stage linear model for continuous outcomes without censoring. Censoring only affects the estimation procedure, not the interpretation of the underlying causal model. In addition, the normality assumption in (3.11) is not essential for the development of our estimation and inferential method. In section 3.2.4, we note that our approach can be extended with minimal modifications to elliptically contoured models that include many useful non-normal models.

3.2.3 Estimation and Inference Procedure

In the absence of censoring, the classic IV methods for continuous outcome described in section 2.1 can be applied to derive consistent estimates of β_1. In the

presence of censoring, estimation and inference for β_1 can be drawn by using the maximum likelihood estimation theory and applying the delta method. However, the asymptotic approximations that the frequentist approaches rely on might not be valid when weak instruments are used (Lawlor et al., 2008). Moreover, there could be a high-dimensional optimization problem in maximizing likelihood when multiple observed confounders are incorporated into the model. To avoid these two problems, we develop a Bayesian approach with MCMC techniques to draw inferences on the endogenous parameter β_1. Furthermore, the Bayesian approach can take advantage of prior information by using informative priors for the parameters.

For the two-stage IV model (3.9)–(3.11), denote $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \rho)$, $\vec{T} = (T_1, ..., T_n)$, $\vec{\delta} = (\delta_1, ..., \delta_n)$, $\vec{X} = (X_1, ..., X_n)$, $\vec{Z} = (Z_1, ..., Z_n)$ and $\vec{G} = (G_1, ..., G_n)$. The likelihood function of observing $(\vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$ is:

$$\begin{aligned} \mathcal{L}(\theta \mid \vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) &= P(\vec{T}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \theta) \cdot P(\vec{X}, \vec{Z}, \vec{G} \mid \theta) \\ &= \prod_{i=1}^{n} f_1(T_i \mid X_i, Z_i, G_i)^{\delta_i} \, S(T_i \mid X_i, Z_i, G_i)^{1-\delta_i} \cdot f_2(X_i, Z_i, G_i) \end{aligned} \tag{3.14}$$

where

$$\begin{aligned} S(T \mid X, Z, G) &= 1 - \Phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2{}' Z - \frac{\sigma_2}{\sigma_1}\rho(X - \alpha_0 - \alpha_1{}' G - \alpha_2{}' Z)}{\sqrt{(1-\rho^2)\sigma_2^2}} \right) \\ f_1(T \mid X, Z, G) &= \phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2{}' Z - \frac{\sigma_2}{\sigma_1}\rho(X - \alpha_0 - \alpha_1{}' G - \alpha_2{}' Z)}{\sqrt{(1-\rho^2)\sigma_2^2}} \right) \\ f_2(X, Z, G) &= \phi\left(\frac{X - \alpha_0 - \alpha_1{}' G - \alpha_2{}' Z}{\sqrt{\sigma_1^2}} \right) \end{aligned}$$

$\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. The detailed derivation of the likelihood is given in the appendix.

We use independent vague priors for the parameters: a normal distribution $N(0, \varsigma^2)$ with large variance ς^2 for each element in α_0, α_1, α_2, β_0, β_1 and β_2, an inverse-gamma distribution Inv-Gamma(γ_1, γ_2) with small shape parameter γ_1 and small scale parameter γ_2 for σ_1^2 and σ_2^2, and a uniform distribution Unif$(-1, 1)$ for ρ. The MCMC method is used to generate samples from the posterior distributions of the parameters: In each iteration, a random walk Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970) is used to update the parameters one by one, while other parameters are fixed at their current states. Highly correlated parameters are updated simultaneously using a multiple-block Metropolis-Hasting algorithm (Chib and Greenberg, 1995) in order to have faster convergence for the Markov chains. Uniform proposal distributions are used for the random walk in our simulations and real data analysis, with widths chosen to obtain appropriate acceptance rates. The detailed MCMC algorithm is provided in the appendix. A sufficiently large amount of MCMC samples are generated from the posterior distribution. Sample mean of a parameter can be used to approximate the posterior mean and serve as an estimation of the parameter. Credible intervals of the parameters can be constructed by using the empirical quartiles of the simulated samples. We implemented the method in R (R Core Team, 2012). Our program is available online at http://www.biostat.ucla.edu/people/gangli/IV_MH.R , along with an example with simulated data at http://www.biostat.ucla.edu/people/gangli/IV_example.R .

Convergence of the MCMC algorithm can be examined visually by graphical methods including trace plots, histograms and autocorrelation plots, and quantitatively by using the Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998). Detailed convergence diagnostics are presented in section 3.4.3 for the real data examples.

3.2.4 Extension to Non-Normal Models

The method developed earlier for the normal IV model (3.9)–(3.11) can be easily extended to a more general IV model with elliptically contoured distributed errors. Specifically, the normality assumption (3.11) can be replaced by the following assumption

$$\begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} EC_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, g \right), \tag{3.15}$$

where a k-dimensional random vector ξ is said to have an elliptically contoured distribution (ECD), denoted by $\xi \sim EC_n(\mu, \Sigma, g)$, if it has the following density function

$$|\Sigma|^{-1/2} g[(x-\mu)'\Sigma^{-1}(x-\mu)]$$

for a given function g. Clearly the k-dimensional multivariate normal distribution is an ECD with $g(x) = (2\pi)^{-\frac{k}{2}} e^{-\frac{x}{2}}$. The ECD also includes many non-normal multivariate distributions such as the multivariate t, the multivariate Cauchy, the multivariate Laplace, the multivariate uniform, scale mixtures of normal distributions, and the multivariate stable distributions. See Fang and Anderson (1990) and Chmielewski (1981), and the references therein for some nice survey of ECD and its applications in various areas, including robust regression (Lange et al., 1989), risk measure (Landsman and Valdez, 2003), hyperspectral imaging data modeling (Marden and Manolakis, 2004), etc.

The family of ECDs share many nice properties of the multivariate normal, including that all marginal distributions and all conditional distributions of an ECD are ECD. With a bivariate ECD assumption on random errors $(\xi_{1i}, \xi_{2i})'$ other than the bivariate normal, we will simply modify the functions $S(T \mid X, Z, G)$, $f_1(T \mid X, Z, G)$ and $f_2(X, Z, G)$ in the likelihood function (3.14) accordingly: $S(T \mid X, Z, G)$ and $f_1(T \mid X, Z, G)$ are derived from the conditional distribution

of ξ_{2i} given ξ_{1i}; $f_2(X, Z, G)$ is derived from the marginal distribution of ξ_{1i}. Both are univariate ECDs in explicit form determined by the given function g.

3.3 Simulation Studies

A simulation study is conducted to assess the performance of our proposed parametric Bayesian IV model with normal error distribution, under frequentist criteria of bias, standard deviation (SD), and coverage probability (CP). Synthetic data is generated following the underlying model with unobserved confounders and measurement errors given by equations (3.12) and (3.13), with a variety of sample sizes, instrument strengths, censoring rates, and effects of intermediate covariate. Specifically, we use a simulation setting that is motivated by the WHI-OS real data example in section 3.4.1.

For the first-stage equation (3.12): A univariate instrument G_i and a univariate unobserved confounder U_i are used, and no observed confounder vector Z_i is considered, i.e. $\alpha_2 = \beta_2 = 0$. G_i and U_i both follow a standard normal distribution $N(0, 1)$; ε_{1i} and ε_{3i} follow normal distributions $N(0, 0.04)$ and $N(0, 0.015)$, respectively. The regression parameters are set as $\alpha_0 = 0.5$, $\alpha_1 = \sqrt{0.005}$, $\alpha_3 = \sqrt{0.04}$. This gives $E(X) = 0.5$ and $Var(X) = 0.1$, while $Cor^2(X, G) = 5\%$, similar to the mean and variance of the intermediate covariate and its correlation with the genetic instrument in the WHI-OS example. We also consider a stronger instrument with $Cor^2(X, G) = 10\%$ by setting $\alpha_1 = \sqrt{0.01}$ and $\varepsilon_{1i} \sim N(0, 0.035)$ while other parameters remain the same. For the second-stage equation (3.13): Different values are used for the endogenous parameter: $\beta_1 = (0, -0.5, -1)$, representing none, small, or moderate causal effects, respectively, of underlying true covariate W on outcome Y. These values of β_1 correspond to acceleration factors of 1, 1.6 and 2.7 when the outcome Y

is a log-transformed survival time and an accelerated failure time model is used. Other regression parameters are set as $\beta_0 = 5 - \alpha_0\beta_1$ and $\beta_3 = -\sqrt{0.05}$, and the random error ε_{2i} follows a normal distribution $N(0, 0.45)$. This gives $E(Y) = 5$ and $Var(Y) = 0.5$ when $\beta_1 = 0$, similar to the time-to-event outcome in the WHI-OS example. The underlying right-censoring time is set to be conditionally independent of Y_i: $C_i = \beta_0 + \beta_1 X_i + \beta_3 U_i + \varepsilon_{ci}$, where $\varepsilon_{ci} \sim N(\mu_c, 2)$ and μ_c is adjusted to give censoring rates $25\%, 50\%$ and 75%. The observed time T_i and censoring indicator δ_i are derived by $T_i = \min(Y_i, C_i)$ and $\delta_i = I[Y_i \leq C_i]$. We use sample size $n = (300, 500, 800)$ for each of the parameter settings.

We first use a regular linear regression survival model (without using IV) to evaluate the bias caused by the unobserved confounder U_i and the measurement error ε_{3i}. We generate 5000 synthetic data sets for each of the simulation settings. The following linear model is applied to each set of observed data (T_i, δ_i, X_i), $i = 1, \ldots, n$:

$$Y_i = \beta_1 X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(\mu, \sigma^2) \tag{3.16}$$

We denote this as the 'simple method'. Estimate and 95% confidence interval of the parameters (β_1, μ and σ^2) are derived based on maximum likelihood estimator and asymptotic normal approximation (Klein and Moeschberger, 2003). This can be easily implemented by using the *survreg* function in R package 'survival' (Therneau, 2013) or the LIFEREG procedure (SAS Institute Inc., 2008) in the SAS® software. Expectation and SD of the parameter estimates are approximated by the sample mean and sample SD of the 5000 estimates from independent simulated data sets. Coverage probability is approximated by the proportion of confidence intervals that cover the true value of the parameter.

We then apply our proposed IV model (3.9)–(3.11) to the simulated data. We generate 2000 synthetic data sets for each of the simulation settings described

earlier. For each data set, the likelihood is constructed using observed data $(T_i, \delta_i, X_i, G_i)$, $i = 1, \ldots, n$ and likelihood function (3.14). Independent vague priors are used: $N(0, 100^2)$ for each of α_0, α_1, β_0, and β_1; Inv-Gamma$(0.001, 0.001)$ for σ_1^2 and σ_2^2, and Unif$(-1, 1)$ for ρ. MCMC samples are generated from the posterior distribution, which is the product of the likelihood and the priors. The detailed MCMC algorithm is described in the Appendix. Specifically, a width of the proposal distribution that gives an acceptance rate between 0.3 and 0.4 is chosen for each parameter. We run $110,000$ iterations and use the first $10,000$ iterations as burn-in to achieve a state of convergence. We thin the resulting chain by taking every 5^{th} sample, in order to reduce autocorrelation. Posterior mean and credible interval are derived for each parameter based on the resulting $20,000$ posterior samples. Similar to simulations using the simple method described earlier, expectation and SD of the parameter estimate are approximated by the sample mean and sample SD of the 2000 posterior means from independent simulated data sets. Coverage probability is approximated by the proportion of credible intervals that cover the true value of the parameter.

Results of the parameter of primary interest, β_1, from the two analysis models are summarized in Table 3.1. The β_1 estimates from the simple method have substantial bias in all simulation settings: bias ≈ 0.45, 0.37 and 0.29 for $\beta_1 = 0$, -0.5 and -1, respectively, insensitive to different sample sizes and different censoring rates. The coverage probabilities are poor, especially when sample size is large or censoring rate is low (e.g. CP ≤ 0.36 when $n = 800$; CP ≤ 0.43 when censoring rate $= 25\%$). This is because the SD is smaller in these situations, while the bias remains large. On the other hand, the bias in β_1 estimation from our IV method is much smaller. For example, with instrument strength $R^2(X, G) = 0.05$ and $n = 800$, the bias of the IV estimate is <0.05. With a stronger instrument of $R^2(X, G) = 0.1$ and $n = 800$, the bias is further reduced to ≤ 0.021. Coverage

probabilities of the 95% credible intervals from the IV method are very close to the nominal level. Moreover, the bias from the IV method is always reduced as sample size increases, under different scenarios of censoring rates, instrument strengths and true values of β_1, suggesting consistency of the parametric Bayesian IV estimation. The bias, SD and coverage probability seem to be insensitive to different values of β_1. Unsurprisingly, the IV method tends to perform better in terms of bias and SD with stronger instrument and/or smaller censoring rate, and the SD decreases as sample size increases.

Using the IV method will result in larger variation in the estimates compared to the simple method, due to the uncertainty from the first-stage model (3.9). Our simulation results show that the β_1 estimates from our IV method have a four-to-five-fold increase in SD with instrument strength $R^2(X, G) = 0.05$ and a roughly threefold increase in SD with instrument strength $R^2(X, G) = 0.1$ compared to the simple method. This is an inevitable trade-off between accuracy and precision. The gains in bias reduction come at the cost of lower statistical power. It is worthwhile to apply the IV method when the major concern lies with bias in parameter estimation, and instruments with reasonable strength are available.

We further investigate the performance of our method by examining all the parameter estimates, with extensive simulations by varying the true values of all parameters. Our method performs well for estimation of all parameters in the IV analysis model: $\alpha_0, \alpha_1, \beta_0, \beta_1, \sigma_1^2, \sigma_2^2$, and ρ, in terms of bias and coverage probability (not reported here). In addition, we have conducted simulations with vague dependent priors. The results are very similar to the ones with vague independent priors (not reported here).

In order to investigate the robustness of our proposed model against deviation

of normality assumption, we conducted another simulation study with different true underlying distributions for the time-to-event outcome Y. The simulation setting is the same as the previous one with $R^2(X, G) = 0.05$, except for the random error terms ε_{2i} in equation (3.10) and ε_{ci} in censoring time C_i. ε_{2i} follows distributions of four different forms: normal, exponential, weibull, and a mixture of normal distributions, while fixing $E(\varepsilon_{2i}) = 0$ and $Var(\varepsilon_{2i}) = 0.45$. ε_{ci} follows the same distribution as ε_{2i}, fixing the censoring rate at 50%. We apply the proposed IV method to each of the data set, similar to described earlier. Results of β_1 estimation are summarized in Table 3.2. Again, when the distribution of ε_{2i} is normal, the IV estimation of β_1 has bias becoming close to 0 as sample size increases, and the coverage probability is close to the nominal level of 95%. These properties also hold for IV estimation of β_1 when ε_{2i} follows the three non-normal distributions. This suggests that the IV method based on normality assumption is quite robust against the deviation from normality.

Table 3.1: β_1 Estimation with and without Instrumental Variable Analysis on Simulated Data with Normal Random Errors

			Simple estimate without IV			IV estimate					
						$R^2(X,G)=0.05$			$R^2(X,G)=0.1$		
n	CR	β_1	Bias	SD	CP	Bias	SD	CP	Bias	SD	CP
300	25%	0	0.444	0.139	0.108	0.031	0.609	0.968	0.021	0.434	0.958
		-0.5	0.378	0.141	0.228	0.035	0.607	0.969	0.012	0.449	0.954
		-1	0.297	0.141	0.432	0.032	0.589	0.975	0.015	0.451	0.953
	50%	0	0.449	0.157	0.186	0.062	0.649	0.969	0.042	0.496	0.960
		-0.5	0.372	0.158	0.351	0.063	0.649	0.963	0.031	0.502	0.959
		-1	0.294	0.163	0.551	0.063	0.650	0.969	0.030	0.488	0.965
	75%	0	0.447	0.206	0.402	0.085	0.776	0.975	0.081	0.621	0.954
		-0.5	0.377	0.207	0.551	0.101	0.791	0.969	0.063	0.602	0.960
		-1	0.300	0.210	0.705	0.103	0.757	0.977	0.080	0.587	0.965
500	25%	0	0.447	0.107	0.013	0.015	0.484	0.955	0.010	0.347	0.950
		-0.5	0.372	0.107	0.067	0.020	0.470	0.963	0.010	0.342	0.955
		-1	0.297	0.110	0.229	0.006	0.485	0.958	0.004	0.339	0.961
	50%	0	0.449	0.122	0.044	0.022	0.519	0.970	0.018	0.385	0.955
		-0.5	0.371	0.125	0.153	0.008	0.526	0.958	0.010	0.370	0.965
		-1	0.296	0.127	0.348	0.014	0.538	0.958	0.019	0.378	0.965
	75%	0	0.438	0.156	0.210	0.064	0.626	0.973	0.055	0.485	0.953
		-0.5	0.371	0.164	0.363	0.066	0.654	0.966	0.029	0.487	0.958
		-1	0.296	0.166	0.552	0.067	0.620	0.970	0.045	0.480	0.959
800	25%	0	0.449	0.083	<0.001	0.004	0.393	0.957	0.005	0.278	0.949
		-0.5	0.374	0.086	0.008	0.013	0.386	0.953	0.004	0.272	0.952
		-1	0.297	0.087	0.071	0.007	0.378	0.958	0.006	0.272	0.956
	50%	0	0.447	0.096	0.003	0.015	0.434	0.955	0.007	0.301	0.956
		-0.5	0.372	0.097	0.032	0.034	0.422	0.964	0.016	0.312	0.952
		-1	0.299	0.099	0.145	0.040	0.412	0.960	0.007	0.302	0.954
	75%	0	0.449	0.125	0.055	0.035	0.507	0.964	0.017	0.383	0.956
		-0.5	0.372	0.126	0.163	0.034	0.513	0.958	0.011	0.373	0.966
		-1	0.296	0.128	0.358	0.049	0.523	0.963	0.021	0.374	0.967

Results are based on 2000 simulations using the IV method and 5000 simulations using the simple method. Bias is calculated as the absolute difference between the sample mean of the β_1 estimates and the true value of β_1. Standard deviation (SD) is calculated as the sample standard deviation of the β_1 estimates. Coverage probability (CP) is the proportion of 95% credible intervals (for IV estimates) or confidence intervals (for simple estimates) that cover β_1.

Table 3.2: β_1 Estimation with Instrumental Variable Analysis on Simulated Data with Normal and Non-Normal Random Errors

			IV estimate		
Error Distribution	n	β_1	Bias	SD	CP
Normal	300	0	0.063	0.631	0.970
		-0.5	0.055	0.655	0.962
		-1	0.071	0.653	0.970
	500	0	0.019	0.514	0.961
		-0.5	0.009	0.506	0.962
		-1	0.015	0.509	0.970
	800	0	0.009	0.414	0.955
		-0.5	0.022	0.414	0.962
		-1	0.001	0.403	0.962
Exponential	300	0	0.031	0.431	0.974
		-0.5	0.039	0.436	0.965
		-1	0.052	0.429	0.971
	500	0	0.018	0.367	0.958
		-0.5	0.026	0.356	0.957
		-1	0.001	0.359	0.964
	800	0	0.001	0.279	0.960
		-0.5	0.011	0.281	0.963
		-1	0.011	0.286	0.959
Weibull	300	0	0.080	0.714	0.969
		-0.5	0.061	0.744	0.974
		-1	0.055	0.744	0.966
	500	0	0.012	0.583	0.967
		-0.5	0.022	0.591	0.962
		-1	0.025	0.595	0.957
	800	0	0.004	0.486	0.955
		-0.5	0.016	0.476	0.956
		-1	0.017	0.470	0.959
Normal Mixture	300	0	0.061	0.649	0.979
		-0.5	0.045	0.662	0.958
		-1	0.040	0.629	0.976
	500	0	0.009	0.539	0.960
		-0.5	0.031	0.538	0.959
		-1	0.009	0.533	0.958
	800	0	0.009	0.433	0.951
		-0.5	0.001	0.431	0.955
		-1	0.020	0.421	0.952

'Error Distribution' refers to distribution of ε_{2i} in equation (3.10) and ε_{ci} in censoring time C_i. Distribution 'Normal Mixture' is a mixture of two normal distributions $N(-.63, .05) \cdot 0.5 + N(.63, .05) \cdot 0.5$. Results are based on 2000 simulations. Censoring rate is 50% and instrument strength is $R^2(X, G) = 0.05$. Bias is calculated as the absolute difference between the sample mean of the β_1 estimates and the true value of β_1. Standard deviation (SD) is calculated as the sample standard deviation of the β_1 estimates. Coverage probability (CP) is the proportion of 95% credible intervals that cover β_1.

3.4 Real Data Examples

3.4.1 Women's Health Initiative Observational Study

We illustrate the proposed IV method using two real data examples. The first one is a prospective case-control study nested within the Women's Health Initiative Observational Study (WHI-OS). In this study, we want to investigate the effect of high-sensitivity C-reactive protein (hsCRP) on development of diabetes. hsCRP is an inflammatory marker that has been positively associated with diabetes (Han et al., 2002; Freeman et al., 2002; Liu et al., 2007). However, whether lowering hsCRP level will result in diabetes prevention is uncertain. Therefore, we apply the Mendelian Randomization (MR) method (i.e. IV analysis with genetic instruments) with time-to-event outcome to make inference about the causal effect of hsCRP on diabetes and account for potential impact of unobserved confounders and measurement errors.

In the WHI-OS, 82069 postmenopausal women (50-59 years of age) with no history of diabetes were followed-up for a mean of 5.5 years. 1584 cases of diabetes were identified and matched with 2198 controls (by age, ethnicity, clinical center, time of blood draw, and length of follow-up). We focus on the subgroup of whites (954 cases and 968 controls) to avoid the potential problem of population stratification. Time to diabetes diagnosis from baseline was recorded for each case, and time to last visit from baseline for each control. Plasma concentration of hsCRP is measured for each subject. Descriptive statistics of baseline characteristics are summarized by case-control status in Table 3.3. More detailed descriptions of the study are given in Liu et al. (2007) and Chan et al. (2011). 13 haplotype-tagging single-nucleotide polymorphisms (tSNPs) across 2.3 kb of the *CRP* (C-reactive protein, pentraxin-related) genes that had been shown to

account for most of the genetic variation within the CRP locus are used as instruments. Details of selection of the tSNPs are given in Lee et al. (2009). None of the 13 tSNPs shows indication of Hardy-Weinberg disequilibrium (all p-values from a Hardy-Weinberg equilibrium test > 0.05 after Bonferroni correction).

We first apply a simple method similar to what has been described in section 3.3 to estimate the association between hsCRP and time to diabetes diagnosis:

$$Y_i = \beta_1 X_i + \beta_2{}' Z_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(\mu, \sigma^2) \tag{3.17}$$

$i = 1, \ldots, n$. For each subject i, Y_i is the log-transformed time to diabetes diagnosis (in days); X_i is the log-transformed hsCRP level; Z_i is a vector of observed potential confounders including age, body mass index, cigarette smoking, alcohol intake, hormone-replacement therapy, family history of diabetes and physical activity. The censoring indicator δ_i is 1 for cases and 0 for controls. The observed time variable T_i is log-transformed time to diabetes diagnosis for cases and log-transformed time to last visit for controls. Since the outcome Y_i is a log-transformed survival time, this model is a log-normal accelerated failure time model. By using the SAS® procedure LIFEREG (SAS Institute Inc., 2008), β_1 has an estimate (SE) of -0.446 (0.089) with a p-value $<.001$. This correspond to a 95% confidence interval of $(-0.621, -0.272)$ as summarized in Table 3.4. This significant negative association is consistent with the previous finding by Liu et al. (2007), who analyzed the same data set and found that hsCRP was significantly associated with increased diabetes risk (odds ratio = .16 with 95% confidence interval $(1.03, 1.30)$ and a p-value $<.001$).

We then apply the proposed parametric IV method with two-stage model (3.9)–(3.11) to estimate the causal effect of hsCRP on time to diabetes diagnosis. For each subject i, instrument G_i is a vector of the 13 tSNPs, where each tSNP is coded as an additive effect model, i.e. coded as either 0, 1, or 2 depending on the

number of minor alleles. Y_i, X_i, Z_i, T_i and δ_i are defined as earlier. The likelihood is constructed using observed data $(T_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots, n$ and likelihood function given by equation (3.14). Independent vague priors are used: $N(0, 100^2)$ for each element of α_0, α_1, α_2, β_0, β_1 and β_2; Inv-Gamma$(0.001, 0.001)$ for σ_1^2 and σ_2^2, and Unif$(-1, 1)$ for ρ. In the MCMC sampling, a width of the proposal distribution that gives an acceptance rate between 0.35 and 0.4 is chosen for each parameter. We generate 30 chains from different initial values, with $1,100,000$ iterations $(100,000$ burn-ins) in each chain. We thin the chains by taking every 10^{th} sample to reduce autocorrelation. A detailed discussion of convergence is given in section 3.4.3. Posterior mean, posterior standard deviation and credible interval are derived for each parameter based on the resulting $3,000,000$ combined samples. Figure 3.2(a) is a histogram of the resulting MCMC samples of β_1. The brackets on the horizontal axis denote the 95% credible interval. The posterior distribution of β_1 is fairly normal, with mean -0.162, standard deviation 0.426 and 95% credible interval $(-0.987, 0.685)$, as summarized in Table 3.4.

Based on these results, we see that although hsCRP is significantly associated with development of diabetes, there is not sufficient evidence of causal effect of hsCRP on time to diabetes diagnosis among white postmenopausal women. This is consistent with the previous finding by Brunner et al. (2008), who applied the MR approach in a case-control study and found that the associations between C-reactive protein (CRP) and diabetes incidence are likely to be noncausal. One possible explanation for the association is that hsCRP level is affected by causal factors of diabetes, such as obesity (Keavney, 2008). On the other hand, the instruments in this IV analysis may not be strong enough to provide sufficient statistical power to detect small effect sizes (partial R-square = 0.028), even though the sample size is reasonably large.

3.4.2 Atherosclerosis Risk in Communities Study

The second data example is a subset in the Atherosclerosis Risk in Communities (ARIC) Study. In this example, we focus on the aspect of measurement error correction of our proposed IV method, and we assume that there is no unobserved confounder (i.e. all confounders are adjusted). Therefore, IV assumption (1) described in section 1.1 is reduced to: Instrument G is independent of measurement errors in intermediate covariate W. The ARIC study is a multi-center prospective cohort study of cardiovascular disease and its risk factors. A total of 15,792 subjects aged 45-64 years were recruited from four US communities in 1987-89. They received 4 clinical examinations at 3-year intervals (Visits 1-4). Medical, social and demographic data were collected at each visit. Hospitalization information was obtained by annual telephone follow-up and active surveillance in the communities. A more detailed description of the study is reported elsewhere (The ARIC Investigators, 1989). In this study, we are interested in estimating the association between systolic blood pressure (SBP) and development of coronary heart disease (CHD), after correcting for potential bias due to measurement error through IV analysis. For each of Visits 1-4, a subject's SBP level is an average of three measurements. We use Visit 2 (1990-92) as baseline, and use the SBP level at Visit 1 (1987-89) as an instrument of the baseline SBP level. We exclude subjects that (1) have missing baseline information, (2) do not have information after baseline, and/or (3) have developed their first CHD event prior to baseline. After the exclusion, our data consists of 12,782 subjects, 768 of which have CHD events during the follow-up. Descriptive statistics of baseline characteristics are summarized in Table 3.5.

For each subject i, outcome Y_i = time to the first CHD event from baseline (Visit 2) in years; censoring indicator $\delta_i = 1$ if the subject has at least one CHD

event during the follow-up, and 0 otherwise; observed time $T_i = Y_i$ if $\delta_i = 1$, and time to the last visit from baseline in years otherwise; covariate of interest X_i = standardized log-transformed SBP level at baseline; instrument G_i = standardized log-transformed SBP level at Visit 1; both X_i and G_i are standardized to have standard deviation 1; Z_i is a vector of observed potential confounders at baseline, including ethnicity (black vs. non-black) and other potential risk factors of CHD developed by the Framingham Heart Study: gender, age, total cholesterol level, high-density lipoprotein cholesterol level, smoking behavior, and diabetes status (Wilson et al., 1998).

Similar to the previous example in Section 3.4.1, we first apply the simple method with equation (4.20) to estimate the association between SBP and time to CHD. β_1 has an estimate (SE) of -0.779 (0.087) with a p-value $<.001$, corresponding to a 95% confidence interval of $(-0.950, -0.608)$ as summarized in Table 3.6. We then apply the proposed parametric IV method with two-stage model (3.9)–(3.11), primarily aiming to correct for potential measurement error bias. Similarly, the likelihood is constructed using observed data $(T_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots, n$ and likelihood function (3.14). Similar independent vague priors are used as described in Section 3.4.1. In the MCMC sampling, a width of the proposal distribution that gives an acceptance rate between 0.2 and 0.25 is chosen for each parameter. Due to the high correlations between some of the parameters ($\text{cor}(\alpha_0, \alpha_1) \simeq -0.95$, $\text{cor}(\beta_0, \beta_1) \simeq -0.97$ in the posterior samples), we used a multiple-block Metropolis-Hasting algorithm (Chib and Greenberg, 1995) to update the parameters: α_0 and α_1 are updated simultaneously; β_0 and β_1 are updated simultaneously. This dramatically improves convergence and results in greatly reduced autocorrelations. We generate 40 chains from different initial values, with $2,100,000$ iterations ($100,000$ burn-ins) in each chain. The chains are thinned to reduce autocorrelation by taking every 20^{th} sample. A detailed discus-

sion of convergence is given in section 3.4.3. Posterior mean, posterior standard deviation and credible interval are derived for each parameter based on the resulting $4,000,000$ combined samples. Figure 3.2(b) is a histogram of the resulting MCMC samples of β_1. The brackets on the horizontal axis denote the 95% credible interval. The posterior distribution of β_1 appears to be normal. It has mean -1.180, standard deviation 0.141 and 95% credible interval $(-1.460, -0.907)$, as summarized in Table 3.6. A standard deviation increase in log-transformed SBP level is associated with an acceleration of 1.18 years in time to the first CHD event. We observe a larger effect size of SBP on CHD development compared to the simple analysis. This result suggests that the effect size of β_1 calculated by the simple method is possibly attenuated by measurement errors in X_i.

In the IV analysis, we assume that the SBP measurement at an earlier visit is an instrument of the SBP measurement at a later visit. This assumption is weaker than the assumption that both the earlier and later measurements are replicates of noisy surrogate (Carroll et al., 2006; Gustafson, 2007). This is because the latter assumption fixes $\alpha_1 = 1$ while the former assumption does not. Note that the instrument G is not required to be independent of the observed confounders Z, since the confounding effects of Z are adjusted when Z is included in both stages of the model (equations (3.9) and (3.10)). Since a subject's SBP level at certain time point is naturally predictive of his/her SBP level three years later, G_i is a strong instrument of X_i (partial R-square = 0.35). Furthermore, measurement of the instrument does not need to be accurate: Measurement errors in instrument G will not violate the IV assumptions. Therefore, the SBP at Visit 1 can still serve as an instrument if it is also subject to measurement errors.

Table 3.3: Baseline characteristics of a white subgroup within the Women's Health Initiative Observational Study (WHI-OS)

Characteristic	Controls (n=968)	Case (n=954)
Age, mean ± SD, year	63.9 ± 6.9	63.9 ± 6.9
BMI, mean ± SD, kg/m^2	26.5 ± 5.1	32.5 ± 6.8
Physical activity, MET-h/wk		
Median	10.5	5.8
Interquartile range	3.8 – 20.6	0.9 – 14.0
Smoking status, %		
Nonsmoker	52.3	49.7
Past smoker	42.6	43.7
Current smoker	5.1	6.6
Alcohol intake, %		
Nondrinker	9.8	13.4
Past drinker	16.5	22.3
Current drinker, <1 drink/week	31.9	40.1
Current drinker, ≥ 1 drink/week	41.8	24.2
Hormone-replacement therapy, %		
Never	37.1	48.5
Past	14.3	15.3
Current	48.6	36.2
Family history of diabetes, %		
Yes	30.2	51.5
No	69.8	48.5
HsCRP, mg/L		
Median	2.05	4.06
Interquartile range	0.91 – 4.24	2.16 – 7.55

BMI = body mass index; hsCRP = high-sensitivity C-reactive protein. Family history of diabetes is defined as self-reported diabetes in a first-degree relative. Medians and interquartile ranges are provided for continuous variables with skewed distributions.

Table 3.4: Instrumental Variable (IV) analysis versus simple method in a subgroup analysis of whites within the Women's Health Initiative Observational Study (WHI-OS)

The simple method uses a log-normal accelerated failure time model to estimate the association between high-sensitivity C-reactive protein (hsCRP) and time to diabetes diagnosis. The IV analysis uses our proposed Bayesian IV method to estimate the causal effect of hsCRP on time to diabetes diagnosis, using 13 selected tSNPs as genetic instruments. Both models adjust for observed potential confounders including age, body mass index, cigarette smoking, alcohol intake, hormone-replacement therapy, family history of diabetes and physical activity.

	estimate of β_1	SE	95% CI
Simple method	-0.446	0.089	(-0.621, -0.272)
IV analysis	-0.162	0.426	(-0.987, 0.685)

CI stands for confidence interval for the simple method and credible interval for the IV analysis.

SE in the IV analysis is estimated by the posterior standard deviation of β_1.

The 13 selected tSNPs are: rs4275453, rs2808634, rs3093059, rs2794521, rs1417938, rs1800947, rs1130864, rs1205, rs3093075, rs3093068, rs2808629, rs2369146, and rs1470515.

Table 3.5: Baseline characteristics of a subset of the Atherosclerosis Risk in Communities (ARIC) Study

Characteristic	Baseline (Visit 2) ($n = 12,782$)
Age, mean ± SD, year	56.9 ± 5.7
Gender, % of Female	57
Ethnicity, % of Black	24.6
Current smoker, %	21.6
Diabetes, %	11
Total cholesterol level, mean ± SD, mmol/L	5.42 ± 1.01
HDL cholesterol level, mean ± SD, mmol/L	1.29 ± 0.43
Systolic blood pressure, mean ± SD, mmHg	145 ± 17.7

HDL = high-density lipoprotein.

Table 3.6: Instrumental Variable (IV) analysis versus simple method in a subset of the Atherosclerosis Risk in Communities (ARIC) Study

The simple method uses a linear regression survival model with normally distributed residuals to estimate the association between standardized log-transformed systolic blood pressure (SBP) level at baseline and time to the first CHD event. The IV analysis uses our proposed Bayesian IV method to estimate this association, by using standardized log-transformed SBP level at Visit 1 as an instrument to correct for potential bias due to measurement errors in baseline SBP. Both models adjust for observed potential confounders including gender, age, total cholesterol level, high-density lipoprotein cholesterol level, smoking behavior, and diabetes status.

	estimate of β_1	SE	95% CI
Simple method	-0.779	0.087	(-0.950, -0.608)
IV analysis	-1.180	0.141	(-1.460, -0.907)

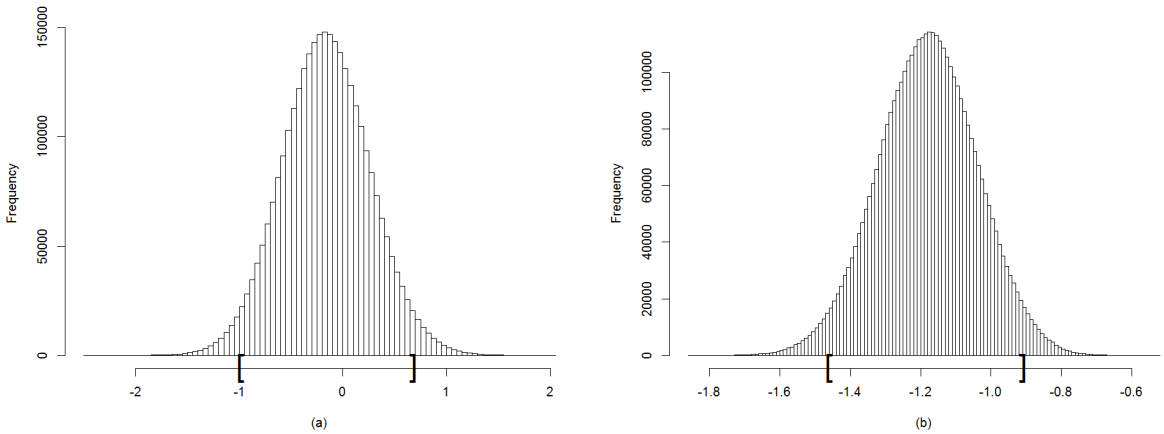
CI stands for confidence interval for the simple method and credible interval for the IV analysis.

SE in the IV analysis is estimated by the posterior standard deviation of β_1.

3.4.3 MCMC Convergence Diagnostics

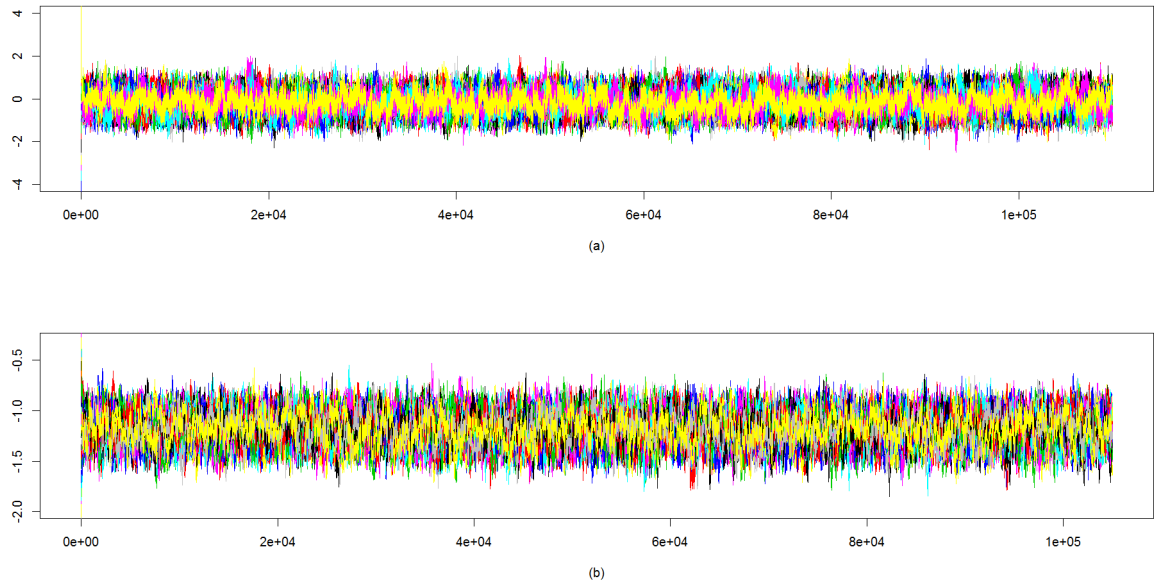
We assess the convergence of MCMC sampling in the two real data examples. Trace plots of parallel chains with diverse initial values are monitored for all the parameters. Figure 3.3 shows the trace plots of β_1: (a) for the WHI-OS data and (b) for the ARIC data. Different chains are marked with different colors. The chains seem to be mixing well and stable over the whole period. We further visually compare the histograms of individual chains to the histogram of combined samples. No obvious difference is observed. Autocorrelation plots of individual chains are monitored for each parameter. Figure 3.4 shows the autocorrelation plots of two representative individual chains for β_1: (a) for the WHI-OS data and (b) for the ARIC data. The autocorrelation is relatively low in the individual chains (generally less than 0.1 after a lag of 1000 for the WHI-OS data and less than 0.1 after a lag of 500 for the ARIC data). Finally, we use the Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998) to quantitatively measure convergence. The 'potential scale reduction factor' (PSRF) is calculated for each parameter, together with its 95% confidence interval. Approximate convergence is diagnosed when the upper limit of PSRF is close to 1. The 95% upper confidence limit of PSRP for β_1 is <1.005 for both real data examples, indicating good convergence properties of the method.

Figure 3.2: Histograms of the posterior samples of β_1 from normal IV model



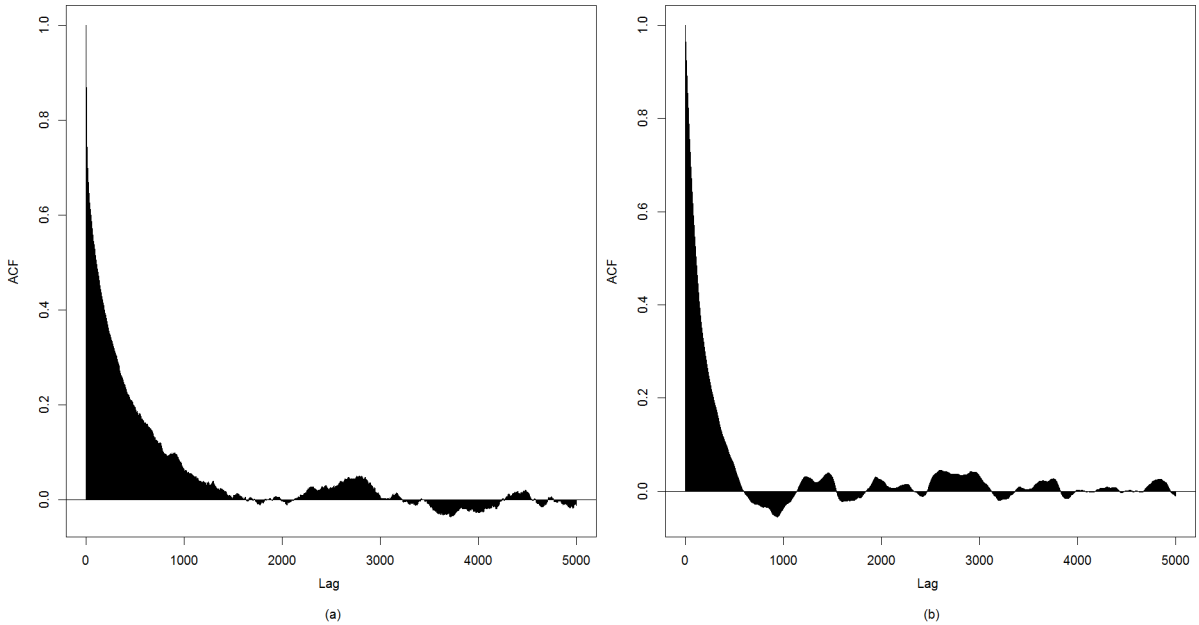
(a) WHI-OS example; (b) ARIC example. Posterior samples after discarding burn-in and thinning. The brackets denote the limits of the 95% credible interval.

Figure 3.3: Trace plots of the posterior samples of β_1 from normal IV model



(a) WHI-OS example; (b) ARIC example. Posterior samples after thinning.

Figure 3.4: Autocorrelation plots of individual chains of β_1 from normal IV model



(a) WHI-OS example; (b) ARIC example. Posterior samples after discarding burn-in and thinning.

CHAPTER 4

A Semiparametric Bayesian Approach for Instrumental Variable Analysis with Censored Time-to-Event Outcome

In Chapter 3, we assume the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution or other parametric distribution such as a bivariate elliptically contoured distribution. This allows a natural extension of the linear two-stage model for uncensored continuous outcomes into an instrumental variable (IV) model for censored time-to-event outcomes with MCMC procedures. This parametric Bayesian IV method requires full specification of a parametric model for the error terms. It also assumes homogeneous error distribution, which may not reflect the real underlying distribution. Thus, we consider to develop a more flexible IV method with less stringent model assumptions. However, it is difficult to handle arbitrary censoring (including left-censoring, interval-censoring and right-censoring) using the classic semiparametric frequentist approaches such as the Buckley-James estimator (Buckley and James, 1979). Therefore, we extend the parametric Bayesian IV approach to a semiparametric Bayesian IV approach by using a Dirichlet process mixture (DPM) model, to allow for heterogeneity in the random error distribution in the presence of arbitrarily censored data.

In this chapter, we first review the Dirichlet process, its application of DPM models, and the existing MCMC sampling algorithms in section 4.1. We then

introduce our semipametric Bayesian IV method in section 4.2.1, with estimation and inference procedure described in section 4.2.2. In section 4.3, we examine the performance of the semipametric IV model through simulation studies, compared with the parametric IV model proposed earlier. In section 4.4, the DPM model is applied to the Women's Health Initiative Observational Study and the Atherosclerosis Risk in Communities Study as illustration .

4.1 Preliminaries

4.1.1 Introduction to Dirichlet Process

First introduced by (Ferguson, 1973), the Dirichlet Process (DP) is a distribution over distributions, i.e. each draw from a Dirichlet process is itself a distribution. A random distribution H is distributed according to a DP with base distribution H_0 and strength parameter ν, if for any finite measurable partition $A_1, \ldots, A_r$ of the parameter space Θ, the measures

$$(H(A_1), \ldots, H(A_r)) \sim \text{Dirichlet}(\nu H_0(A_1), \ldots, \nu H_0(A_r))$$

where the Dirichlet distribution is a distribution over the K-dimensional probability simplex: $\{(\pi_1, \ldots, \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\}$. $(\pi_1, \ldots, \pi_K)$ follows Dirichlet$(\nu_1, \ldots, \nu_K)$ if:

$$P(\pi_1, \ldots, \pi_K) \quad = \quad \frac{1}{B} \prod_{k=1}^{K} \pi_k^{\nu_k - 1} \text{ where } B = \frac{\prod_k \Gamma(\nu_k)}{\Gamma(\sum_k \nu_k)}$$

The Dirichlet distribution is a multidimensional extension of the beta distribution. The base distribution H_0 and strength parameter ν represent mean and inverse-variance of the DP: for any measurable set $A \subset \Theta$,

$$\text{E}[H(A)] = H_0(A), \quad \text{Var}[H(A)] = H_0(A)(1 - H_0(A))/(\nu + 1)$$

For $H \sim \mathrm{DP}(\nu, H_0)$ and $\theta_1, \ldots, \theta_n \overset{\text{i.i.d.}}{\sim} H$, the posterior distribution:

$$H \mid \theta_1, \ldots, \theta_n \sim \mathrm{DP}(\nu + n, \frac{\nu H_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\nu + n}) \tag{4.1}$$

where δ_{θ_i} is a point mass located at θ_i. Thus the predictive distribution for $\theta_{n+1} \mid \theta_1, \ldots, \theta_n$ with H marginalized out is:

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{1}{\nu + n} \left(\nu H_0 + \sum_{i=1}^{n} \delta_{\theta_i} \right) \tag{4.2}$$

the posterior base distribution of H.

One way to visualize the DP is the Pólya urn scheme (Blackwell and Macqueen, 1973). Starting from (4.2), we can construct a distribution over sequences $\theta_1, \theta_2, \ldots$ by iteratively drawing each θ_i given $\theta_1, \ldots, \theta_{i-1}$, as for $n > 1$, the joint distribution

$$P(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} P(\theta_i \mid \theta_1, \ldots, \theta_{i-1})$$

Specifically, each value in Θ is a unique color, and draws $\theta \sim H$ are balls with the drawn value being the color of the ball. In addition there is an urn containing previously seen balls. In the beginning there are no balls in the urn, and we pick a color drawn from H_0, i.e. draw $\theta_1 \sim H_0$, paint a ball with that color, and drop it into the urn. In subsequent steps, say the $n+1$st, we will either, with probability $\frac{\nu}{\nu+n}$, pick a new color (draw $\theta_{n+1} \sim H_0$), paint a ball with that color and drop the ball into the urn, or, with probability $\frac{n}{\nu+n}$, reach into the urn to pick a random ball out (draw θ_{n+1} from the empirical distribution), paint a new ball with the same color and drop both balls back into the urn.

From the Pólya urn scheme, we see that for a long enough sequence of draws from H, the value of any draw will be repeated by another draw, regardless the smoothness of H_0. This implies that H is composed only of a weighted sum of point masses, i.e. it is a discrete distribution.

Besides the Pólya urn scheme, the DP can also be expressed in terms of the "stick-breaking" construction as given in Sethuraman (1994):

$$\begin{aligned} H &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \\ \text{where } \theta_k^* &\overset{\text{i.i.d.}}{\sim} H_0 \\ \pi_k &= \beta_k \prod_{i=1}^{k-1} (1-\beta_i) \\ \beta_k &\overset{\text{i.i.d.}}{\sim} \text{Beta}(1,\nu) \end{aligned} \tag{4.3}$$

Truncating the summation of H in (4.3) at a large integer K results in a model considered in Ishwaran and Zarepour (2002). This reduces H into finite dimensional form as $H = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}$.

4.1.2 Dirichlet Process Mixture Models and MCMC algorithms

The most common application of the DP is Dirichlet Process Mixture (DPM) models, which is becoming increasingly popular in Bayesian applications, due to the growth of MCMC simulation methods. The DPM models use Dirichlet process priors to avoid critical dependence on parametric models and robustify parametric assumptions. Instead of using a pre-specified number of mixture components, a DPM model allows the number of mixture components to be determined by both the prior and the data. It is a general approach which allows model parameters to vary from observation to observation.

We model a set of observations $(Z_1, \ldots, Z_n)$ using a set of latent parameters $(\theta_1, \ldots, \theta_n)$. Each θ_i is drawn independently and identically from H, while each

Z_i has distribution $F(\theta_i)$ parameterized by θ_i:

$$\begin{aligned} Z_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid H &\sim H \\ H &\sim \mathrm{DP}(\nu, H_0) \end{aligned}$$

$i = 1, \ldots, n$.

A variety of MCMC methods have been developed to sample from the posterior distribution of parameters of a DPM model. When H_0 is a conjugate prior for the likelihood given by F, the Gibbs sampling method considered by Escobar (1994) and Escobar and West (1995), the Gibbs sampling method by West et al. (1994) and Bush and MacEachern (1996), the collapsed cluster sampling method by Maceachern (1994) and the blocked Gibbs sampler by Ishwaran and James (2001) can be used. The first three methods exploit the Pólya urn scheme, whilst the last method considers the truncated stick-breaking process. When H_0 is a non-conjugate prior, the "no-gaps" algorithm of Maceachern and Müller (1998) based on the Pólya urn scheme, and Metropolis-Hasting algorithms with different modifications (Neal, 2000; Jain and Neal, 2007) can be applied. Here we illustrate one algorithm with conjugate priors (Escobar, 1994; Escobar and West, 1995) and two algorithms for non-conjugate priors (algorithms 5 and 8 in Neal, 2000).

Let the state of the Markov Chain consist of $\theta = (\theta_1, \ldots, \theta_n)$. For conjugate prior H_0, the Gibbs sampling method can be used to sample from the posterior distribution of θ. The Escobar (1994) approach is to repeatedly draw values for each θ_i from its conditional distribution given both data Z and θ_j for $j \neq i$ (denoted as θ_{-i}). Since θ_i's are exchangeable, from (4.2) we have the conditional

distribution:

$$\theta_i \mid \theta_{-i}, Z_i \sim \sum_{j \neq i} q_{ij} \delta(\theta_j) + r_i S_i \tag{4.4}$$

$i = 1, \ldots, n$. Here, S_i is the posterior distribution for θ based on the prior H_0 and the single observation Z_i with likelihood $F(Z_i, \theta)$. The values of q_{ij} and r_i are defined by:

$$q_{ij} = b\, F(Z_i, \theta_j) \tag{4.5}$$

$$r_i = b\, \nu \int F(Z_i, \theta)\, dH_0(\theta) \tag{4.6}$$

where b is such that $\sum_{j \neq i} q_{ij} + r_i = 1$. Computing the integral r_i and sampling from R_i must be feasible operations for this algorithm. This will generally so when H_0 is the conjugate prior for F.

For non-conjugate prior H_0, we first introduce an equivalent form of the parameters θ. Let c_i indicate which "latent class" is associated with observation Z_i, with the numbering of the c_i being no significance. For each class, c, the parameter θ_c determines the distribution of observations from that class: $Z_i \mid c_i, \theta_C \sim F(\theta_{c_i})$, where $\theta_C = (\theta_c : c \in \{c_1, \ldots, c_n\})$, i.e. all distinct values of θ. The conditional distribution of c_i given c_j for $j \neq i$ (denoted as c_{-i}) is:

$$\begin{aligned} \text{If } c = c_j \text{ for some } j \neq i\text{:}\ \ P(c_i = c \mid c_{-i}) &= \frac{n_{-i,c}}{n - 1 + \nu} \\ P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}) &= \frac{\nu}{n - 1 + \nu} \end{aligned} \tag{4.7}$$

where $n_{-i,c}$ denotes the number of c_j that $c_j = c$ and $j \neq i$. One Metropolis-Hasting algorithm to sample from the posterior distribution of $c = (c_1, \ldots, c_n)$ and θ_C is to repeatedly sample as follows:

1. For $i = 1, \ldots, n$, repeat the following update of c_i R times: Draw a candidate, c_i^* from the conditional distribution for c_i by (4.7). If this c_i^* is not

in $\{c_1, \ldots, c_n\}$, choose a value for $\theta_{c_i^*}$ from H_0. Replace c_i with c_i^* with acceptance probability:

$$a(c_i^*, c_i) = \min\left[1, \frac{F(Z_i, \theta_{c_i^*})}{F(Z_i, \theta_{c_i})}\right] \tag{4.8}$$

Otherwise keep c_i the same.

2. For all $c \in (c_1, \ldots, c_n)$: draw a new value from the posterior distribution $\theta_c \mid Z_i$ such that $c_i = c$.

This is the algorithm 5 in Neal (2000). If the integer R is greater than one, it is possible to save computation time by reusing values of F that were previously computed. Neal (2000) uses $R = 4$ to examine the performance of the algorithm.

Another commonly used MCMC algorithm is sampling with auxiliary parameters (Algorithm 8 in Neal, 2000). Again, let the state of the Markov Chain consist of $\{c_1, \ldots, c_n\}$ and $\{\theta_c : c \in c_1, \ldots, c_n\}$. Let m be a prefixed number of auxiliary parameters. Repeatedly sample as follows:

1. For $i = 1, ..., n$: Let k^- be the number of distinct c_j for $j \neq i$. Label these c_j with values in $\{1, ..., k^-\}$. Let $h = k^- + m$.

2. If $c_i = c_j$ for some $j \neq i$, draw m values independently from H_0 as $\{\theta_{k^-+1}, ..., \theta_h\}$.

3. If $c_i \neq c_j$ for all $j \neq i$, let c_i have the label $k^- + 1$, and draw $m - 1$ valus independently from H_0 as $\{\theta_{k^-+2}, ..., \theta_h\}$.

4. Draw a new value for c_i from $\{1, ..., h\}$ using the following probabilities:
$$P(c_i = c | c_{-i}, \xi_i, \theta_1, ..., \theta_h) = \begin{cases} b\, n_{-i,c}\, F(\xi_i|\theta_c) & \text{for } 1 \leq c \leq k^- \\ b\, \nu/m\, F(\xi_i|\theta_c) & \text{for } k^- < c \leq h \end{cases}$$
where $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c. b is the normalizing constant.

5. For all $c \in \{c_1, ..., c_n\}$: Draw a new value from $\theta_c|\xi_i$ such that $c_i = c$.

The m temporary auxiliary parameters drawn from H_0 are used to approximate the integral in 4.6. With the auxiliary parameters, models with non-conjugate priors can apply Gibbs sampling to update $\{c_1, \ldots, c_n\}$. Neal (2000) showed by simulation that this algorithm is more efficient in terms of autocorrelation time compared to the previous algorithm for models with non-conjugate priors.

4.2 A Semiparametric Bayesian Instrumental Variable Model

4.2.1 The Model and Data

For each subject i, let Y_i be the time-to-event outcome variable, W_i be an unobserved continuous covariate subject to measurement errors, X_i be an observed surrogate of W_i, Z_i be a vector of observed confounders, U_i be a vector of unobserved confounders, and G_i be a vector of instruments, $i = 1, \ldots, n$.

We consider the following two-stage linear model:

$$X_i = \alpha_1{}'G_i + \alpha_2{}'Z_i + \xi_{1i} \tag{4.9}$$

$$Y_i = \beta_1 X_i + \beta_2{}'Z_i + \xi_{2i} \tag{4.10}$$

where the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution with Dirichlet process (DP) prior:

$$(\xi_{1i}, \xi_{2i})' \sim N_2(\mu_i, \Sigma_i) \tag{4.11}$$

$$(\mu_i, \Sigma_i) \sim \text{i.i.d. } H \tag{4.12}$$

$$H \sim \text{DP}(\nu, H_0) \tag{4.13}$$

The variable X_i is often referred to as an endogenous variable in the econometrics literature. The endogenous parameter β_1 is the parameter of primary interest. For each subject i, $\mu_i = (\mu_{1i}, \mu_{2i})'$, and $\Sigma_i = \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix}$. $DP(\nu, H_0)$ in (4.13) is a Dirichlet process prior with strength parameter ν and base distribution H_0. H is a random discrete distribution that has the same support as H_0, where H_0 is usually a continuous distribution. This discreteness of H randomly clusters different (μ_i, Σ_i) together: The parameters μ_i and Σ_i are the same within one

cluster and different across clusters. Note that the marginal distribution of any (μ_i, Σ_i) (by marginalizing out H) is H_0. A detailed introduction of DP prior can be found in section 4.1.1. As a result, the total number of clusters, denoted as k, is random. The posterior distribution of k is determined by both the strength parameter ν and the data. Therefore, the DP prior enables the model to better capture heterogeneity in the error distribution, without using a pre-specified number of clusters. This can relax the parametric assumption of a specific distribution for the previous IV model introduced in Chapter 3, and address for potential heterogeneous clustering problems.

Similar to the parametric model in Chapter 3, this model can be used to adjust for unobserved confounders and measurement errors simultaneously, based on the following derivation: With assumption of linear relationships among the variables Y_i, W_i, U_i, Z_i and G_i, the underlying structure of IV analysis in Figure 3.1 can be modeled by

$$W_i = \alpha_0 + \alpha_1{}'G_i + \alpha_2{}'Z_i + \alpha_3{}'U_i + \varepsilon_{1i} \tag{4.14}$$

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2{}'Z_i + \beta_3{}'U_i + \varepsilon_{2i} \tag{4.15}$$

$$X_i = W_i + \varepsilon_{3i} \tag{4.16}$$

$i = 1, \ldots n$, where ε_{1i}, ε_{2i}, and ε_{3i} are random error in the intermediate covariate W_i, random error in time-to-event outcome Y_i, and measurement error in W_i, respectively. We replace the unobserved W_i in equations (4.14) and (4.15) by $W_i = X_i - \varepsilon_{3i}$ from equation (4.16), and combine the unobserved confounders U_i, random errors $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})$ and intercepts (α_0, β_0) into the correlated random error terms (ξ_{1i}, ξ_{2i}):

$$\xi_{1i} = \alpha_0 + \alpha_3{}'U_i + \varepsilon_{1i} + \varepsilon_{3i} \quad \text{and} \quad \xi_{2i} = \beta_0 + \beta_3{}'U_i + \varepsilon_{2i} - \beta_1\varepsilon_{3i}$$

This gives the two-stage linear model given by equations (4.9) and (4.10).

In time-to-event data that is subject to censoring, the outcome Y_i is not always observed. Let L_i and R_i be two censoring process with $L_i \leq R_i$. Y_i is observed if and only if $L_i = Y_i = R_i$. The censoring indicator $\delta_i = 1$ if $Y_i < L_i$ (left-censored); 2 if $L_i \leq Y_i \leq R_i$ and $L_i < R_i$ (interval-censored); 3 if $Y_i > R_i$ (right-censored); 4 if $L_i = Y_i = R_i$ (event). The primary aim is to estimate the causal effect of W_i on Y_i, i.e. parameter β_1, based on the observed censored data consisting of n independent and identically distributed observations $(L_i, R_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots n$.

4.2.2 Estimation and Inference Procedure

We develop an MCMC procedure to draw inferences on the endogenous parameter β_1. Let $\vec{C} = \{c_1, \ldots, c_n\}$ be the latent class (or "cluster") indicator, and $\theta_C = \{\theta_c : c \in c_1, \ldots, c_n\}$, where $\theta_c = \{\mu_{1c}, \mu_{2c}, \sigma^2_{1c}, \sigma^2_{2c}, \rho_c\}$, i.e. θ_C consists of all distinct values of $\theta_i = \{\mu_{1i}, \mu_{2i}, \sigma^2_{1i}, \sigma^2_{2i}, \rho_i\}$ and $\vec{C}$ is a vector of indicators that maps the individuals to the clusters. Note that the numbering of C can be arbitrary. We denote the total number of clusters as k. For the two-stage IV model (4.9)–(4.13), we denote the parameters as $\Theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_C, \vec{C})$. The observed data consists of $(\vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$, where $\vec{L} = (L_1, ..., L_n)$, $\vec{R} = (R_1, ..., R_n)$, $\vec{\delta} = (\delta_1, ..., \delta_n)$, $\vec{X} = (X_1, ..., X_n)$, $\vec{Z} = (Z_1, ..., Z_n)$ and $\vec{G} = (G_1, ..., G_n)$. Due to censoring of the event times $\vec{Y}$, the likelihood function cannot be derived based on the bivariate distribution given by (4.11) directly. We construct the likelihood function by using the marginal likelihood of the first-stage model (4.9) and the conditional likelihood of the second-stage model (4.10). The likelihood function

is:

$$\begin{aligned}\mathcal{L}(\Theta \mid \vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) &= P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \cdot P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) \\ &= \prod_{i=1}^{n} f_{1i}(X_i, Z_i, G_i) \cdot [1 - S_i(L_i \mid X_i, Z_i, G_i)]^{I\{\delta_i=1\}} \\ &\quad \cdot [S_i(L_i \mid X_i, Z_i, G_i) - S_i(R_i \mid X_i, Z_i, G_i)]^{I\{\delta_i=2\}} \\ &\quad \cdot S_i(R_i \mid X_i, Z_i, G_i)^{I\{\delta_i=3\}} \cdot f_{2i}(L_i \mid X_i, Z_i, G_i)^{I\{\delta_i=4\}}\end{aligned} \tag{4.17}$$

where

$$\begin{aligned} f_{1i}(X, Z, G) &= \phi\left(\frac{X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z}{\sqrt{\sigma_{1i}^2}}\right) \\ f_{2i}(Y \mid X, Z, G) &= \phi\left(\frac{Y - \mu_{2i} - \beta_1 X - \beta_2{}'Z - \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z)}{\sqrt{(1-\rho_i{}^2)\sigma_{2i}^2}}\right) \\ S_i(Y \mid X, Z, G) &= 1 - \Phi\left(\frac{Y - \mu_{2i} - \beta_1 X - \beta_2{}'Z - \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z)}{\sqrt{(1-\rho_i{}^2)\sigma_{2i}^2}}\right) \end{aligned}$$

$i = 1, \ldots n$. $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. The detailed derivation of the likelihood is given in the appendix. Note that functions $f_{1i}(\cdot)$, $f_{2i}(\cdot)$ and $S_i(\cdot)$ are specifically for subject i, since the subjects have different distribution parameters given by the DP prior.

Due to the components involving survival function $S_i(\cdot)$ in the likelihood function, conjugate priors are not available for parameter elements in Θ in the presence of censoring. We propose to use vague and slightly informative independent priors for elements in Θ and the strength parameter ν, and develop an MCMC procedure to generate samples from the joint posterior distribution as summarized below. We use independent priors for simplicity in implementation. Furthermore, using vague joint priors will generate similar results, since the posterior distribu-

tion will be primarily driven by the data. The MCMC algorithm is described in details in the appendix.

For each parameter element in $(\alpha_1, \alpha_2, \beta_1, \beta_2)$: A normal distribution $N(0, \varsigma^2)$ with large variance ς^2 is used as prior distribution. In each MCMC iteration, a regular random walk Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970) is used to update the parameter, while other parameters are fixed at their current states.

Since the individual parameters θ_i's are grouped into k clusters, we can update the cluster indicators $\vec{C}$ and corresponding parameters θ_C, instead of updating θ_i's. The algorithm 8 in Neal (2000) is used for the MCMC sampling. For cluster indicators $\vec{C}$: In each MCMC iteration, we reassign the cluster status of the subjects by updating $(c_1, \ldots, c_n)$ one by one. For each update of c_i, the subject i will be re-assigned to either an existing cluster or a new cluster. The probability of being assigned to a new cluster involves the base distribution H_0 of the DP prior. We set H_0 as a product of independent slightly informative priors for elements in θ_i, i.e. $H_0 = \pi(\mu_{1i})\pi(\mu_{2i})\pi(\sigma_{1i}^2)\pi(\sigma_{2i}^2)\pi(\rho_i)$. Here 'slightly informative' means that the chosen priors spread out and properly cover the reasonable values for the parameters. We use normal distributions for $\pi(\mu_{1i})$ and $\pi(\mu_{2i})$, inverse-gamma distributions for $\pi(\sigma_{1i}^2)$ and $\pi(\sigma_{2i}^2)$, and a uniform distribution Unif$(-1, 1)$ for $\pi(\rho_i)$.

For cluster parameters θ_C: After all cluster indicators in $\vec{C}$ are updated, the corresponding parameters θ_C are updated for each cluster, using data within the cluster only. For this step, we propose to use the regular random walk Metropolis-Hasting method, with vague independent priors: a normal distribution $N(0, \varsigma^2)$ with large variance ς^2 for μ_{1c} and μ_{2c}, an inverse-gamma distribution Inv-Gamma(γ_1, γ_2) with small shape parameter γ_1 and small scale parameter γ_2

for σ_{1c}^2 and σ_{2c}^2, and again a uniform distribution Unif$(-1, 1)$ for ρ_c. We use different priors than H_0 in this step because we want to use priors as non-informative as possible to evaluate the performance of our proposed method under frequentist criteria.

For strength parameter ν: The parameter ν affects the model by affecting the number of clusters, k. Given a fixed sample size n, clusters k increases as ν increases. We will use a prior that gives a reasonable marginal prior for k, $P(k \mid \nu, n)$, that properly spreads out on the reasonable values of k. Antoniak (1974) gives an explicit expression for the marginal distribution of k:

$$P(k \mid \nu, n) = a_n(k)\, n!\, \nu^k \frac{\Gamma(\nu)}{\Gamma(\nu + n)}$$

$k = 1, 2, \ldots, n$, where $a_n(k)$ is a normalizing constant. We use the prior distribution proposed by Conley et al. (2008):

$$P(\nu) \propto \left(\frac{\overline{\nu} - \nu}{\overline{\nu} - \underline{\nu}}\right)^{\omega} \cdot I(\underline{\nu} < \nu < \overline{\nu}) \tag{4.18}$$

where $\underline{\nu}$ and $\overline{\nu}$ are chosen to give small k and large k, respectively. ω is a constant chosen to control the shape of the prior. Therefore, we have the posterior distribution of ν given by $P(\nu \mid k, n) \propto P(\nu)P(k \mid \nu, n)$. Similarly, we use the random walk Metropolis-Hasting method to update ν in each iteration. For all Metropolis-Hasting algorithms mentioned above, uniform proposal distributions are used for the random walk in our simulations and real data analysis, with widths chosen to obtain appropriate acceptance rates.

By iterating the procedure described above, a sufficiently large amount of MCMC samples can be generated from the posterior distribution. Posterior mean of a parameter can be used as an estimation of the parameter. Credible intervals of the parameters can be constructed by using the empirical quartiles of the simulated samples. Convergence of the MCMC algorithm can be examined

visually by graphical methods including trace plots and histograms, and quantitatively by using the Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998). Detailed convergence diagnostics are presented in section 4.5 for the real data examples. We implemented this method in C programming language, due to its relatively fast process in large number of iterations.

4.3 Simulation Studies

Two simulation studies are conducted to assess the performance of our proposed semiparametric Bayesian IV model with Dirichlet process mixture (DPM) error distribution, under frequentist criteria of bias, standard deviation (SD), coverage probability (CP), and width of credible interval (CI). Synthetic data is generated following the underlying model with unobserved confounders and measurement errors given by equations (4.14) to (4.16), with different underlying error distributions, sample sizes, and effects of intermediate covariate. Specifically, we use a simulation setting that is motivated by the WHI-OS real data example in section 4.4.1.

For equation (4.14): A univariate instrument G_i and a univariate unobserved confounder U_i are used, and no observed confounder vector Z_i is considered, i.e. $\alpha_2 = \beta_2 = 0$. G_i and U_i both follow a standard normal distribution $N(0, 1)$; ε_{1i} follows normal distributions $N(0, 0.04)$. The regression parameters are set as $\alpha_0 = 0.5$, $\alpha_1 = \sqrt{0.005}$, $\alpha_3 = \sqrt{0.04}$. For equation (4.16): The measurement error ε_{3i} follows normal distributions $N(0, 0.015)$. This gives $E(X) = 0.5$ and $Var(X) = 0.1$, while $Cor^2(X, G) = 5\%$, similar to the mean and variance of the intermediate covariate and its correlation with the genetic instrument in the WHI-OS example. For equation (4.15): Different values are used for the endogenous parameter: $\beta_1 = (0, -0.5, -1)$, representing none, small, or moderate causal effects, respectively, of underlying true covariate W on outcome Y. These values of β_1 correspond to acceleration factors of 1, 1.6 and 2.7 when the outcome Y is a log-transformed survival time and an accelerated failure time model is used. Other regression parameters are set as $\beta_0 = 5 - \alpha_0\beta_1$ and $\beta_3 = -\sqrt{0.05}$. The random error ε_{2i} is set to have mean 0 and variance 0.45, following different distributions including normal, exponential and two sets of mixtures of normal

distributions. This gives $E(Y) = 5$ and $Var(Y) = 0.5$ when $\beta_1 = 0$, similar to the time-to-event outcome in the WHI-OS example.

In the first simulation study, only right censoring is considered. An underlying right-censoring process, T_i, is conditionally independent of Y_i: $T_i = \beta_0 + \beta_1 X_i + \beta_3 U_i + \varepsilon_{Ti}$. If $Y_i \leq T_i$, then the event time Y_i is observed, $L_i = R_i = Y_i$ and censoring indicator $d_i = 4$; else if $Y_i > T_i$, then Y_i is right-censored, $L_i = R_i = T_i$ and $d_i = 3$. We let ε_{Ti} follow the same distribution as ε_{2i}, giving a 50% right-censoring rate and a 50% event rate, mimicking the WHI-OS example. We use sample size $n = (100, 300, 500)$ for each of the parameter settings.

We first use a regular linear regression survival model (without using IV) to evaluate the bias caused by the unobserved confounder U_i and the measurement error ε_{3i}. We generate 5000 synthetic data sets for each of the simulation settings. The following linear model is applied to each set of observed data $(L_i, R_i, \delta_i, X_i)$, $i = 1, \ldots, n$:

$$Y_i = \beta_1 X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(\mu, \sigma^2) \tag{4.19}$$

We denote this as the 'simple method'. Estimate and 95% confidence interval of the parameters (β_1, μ and σ^2) are derived based on maximum likelihood estimator and asymptotic normal approximation (Klein and Moeschberger, 2003). This can be easily implemented by using the *survreg* function in R package 'survival' (Therneau, 2013) or the LIFEREG procedure (SAS Institute Inc., 2008) in the SAS® software. Expectation and SD of the parameter estimates are approximated by the sample mean and sample SD of the 5000 estimates from independent simulated data sets. Coverage probability is approximated by the proportion of confidence intervals that cover the true value of the parameter.

We then apply our proposed DPM IV model (4.9)–(4.13) to the simulated data. We generate 2000 synthetic data sets for each of the simulation settings

described earlier. For each data set, the likelihood is constructed using observed data $(L_i, R_i, \delta_i, X_i, G_i)$, $i = 1, \ldots, n$ and likelihood function (4.17). Independent vague priors are used for α_1, β_1, and each parameter in θ_C in the cluster parameter update steps: $N(0, 100^2)$ for α_1, β_1, μ_{1c} and μ_{2c}; Inv-Gamma$(0.001, 0.001)$ for σ^2_{1c} and σ^2_{2c}; Unif$(-1, 1)$ for ρ_c, $c = 1, \ldots, k$. Independent slightly informative priors are used for θ_i (i.e. for the base distribution H_0) in the cluster indicator update steps: Let μ_{1t}, μ_{2t}, σ^2_{1t} and σ^2_{2t} denote the true population means of μ_{1i}, μ_{2i}, σ^2_{1i} and σ^2_{2i}, respectively. We use normal distributions with standard deviations equal to 10 and means equal to μ_{1t} and μ_{2t} for μ_{1i} and μ_{2i}, respectively; Inverse-gamma distributions with standard deviations equal to 5 and means equal to σ^2_{1t} and σ^2_{2t} for σ^2_{1i} and σ^2_{2i}, respectively; and Unif$(-1, 1)$ for ρ_i, $i = 1, \ldots, n$. Prior distribution given by (4.18) is used for the strength parameter ν, with $\underline{\nu}$ and $\overline{\nu}$ set as extreme values of ν that will give mode of k equal to 1 and 15, respectively. Specifically, $\underline{\nu} = 0.01$ and $\overline{\nu} = (5, 3.3, 2.9)$ for $n = (100, 300, 500)$, respectively. ω is set as 0.8.

MCMC samples are generated from the posterior distribution. Detailed MCMC algorithm can be found in the appendix. For each simulated data set, we run $110,000$ iterations and use the first $10,000$ iterations as burn-in. The resulting chain is thinned by taking every 5^{th} sample to reduce autocorrelation. Posterior mean, posterior SD and credible interval are derived for α_1, β_1, ν and k based on the resulting $20,000$ posterior samples. Similar to simulations using the simple method described earlier, expectation and SD of the parameter estimate are approximated by the sample mean and sample SD of the 2000 posterior means from independent simulated data sets. Coverage probability is approximated by the proportion of credible intervals that cover the true value of the parameter.

In addition, we also apply the parametric Bayesian IV model with normal error

distribution (3.6)–(3.8) introduced in Chapter 3 to the 2000 synthetic data sets. Similar independent vague priors are used for the parameters as for the DPM IV model: $N(0, 100^2)$ for each of α_0, α_1, β_0, and β_1; Inv-Gamma$(0.001, 0.001)$ for σ_1^2 and σ_2^2, and Unif$(-1, 1)$ for ρ. The MCMC procedure described in Chapter 3 and appendix is used to generate posterior samples for each parameter. Expectation, SD and coverage probability are derived for each parameter following the same procedure as for the DPM IV model.

Results of the parameter of primary interest, β_1, from the three analysis models are summarized in Table 4.1. The β_1 estimates from the simple method have substantial bias in all simulation settings: bias ≈ 0.45, 0.37 and 0.29 for $\beta_1 = 0$, -0.5 and -1, respectively, insensitive to different sample sizes and different error distributions. The coverage probabilities are much lower than the nominal level of 95%, especially when sample size is large (e.g. CP $< 35\%$ when $n = 500$). This is because the SD gets smaller as n increases, while the bias remains large.

When the underlying error distribution is normal (i.e. the parametric IV model is correct), the two proposed IV methods have similar performance in terms of bias, SD, width of 95% and coverage probability of β_1 estimation. Both IV methods have much smaller bias in β_1 estimation than the simple method. For example, when $n = 500$ and $\beta_1 = 0$, both IV methods have bias < 0.02, while the simple method has bias ≈ 0.45. Coverage probabilities of the 95% CIs from both IV methods are slightly higher but close to the nominal level. For both IV methods, bias reduces to close to 0 as sample size n increases.

When the underlying error distribution is non-normal (exponential or mixtures of normal distributions), both IV methods show robustness against deviation of normality assumption: Bias is again largely reduced compared to the simple method, and coverage probability remains close to the nominal level of

95%. While both IV methods have similar good coverage probabilities, the DPM IV model has smaller mean width of 95% CIs, as well as smaller SD of the β_1 estimates, than the normal IV model. For example, the mean 95% CI width of β_1 from the DPM IV model is about one-third shorter when $n = 500$ and error distribution is exponential, and more than half shorter when $n = 500$ and error distribution is Normal Mixture 1 (a mixture of two normal distributions with different means and same variance), compared to the mean 95% CI width of β_1 from the normal IV model. This implies that the DPM IV model can gain more precision, therefore has more statistical power in hypothesis testing of parameter β_1, than the normal IV model, when the true underlying random error is non-normal. Although bias from the DPM IV model appears to be larger than from the normal IV model when $n = 300$ and $n = 500$, this only implies that the posterior mean is not as good an estimator for β_1 in the DPM IV model as in the normal IV model. This is possibly due to the non-normal posterior distribution of β_1 in the DPM IV model. Nevertheless, biases from both IV methods reduce to close to 0 as sample size n increases.

Results of the strength parameter ν and total number of clusters k are summarized in Table 4.2. When the true random error is normal, the average of k is close to 1, indicating that the DPM IV model only consists of one cluster in most of the iterations, which makes the DPM IV model reduce to the normal IV model. This explains why the two IV methods have similar performance. When the true random error is a mixture of two normal distributions (same means with different variance, or different means with same variance), the average of k is close to 2 when $n = 300$ and $n = 500$, indicating that the DPM IV model finds the correct number of clusters. When the true random error is exponential, the average of k is close to 2 when $n = 300$ and $n = 500$. This indicates that the skewed 'tail' of the distribution can be approximated by one additional normal

distribution. These explains how the DPM IV model gains efficiency compared to the normal IV model.

Using the IV methods will result in larger variation in the estimates compared to the simple method, due to the uncertainty from the first-stage model (4.9). The β_1 estimates from the DPM IV model have a two-to-five-fold increase in SD compared to the simple method in our simulation results. This is the cost of using IV methods to reduce bias and make causal inference.

In addition, although two IV methods result in similar mean CI widths when the random error is normal, it does not guarantee that the CI widths from the two methods are always similar in specific simulated data. Figure 4.1 (a) shows the CI width ratio (normal IV model vs. DPM IV model) for the simulation setting of normal random errors, $n = 500$ and $\beta_1 = 0$. 51% of the CI width ratios are smaller than 1, and 1% of the CI width ratios are smaller than 0.7. Similarly, although the semiparametric Bayesian IV method results in shorter mean CI widths than the parametric IV method when the random error is non-normal, it does not guarantee that the CI width from the DPM IV model is always shorter. Figure 4.1 (b) shows the CI width ratio (normal IV model vs. DPM IV model) for the simulation setting of Normal Mixture 2 (a mixture of two normal distributions with same means and different variances) random errors, $n = 500$ and $\beta_1 = 0$. 1% of the CI width ratios are smaller than 1.

We conduct a second simulation study to further examine the performance of the proposed DPM IV model on time-to-event data with arbitrary censoring (left censoring, interval censoring and right censoring). The simulation settings are similar as in the first simulation study except for the censoring processes. Two

underlying censoring processes, T_{1i} and T_{2i}, are conditionally independent of Y_i:

$$T_{1i} = \beta_0 + \beta_1 X_i + \beta_3 U_i + \varepsilon_{T1i}$$
$$T_{2i} = \beta_0 + \beta_1 X_i + \beta_3 U_i + \varepsilon_{T2i}$$

where ε_{T1i} and ε_{T2i} independently follow the same distribution as ε_{2i}, $i = 1, \ldots, n$. The left-censoring time L_i and right-censoring time R_i are set as $L_i = min(T_{1i}, T_{2i})$ and $R_i = max(T_{1i}, T_{2i})$. This gives 1/3 of left-censoring, 1/3 of interval-censoring, and 1/3 of right-censoring. 2000 synthetic data sets are generated for each of the simulation settings. Similar priors and MCMC procedures are used to generate posterior samples for the parameters.

Inferences of β_1, ν and k are summarized in Table 4.3. The results are generally consistent with the ones in the first simulation study: For three different distributions (normal, exponential and mixture normals), biases of β_1 estimation are small and decrease as n increases, and the coverage probabilities are close to the nominal level of 95%. This shows that our proposed semiparametric Bayesian IV model performs well for arbitrarily censored time-to-event data, which is difficult to handle using the classic frequentist approaches.

The average number of clusters k is close to 1 for normal errors, and is around 1.4 for exponential errors. The average k is not correctly specified when the error distribution is a mixture of two normal distributions (mean $k < 1.1$). This is probably due to the simulation setting that all Y_i's are censored, and the censored data does not provide sufficient information for the DPM IV model to identify different clusters. However, the average k increases to close to 2 when we increase the event rate (results not reported here).

Table 4.1: β_1 estimation with and without Instrumental Variable analysis on simulated right-censored data

Error			Simple estimate			Normal IV estimate				DPM IV estimate			
Distribution	n	β_1	Bias	SD	CP	Bias	SD	Wid	CP	Bias	SD	Wid	CP
Normal	100	0	0.449	0.264	0.578	0.156	0.817	5.82	0.982	0.135	0.848	5.79	0.982
		-0.5	0.377	0.273	0.689	0.125	0.838	5.88	0.984	0.152	0.830	5.87	0.986
		-1	0.301	0.276	0.792	0.113	0.858	5.98	0.984	0.114	0.854	6.05	0.990
	300	0	0.449	0.149	0.147	0.066	0.621	3.09	0.973	0.056	0.634	3.17	0.970
		-0.5	0.371	0.150	0.312	0.008	0.642	3.15	0.975	0.023	0.648	3.18	0.967
		-1	0.298	0.159	0.512	0.032	0.634	3.26	0.975	0.021	0.649	3.25	0.979
	500	0	0.448	0.115	0.029	0.007	0.501	2.26	0.970	0.018	0.504	2.29	0.962
		-0.5	0.371	0.118	0.114	0.024	0.522	2.32	0.964	0.039	0.518	2.30	0.963
		-1	0.299	0.122	0.301	0.006	0.523	2.38	0.965	0.027	0.528	2.36	0.963
Exponential	100	0	0.445	0.180	0.277	0.192	0.569	4.03	0.983	0.168	0.543	3.57	0.981
		-0.5	0.371	0.184	0.452	0.158	0.574	4.05	0.982	0.155	0.539	3.67	0.979
		-1	0.297	0.194	0.640	0.110	0.609	4.29	0.987	0.075	0.582	4.00	0.985
	300	0	0.446	0.103	0.011	0.045	0.453	2.14	0.960	0.106	0.327	1.54	0.951
		-0.5	0.372	0.105	0.060	0.046	0.444	2.19	0.972	0.070	0.358	1.74	0.968
		-1	0.295	0.108	0.231	0.028	0.466	2.28	0.969	0.039	0.410	1.99	0.970
	500	0	0.448	0.079	0.000	0.017	0.347	1.56	0.972	0.083	0.248	1.10	0.947
		-0.5	0.373	0.081	0.005	0.006	0.366	1.59	0.955	0.064	0.267	1.20	0.965
		-1	0.297	0.085	0.060	0.004	0.378	1.66	0.964	0.040	0.304	1.35	0.960
Normal	100	0	0.451	0.274	0.599	0.193	0.795	5.99	0.990	0.206	0.697	4.54	0.973
Mixture 1		-0.5	0.373	0.279	0.716	0.137	0.867	6.05	0.982	0.183	0.787	5.00	0.981
		-1	0.296	0.281	0.807	0.124	0.845	6.10	0.990	0.145	0.815	5.67	0.988
	300	0	0.445	0.155	0.183	0.024	0.644	3.24	0.969	0.174	0.267	1.32	0.938
		-0.5	0.373	0.155	0.334	0.029	0.638	3.30	0.974	0.149	0.280	1.45	0.966
		-1	0.297	0.162	0.537	0.021	0.695	3.36	0.969	0.113	0.345	1.73	0.975
	500	0	0.449	0.118	0.035	0.031	0.526	2.35	0.960	0.114	0.227	1.05	0.943
		-0.5	0.374	0.122	0.132	0.019	0.531	2.36	0.964	0.096	0.231	1.13	0.964
		-1	0.293	0.122	0.340	0.015	0.546	2.39	0.957	0.090	0.264	1.27	0.966
Normal	100	0	0.445	0.265	0.558	0.190	0.809	5.68	0.983	0.192	0.669	4.63	0.981
Mixture 2		-0.5	0.372	0.265	0.681	0.138	0.829	5.93	0.990	0.166	0.725	4.96	0.984
		-1	0.303	0.275	0.771	0.124	0.846	5.90	0.984	0.121	0.762	5.09	0.981
	300	0	0.447	0.150	0.156	0.055	0.647	3.14	0.968	0.099	0.434	2.09	0.963
		-0.5	0.373	0.152	0.298	0.019	0.658	3.16	0.962	0.085	0.442	2.21	0.968
		-1	0.295	0.156	0.514	0.047	0.656	3.23	0.968	0.079	0.485	2.35	0.962
	500	0	0.446	0.114	0.032	0.025	0.536	2.29	0.951	0.056	0.358	1.56	0.955
		-0.5	0.370	0.115	0.117	0.027	0.507	2.29	0.963	0.048	0.348	1.63	0.969
		-1	0.297	0.119	0.291	0.003	0.543	2.35	0.958	0.033	0.371	1.70	0.965

Results are based on 2000 simulations using the two IV methods and 5000 simulations using the simple method. 'Error Distribution' refers to distribution of ε_{2i} in equation (4.15) and ε_{Ti} in censoring time T_i. Right censoring rate is 50% and instrument strength is $R^2(X, G) = 0.05$. Bias is the absolute difference between the sample mean of the β_1 estimates and the true value of β_1. Standard deviation (SD) is the sample standard deviation of the β_1 estimates. Coverage probability (CP) is the proportion of 95% credible intervals (for IV estimates) or confidence intervals (for simple estimates) that cover β_1. Wid is the mean width of 95% credible intervals. Normal Mixture 1 is a mixture of two normal distributions $N(-.63, .05) \cdot 0.5 + N(.63, .05) \cdot 0.5$. Normal Mixture 2 is a mixture of two normal distributions $N(0, 0.335^2) \cdot 0.8 + N(0, 1.34^2) \cdot 0.2$.

Table 4.2: Simulation results of strength parameter ν and number of clusters k of the Dirichlet process mixture IV model

Error			Strength parameter ν		Number of clusters k	
Distribution	n	β_1	Mean	SD	Mean	SD
Normal	100	0	0.225	0.012	1.041	0.056
		-0.5	0.226	0.015	1.042	0.068
		-1	0.225	0.015	1.042	0.068
	300	0	0.179	0.008	1.034	0.050
		-0.5	0.179	0.009	1.036	0.055
		-1	0.179	0.008	1.034	0.049
	500	0	0.164	0.007	1.033	0.046
		-0.5	0.164	0.007	1.034	0.042
		-1	0.164	0.006	1.032	0.041
Exponential	100	0	0.299	0.087	1.385	0.402
		-0.5	0.286	0.081	1.321	0.378
		-1	0.273	0.074	1.263	0.341
	300	0	0.334	0.052	1.984	0.311
		-0.5	0.318	0.060	1.890	0.362
		-1	0.290	0.068	1.714	0.418
	500	0	0.331	0.039	2.166	0.247
		-0.5	0.323	0.034	2.115	0.217
		-1	0.306	0.040	2.004	0.265
Normal	100	0	0.312	0.091	1.445	0.422
Mixture 1		-0.5	0.290	0.085	1.340	0.393
		-1	0.259	0.062	1.196	0.290
	300	0	0.345	0.017	2.058	0.101
		-0.5	0.343	0.023	2.044	0.135
		-1	0.332	0.042	1.978	0.253
	500	0	0.313	0.014	2.053	0.087
		-0.5	0.313	0.014	2.054	0.086
		-1	0.313	0.015	2.049	0.096
Normal	100	0	0.331	0.097	1.532	0.448
Mixture 2		-0.5	0.318	0.096	1.472	0.445
		-1	0.310	0.093	1.433	0.431
	300	0	0.342	0.036	2.036	0.218
		-0.5	0.338	0.041	2.008	0.246
		-1	0.331	0.049	1.965	0.296
	500	0	0.319	0.016	2.086	0.100
		-0.5	0.318	0.015	2.079	0.097
		-1	0.316	0.021	2.068	0.139

Results are based on 2000 simulations using the DPM IV model.

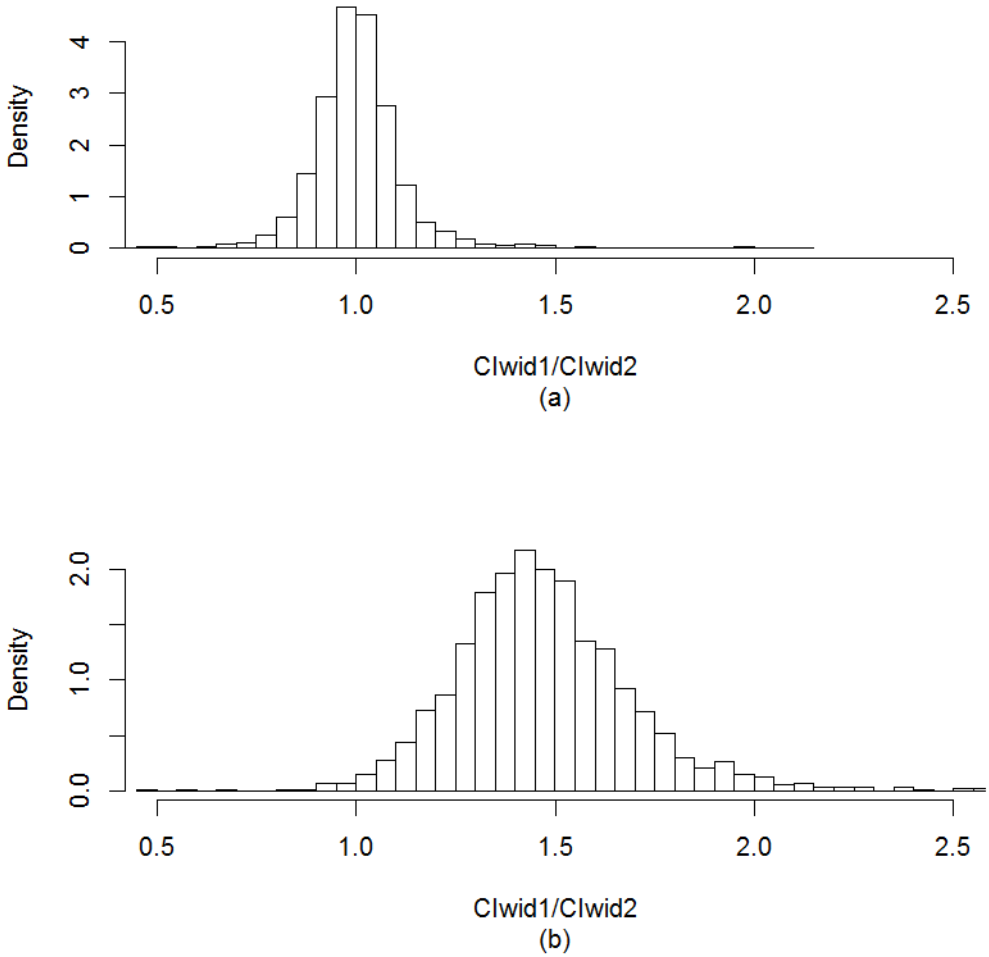
'Error Distribution' refers to distribution of ε_{2i} in equation (4.15) and ε_{Ti} in censoring time T_i. Right censoring rate is 50% and instrument strength is $R^2(X, G) = 0.05$.

Mean and SD are the sample mean and sample standard deviation of the 2000 posterior means, respectively.

Normal Mixture 1 is a mixture of two normal distributions $N(-.63, .05) \cdot 0.5 + N(.63, .05) \cdot 0.5$.

Normal Mixture 2 is a mixture of two normal distributions $N(0, 0.335^2) \cdot 0.8 + N(0, 1.34^2) \cdot 0.2$.

Figure 4.1: Histograms of credible interval width ratios



95% CI width ratio CIwid1/CIwid2 for (a)Normal error distribution, $n = 500$, $\beta_1 = 0$; (b)Normal Mixture 2 error distribution, $n = 500$, $\beta_1 = 0$, from the first simulation study. CIwid1 is 95% credible width by the normal IV model, CIwid2 is 95% credible interval width by the DPM IV model. Normal Mixture 2 is a mixture of two normal distributions $N(0, 0.335^2) \cdot 0.8 + N(0, 1.34^2) \cdot 0.2$.

Table 4.3: Simulation results of the Dirichlet process mixture IV model for simulated data with arbitrary censoring

Error			β_1				Strength parameter ν		Number of clusters k	
Distribution	n	β_1	Bias	SD	Width	CP	Mean	SD	Mean	SD
Normal	300	0	0.052	0.681	3.359	0.971	0.182	0.011	1.055	0.068
		-0.5	0.055	0.697	3.377	0.970	0.182	0.010	1.051	0.062
		-1	0.030	0.689	3.457	0.975	0.182	0.011	1.055	0.065
	500	0	0.039	0.546	2.437	0.955	0.167	0.007	1.050	0.048
		-0.5	0.021	0.556	2.481	0.959	0.167	0.009	1.052	0.058
		-1	0.020	0.558	2.518	0.965	0.167	0.008	1.051	0.050
Exponential	300	0	0.055	0.580	2.822	0.964	0.225	0.060	1.318	0.364
		-0.5	0.025	0.598	2.917	0.965	0.218	0.055	1.272	0.334
		-1	0.029	0.608	3.058	0.972	0.208	0.048	1.210	0.294
	500	0	0.048	0.436	1.980	0.962	0.232	0.063	1.496	0.427
		-0.5	0.017	0.472	2.097	0.957	0.222	0.060	1.426	0.408
		-1	0.010	0.485	2.219	0.966	0.207	0.055	1.324	0.371
Normal	300	0	0.074	0.799	3.836	0.971	0.183	0.016	1.060	0.096
Mixture 1		-0.5	0.036	0.785	3.863	0.965	0.182	0.013	1.055	0.078
		-1	0.044	0.798	3.959	0.966	0.181	0.008	1.050	0.050
	500	0	0.038	0.610	2.732	0.967	0.173	0.029	1.093	0.193
		-0.5	0.021	0.655	2.796	0.949	0.168	0.017	1.062	0.112
		-1	0.017	0.636	2.851	0.961	0.167	0.010	1.051	0.070

Results are based on 2000 simulations using the DPM IV model.

'Error Distribution' refers to distribution of ε_{2i}, ε_{T1i} and ε_{T2i}. Censoring rate is 1/3 for each of left-censoring, interval-censoring and right-censoring. Instrument strength is $R^2(X, G) = 0.05$.

Bias is the absolute difference between the sample mean of the β_1 estimates and the true value of β_1. Mean and SD are the sample mean and sample standard deviation of the 2000 posterior means, respectively. Coverage probability (CP) is the proportion of 95% credible intervals (for IV estimates) or confidence intervals (for simple estimates) that cover β_1. Width is the mean width of 95% credible intervals.

Normal Mixture 1 is a mixture of two normal distributions $N(-.63, .05) \cdot 0.5 + N(.63, .05) \cdot 0.5$.

4.4 Real Data Examples

We illustrate the proposed semiparametric Bayesian IV method using two real data examples. The first one is a prospective case-control study nested within the Women's Health Initiative Observational Study (WHI-OS). The second one is the Atherosclerosis Risk in Communities (ARIC) Study.

4.4.1 Women's Health Initiative Observational Study

In the WHI-OS, our aim is to investigate the effect of high-sensitivity C-reactive protein (hsCRP) on time-to-development of diabetes. hsCRP is an inflammatory marker that has been shown to have positive association with diabetes (Han et al., 2002; Freeman et al., 2002; Liu et al., 2007). However, it is uncertain whether lowering hsCRP level will help prevent diabetes. Therefore, we perform instrumental variable analysis using genetic instruments to make inference about the causal effect of hsCRP on diabetes development, while accounting for potential impact of unobserved confounders and measurement errors.

82069 postmenopausal women (50-59 years of age) with no history of diabetes were followed-up for a mean of 5.5 years in the WHI-OS. 1584 cases of diabetes were identified and matched with 2198 controls by age, ethnicity, clinical center, time of blood draw, and length of follow-up. We focus on the subgroup of whites (954 cases and 968 controls) to avoid the potential problem of population stratification. Time to diabetes diagnosis from baseline was recorded for each case, and time to last visit from baseline was recorded for each control. Plasma concentration of hsCRP is measured for each subject. Descriptive statistics of baseline characteristics are summarized by case-control status in Table 3.3 in Chapter 3. More detailed descriptions of the study can be found in Liu et al. (2007) and

Chan et al. (2011).

In order to perform the IV analysis, 13 haplotype-tagging single-nucleotide polymorphisms (tSNPs) across 2.3 kb of the *CRP* (C-reactive protein, pentraxin-related) genes that had been shown to account for most of the genetic variation within the CRP locus are used as instruments. Details of selection of the tSNPs are given in Lee et al. (2009). None of the 13 tSNPs shows indication of Hardy-Weinberg disequilibrium (all p-values from a Hardy-Weinberg equilibrium test > 0.05 after Bonferroni correction).

We first apply a 'simple method' without using IV, similar to the one in section 4.3, to estimate the association between hsCRP and time to diabetes diagnosis:

$$Y_i = \beta_1 X_i + \beta_2{}' Z_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(\mu, \sigma^2) \tag{4.20}$$

$i = 1, \ldots, n$. For each subject i, Y_i is the log-transformed time to diabetes diagnosis (in days); X_i is the log-transformed hsCRP level; Z_i is a vector of observed potential confounders including age, body mass index, cigarette smoking, alcohol intake, hormone-replacement therapy, family history of diabetes and physical activity. For the cases, the censoring indicator $\delta_i = 4$ (event) and the observed left and right censoring times $L_i = R_i =$ log-transformed time to diabetes diagnosis (in days); for the controls, the censoring indicator $\delta_i = 3$ (right-censored) and the observed left and right censoring times $L_i = R_i =$ log-transformed time to last visit (in days). Since the outcome Y_i is a log-transformed survival time, this model is a log-normal accelerated failure time model. By using the SAS® procedure LIFEREG (SAS Institute Inc., 2008), β_1 has an estimate (SE) of -0.446 (0.089) with a p-value $<.001$, and a corresponding 95% confidence interval of $(-0.621, -0.272)$ as summarized in Table 4.4. This significant negative association between hsCRP and diabetes development is consistent with the results reported by Liu et al. (2007), who analyzed the same data set using the traditional Mendelian Randomiza-

tion method and found that hsCRP was significantly associated with increased diabetes risk (odds ratio = .16 with 95% confidence interval $(1.03, 1.30)$ and a p-value <.001).

We then apply the parametric Bayesian IV model with normal error distribution (3.6)–(3.8) introduced in Chapter 3 to the data. Similar vague priors and MCMC procedure as described in Chapter 3 and appendix are used to generate posterior inferences for each parameter. β_1 has posterior mean (posterior SD) of -0.162 (0.426) and 95% credible interval of $(-0.987, 0.685)$, as summarized in Table 4.4.

We further apply the proposed semiparametric Bayesian IV method with DPM error distribution (4.9)–(4.13) to estimate the causal effect of hsCRP on time to diabetes diagnosis. For each subject i, instrument G_i is a vector of the 13 tSNPs, where each tSNP is coded as an additive effect model, i.e. coded as either 0, 1, or 2 representing the number of minor alleles. L_i, R_i, δ_i, X_i and Z_i are defined as earlier. The likelihood is constructed using observed data $(L_i, R_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots, n$ and likelihood function given by equation (4.17). Independent vague and slightly informative priors are used: $N(0, 100^2)$ for each parameter element in (α_1, α_2, β_1, β_2); $N(0, 100^2)$ for μ_{1c} and μ_{2c}, Inv-Gamma$(0.001, 0.001)$ for σ^2_{1c} and σ^2_{2c}, and Unif$(-1, 1)$ for ρ_c, in the cluster parameter update steps, $c = 1, \ldots, k$; $N(\mu_{1t}, 3^2)$ for μ_{1i}, $N(\mu_{2t}, 10^2)$ for μ_{2i}, Inverse-gamma distribution with mean σ^2_{1t} and SD 1 for σ^2_{1i}, Inverse-gamma distribution with mean σ^2_{2t} and SD 5 for σ^2_{2i}, and Unif$(-1, 1)$ for ρ_i, in the cluster indicator update steps as the base distribution H_0, $i = 1, \ldots, n$, where μ_{1t}, μ_{2t}, σ^2_{1t} and σ^2_{2t} are posterior means of corresponding parameters from the parametric IV model described earlier. Prior distribution given by (4.18) is used for the strength parameter ν, with $\underline{\nu} = 0.01$ and $\overline{\nu} = 2.3$, which are extreme values of ν that will give mode of k equal to 1 and 15, respec-

tively. ω is set as 0.8.

MCMC procedure described in section 4.2.2 and appendix is used to generate posterior samples for the parameters. We generate 40 chains from different initial values, with $1,100,000$ iterations ($100,000$ burn-ins) in each chain. We thin the chains by taking every 10^{th} sample to reduce autocorrelation. A detailed discussion of convergence is given in section 4.5. Posterior mean, posterior standard deviation and credible interval are derived for each parameter element in $(\alpha_1, \alpha_2, \beta_1, \beta_2, \nu, k)$, based on the resulting $4,000,000$ combined samples. Figure 4.3(a) is a histogram of the resulting MCMC samples of β_1. The brackets on the horizontal axis denote the 95% credible interval. The posterior distribution of β_1 has mean 0.230, standard deviation 0.668 and 95% credible interval $(-1.144, 1.629)$, as summarized in Table 4.4. The conclusion is consistent with the previous IV analysis using the normal IV model: No statistically significant causal effect of hsCRP on time to diabetes diagnosis is detected.

The posterior mean (SD) of the strength parameter ν is 0.760 (0.348), and the posterior mean (SD) of the number of clusters k is 5.74 (1.13). We approximate the posterior distribution of bivariate random errors (ξ_1, ξ_2) by using the cluster parameters θ_C and cluster indicators $\vec{C}$ in the last samples of the 40 chains. Figure 4.2 shows the contour plot of the bivariate density. Although the posterior mean of k suggests that the error distribution consists of multiple clusters, the contour plot shows that the error distribution is close to a bivariate normal distribution. This could explain why the DPM IV model results in a longer CI width than the normal IV model.

The results from the three methods suggest that although hsCRP is significantly associated with diabetes development, there is not sufficient evidence of causal effect of hsCRP on time to diabetes diagnosis among white postmenopausal

women. This is consistent with the findings by Brunner et al. (2008), who applied the traditional Mendelian Randomization approach in a case-control study and found that the associations between C-reactive protein (CRP) and diabetes incidence are likely to be noncausal. One possible explanation for the association is that hsCRP level is affected by causal factors of diabetes, such as obesity (Keavney, 2008). On the other hand, even though the sample size of the WHI-OS subgroup analysis is reasonably large, the instruments in our IV analyses may not be strong enough to provide sufficient statistical power to detect small effect sizes (partial R-square = 0.028).

Table 4.4: Two Bayesian approaches of Instrumental Variable (IV) analysis versus simple method in a subgroup analysis of whites within the Women's Health Initiative Observational Study

	estimate of β_1	SE	95% CI
Simple method	-0.446	0.089	(-0.621, -0.272)
Parametric IV method	-0.162	0.426	(-0.987, 0.685)
Semiparametric IV method	0.230	0.668	(-1.144, 1.629)

Simple method: A log-normal accelerated failure time model without using IV to estimate the association between high-sensitivity C-reactive protein (hsCRP) and time to diabetes diagnosis.

Parametric IV method: IV model with normal error distribution.

Semiparametric IV method: IV model with Dirichlet process mixture errors.

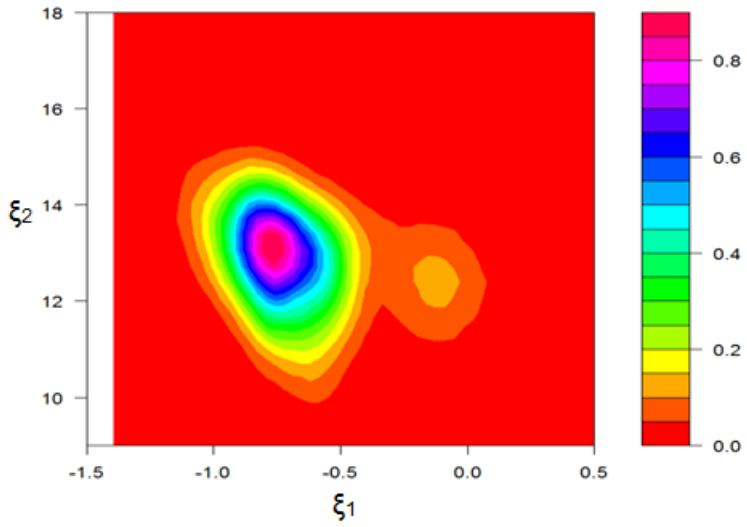
Both IV models estimate the causal effect of hsCRP on time to diabetes diagnosis, using the following 13 selected tSNPs as genetic instruments: rs4275453, rs2808634, rs3093059, rs2794521, rs1417938, rs1800947, rs1130864, rs1205, rs3093075, rs3093068, rs2808629, rs2369146, and rs1470515.

Estimate and SE of β_1 from each of the IV models are posterior mean and posterior standard deviation, respectively.

CI: Confidence interval for the simple model and credible interval for the IV models.

All three models adjust for observed potential confounders including age, body mass index, cigarette smoking, alcohol intake, hormone-replacement therapy, family history of diabetes and physical activity.

Figure 4.2: Density contour plot of random errors (ξ_1, ξ_2) of the Dirichlet process mixture model for the Women's Health Initiative Observational Study



4.4.2 Atherosclerosis Risk in Communities Study

The second real data example is the Atherosclerosis Risk in Communities (ARIC) Study. In this example, we focus on the aspect of measurement error correction of our proposed IV method, and we assume that there is no unobserved confounder (i.e. all confounders are adjusted). Therefore, IV assumption (1) described in section 1.1 is reduced to: Instrument G is independent of measurement errors in intermediate covariate W. The ARIC study is a multi-center prospective cohort study of cardiovascular disease and its risk factors. A total of 15,792 subjects aged 45-64 years were recruited from four US communities in 1987-89. They were planned to receive 4 clinical examinations at 3-year intervals (Visits 1-4). Medical, social and demographic data were collected at each visit. Hospitalization information was obtained by annual telephone follow-up and active surveillance in the communities. A more detailed description of the study is reported elsewhere (The ARIC Investigators, 1989). In this example, we focus on estimating the association between systolic blood pressure (SBP) and development of coronary heart disease (CHD) while correcting for potential bias due to measurement error through IV analysis. For each of Visits 1-4, a subject's SBP level is an average of three measurements. We use Visit 2 (1990-92) as baseline, and use the SBP level at Visit 1 (1987-89) as an instrument of the baseline SBP level. We exclude subjects that (1) have missing baseline information, (2) do not have information after baseline, and/or (3) have developed their first CHD event prior to baseline. After the exclusion, our data consists of 12,782 subjects, 768 of which have CHD events during the follow-up. Descriptive statistics of baseline characteristics are summarized in Table 3.5 in Chapter 3.

For each subject i, outcome Y_i is time to the first CHD event from baseline (Visit 2) in years; covariate of interest X_i is the standardized log-transformed

SBP level at baseline; instrument G_i is the standardized log-transformed SBP level at Visit 1; both X_i and G_i are standardized to have standard deviation 1; Z_i is a vector of observed potential confounders at baseline, including ethnicity (black vs. non-black) and other potential risk factors of CHD developed by the Framingham Heart Study: gender, age, total cholesterol level, high-density lipoprotein cholesterol level, smoking behavior, and diabetes status (Wilson et al., 1998). If subject i has at least one CHD event during the follow-up, censoring indicator $\delta_i = 4$ (event) and the observed left and right censoring times $L_i = R_i = Y_i$; otherwise censoring indicator $\delta_i = 3$ (right-censored) and the observed left and right censoring times $L_i = R_i =$ time to the last visit from baseline in years.

Similar to the previous example in section 4.4.1, we first apply the simple method with equation (4.20) without using IV to estimate the association between SBP and time to CHD. β_1 has an estimate (SE) of -0.779 (0.087) with a p-value $<.001$, corresponding to a 95% confidence interval of $(-0.950, -0.608)$ as summarized in Table 4.5. We then apply the parametric Bayesian IV model with normal error distribution (3.6)–(3.8) to the data. Similar vague priors and MCMC procedure as described in Chapter 3 and appendix are used to generate posterior inferences for each parameter. β_1 has posterior mean (posterior SD) of $-1.180(0.141)$ and 95% credible interval of $(-1.460, -0.907)$, as summarized in Table 4.5.

We further apply the proposed semiparametric Bayesian IV model with DPM error distribution (4.9)–(4.13) to estimate the association between SBP and time to CHD, primarily aiming to correct for potential measurement error bias. Similar to section 4.4.1, the likelihood is constructed using observed data $(L_i, R_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots, n$ and likelihood function given by equation (4.17). Independent vague and slightly informative priors are used: $N(0, 1000^2)$ for each parameter ele-

ment in (α_1, α_2, β_1, β_2); $N(0, 1000^2)$ for μ_{1c} and μ_{2c}, Inv-Gamma$(0.0001, 0.0001)$ for σ^2_{1c} and σ^2_{2c}, and Unif$(-1, 1)$ for ρ_c, in the cluster parameter update steps, $c = 1, \ldots, k$; $N(\mu_{1t}, 10^2)$ for μ_{1i}, $N(\mu_{2t}, 100^2)$ for μ_{2i}, Inverse-gamma distribution with mean σ^2_{1t} and SD 5 for σ^2_{1i}, Inverse-gamma distribution with mean σ^2_{2t} and SD 20 for σ^2_{2i}, and Unif$(-1, 1)$ for ρ_i, in the cluster indicator update steps as the base distribution H_0, $i = 1, \ldots, n$, where μ_{1t}, μ_{2t}, σ^2_{1t} and σ^2_{2t} are posterior means of corresponding parameters from the parametric IV model described earlier. Prior distribution given by (4.18) is used for the strength parameter ν, with $\underline{\nu} = 0.01$ and $\overline{\nu} = 1.7$, which are extreme values of ν that will give mode of k equal to 1 and 15, respectively. ω is set as 0.8.

MCMC procedure described in section 4.2.2 and appendix is used to generate posterior samples for the parameters. We generate 50 chains from different initial values, with $2,200,000$ iterations ($200,000$ burn-ins) in each chain. The chains are thinned to reduce autocorrelation by taking every 20^{th} sample. A detailed discussion of convergence is given in section 4.5. Posterior mean, posterior standard deviation and credible interval are derived for each parameter element in $(\alpha_1, \alpha_2, \beta_1, \beta_2, \nu, k)$, based on the resulting $5,000,000$ combined samples. Figure 4.3(b) is a histogram of the resulting MCMC samples of β_1. The brackets on the horizontal axis denote the 95% credible interval. The posterior distribution of β_1 has mean -1.153, standard deviation 0.141 and 95% credible interval $(-1.432, -0.874)$, as summarized in Table 4.5. The β_1 estimate indicates that a standard deviation increase in log-transformed SBP level is associated with an acceleration of 1.15 years in time to the first CHD event. We observe a larger effect size of SBP on CHD development compared to the simple analysis. This result suggests that the effect size of β_1 calculated by the simple method is possibly attenuated by measurement errors in X_i.

The results from the two IV methods are very similar, indicating that the random error distribution is close to bivariate normal. For the DPM IV model, the posterior mean (SD) of the strength parameter ν is 0.201 (0.143), and the posterior mean (SD) of the number of clusters k is 2.05 (0.22). Although there are on average about 2 clusters, the posterior samples of θ_C and $\vec{C}$ show that one cluster has much smaller weight (number of subjects) than the other, and the bivariate normal error distributions of the two clusters have similar means (μ_{1c}, μ_{2c}) and different covariance matrices (detailed results not reported here).

In the two IV analyses, the SBP measurement at an earlier visit is assumed to be an instrument of the SBP measurement at a later visit. This assumption is weaker than the assumption that both the earlier and later measurements are replicates of noisy surrogate (Carroll et al., 2006; Gustafson, 2007). This is because the latter assumption fixes $\alpha_1 = 1$ while the former assumption does not. Note that the instrument G is not required to be independent of the observed confounders Z, since the confounding effects of Z are adjusted when Z is included in both stages of the model (equations (4.9) and (4.10)). Since a subject's SBP level at certain time point is naturally predictive of his/her SBP level three years later, G_i is a strong instrument of X_i (partial R-square = 0.35). Furthermore, measurement of the instrument does not need to be accurate: Measurement errors in instrument G will not violate the IV assumptions. Therefore, the SBP at Visit 1 can still serve as an instrument if it is also subject to measurement errors.

Table 4.5: Two Bayesian approaches of Instrumental Variable (IV) analysis versus simple method in the Atherosclerosis Risk in Communities (ARIC) Study

	estimate of β_1	SE	95% CI
Simple method	-0.779	0.087	(-0.950, -0.608)
Parametric IV method	-1.180	0.141	(-1.460, -0.907)
Semiparametric IV method	-1.153	0.141	(-1.432, -0.874)

Simple method: a linear regression survival model with normally distributed residuals to estimate the association between standardized log-transformed systolic blood pressure (SBP) level at baseline and time to the first CHD event.

Parametric IV method: IV model with normal error distribution.

Semiparametric IV method: IV model with Dirichlet process mixture errors.

Both IV analyses estimate the association between standardized log-transformed SBP level at baseline and time to first CHD by using standardized log-transformed SBP level at Visit 1 as an instrument to correct for potential bias due to measurement errors in baseline SBP.

Estimate and SE of β_1 from each of the IV models are posterior mean and posterior standard deviation, respectively.

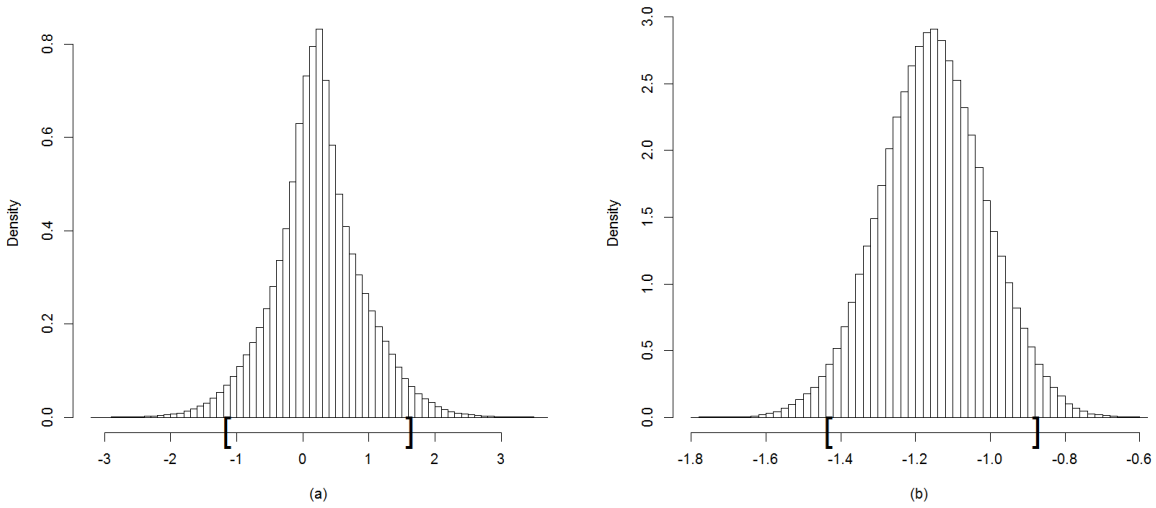
CI: Confidence interval for the simple model and credible interval for the IV models.

All three models adjust for observed potential confounders including gender, age, total cholesterol level, high-density lipoprotein cholesterol level, smoking behavior, and diabetes status.

4.5 MCMC Convergence Diagnostics

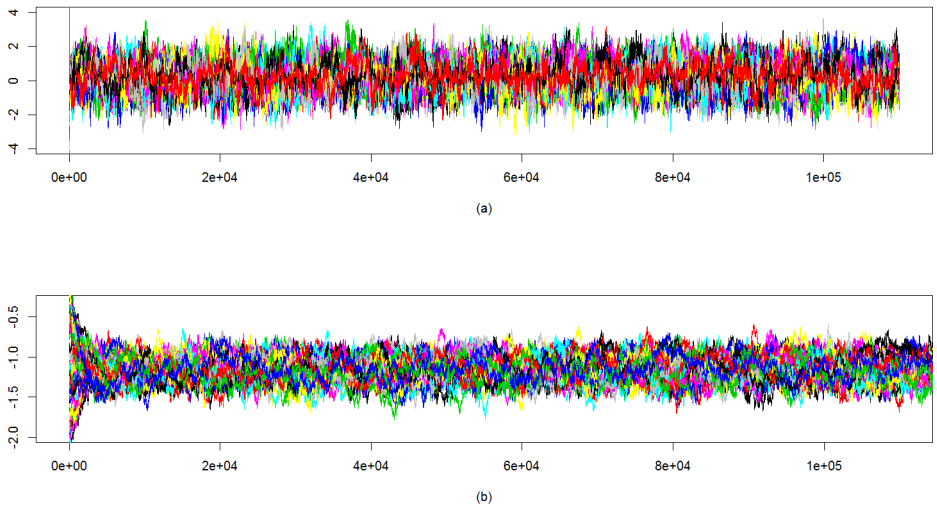
We assess the convergence of MCMC sampling of the semiparametric Bayesian IV method in the two real data examples. Trace plots of parallel chains with diverse initial values are monitored for each parameter in $(\alpha_1, \alpha_2, \beta_1, \beta_2, \nu)$. Figure 4.4 shows the trace plots of β_1: (a) for the WHI-OS data and (b) for the ARIC data. Different chains are marked with different colors. The chains appear to be mixing well and stable over the whole period. We also use the Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998) to measure convergence. The 'potential scale reduction factor' (PSRF) is calculated for each parameter, together with its 95% confidence interval. Approximate convergence is diagnosed when the upper limit of PSRF is close to 1. The 95% upper confidence limit of PSRP for β_1 is 1.015 for the WHI-OS data and 1.038 for the ARIC data, indicating good convergence properties of the method.

Figure 4.3: Histograms of the posterior samples of β_1 from DPM IV model



(a) WHI-OS example; (b) ARIC example. Posterior samples after discarding burn-in and thinning. The brackets denote the limits of the 95% credible interval.

Figure 4.4: Trace plots of the posterior samples of β_1 from DPM IV model



(a) WHI-OS example; (b) ARIC example. Posterior samples after thinning.

CHAPTER 5

Discussion

We have developed two Bayesian approaches for IV analysis to examine the causal effect of an intermediate covariate on a censored time-to-event outcome, in the presence of unobserved confounders and/or measurement errors in the intermediate covariate. We show by simulations that both proposed methods largely reduce bias in estimation and greatly improves coverage probability of the endogenous variable parameter, compared to the 'simple method' where the unobserved confounders and measurement errors are ignored. The parametric Bayesian approach with normality assumption is shown to be robust against deviation from its parametric assumption. However, when the error distribution is non-normal, the semiparametric Bayesian approach with Dirichlet process mixtures has higher precision in estimating the endogenous parameter compared to the parametric Bayesian approach.

Other advantages of the two Bayesian approaches include: (1) It is straightforward to incorporate different types of censoring into these models, which is difficult for the semiparametric methods in the frequentist framework; (2) They do not rely on asymptotic approximations that might be invalid for weak instruments (Lawlor et al., 2008); (3) Prior information can be incorporated by using informative priors for the parameters.

These two methods work well in a variety of settings, provided that the instrumental assumptions described early in introduction are satisfied. It is generally

difficult to validate the instrumental assumptions statistically, since confounder U is unobserved. One possible solution is to extend the Instrumental Variable Bayesian Model Averaging (IVBMA) method proposed by Eicher et al. (2009) to time-to-event outcome. The IVBMA estimate accounts for model uncertainty by taking the weighted average of IV estimates from different potential models, weighted by both the first and second stage posterior model probabilities. The Bayesian Sargan test based on the IVBMA framework could be used to detect violation of the IV assumptions.

Appendix

A1. Likelihood Derivation for the Parametric Bayesian Approach

We use notations defined in Section 3.2.

For time-to-event data subject to right-censoring, the likelihood of observing $(\vec{T}, \vec{\delta})$ is:

$$\mathcal{L} = \prod_{i=1}^{n} f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} \tag{.1}$$

where $S(y) = Pr(Y > y)$ is the survival distribution function and $f(y) = -\frac{d}{dy}S(y)$ is the survival time density function.

Based on the two-stage IV model (3.9) and (3.10), the likelihood function of observing $(\vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$ can be written as:

$$\mathcal{L}(\theta \mid \vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) \quad = \quad P(\vec{T}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \theta) \cdot P(\vec{X}, \vec{Z}, \vec{G} \mid \theta) \tag{.2}$$

where the second part is the marginal likelihood of the first-stage model (3.9), and the first part is the conditional likelihood of the second-stage model (3.10).

For the first part: from the bivariate normality assumption of ξ_1 and ξ_2 following equation (3.11), the conditional distribution of ξ_{2i} given ξ_{1i} is:

$$\xi_{2i} \mid \xi_{1i} \sim N\left(\frac{\sigma_2}{\sigma_1}\rho\,\xi_{1i},\ (1-\rho^2)\sigma_2^2\right)$$

$i = 1, \ldots n$. Since $\xi_{1i} = X_i - \alpha_0 - \alpha_1{}'G_i - \alpha_2{}'Z_i$ from the first-stage model (3.9), the conditional distribution becomes:

$$\xi_{2i} \mid X_i, Z_i, G_i \sim N\left(\frac{\sigma_2}{\sigma_1}\rho(X_i - \alpha_0 - \alpha_1{}'G_i - \alpha_2{}'Z_i),\ \ (1-\rho^2)\sigma_2^2\right)$$

Therefore, given $\vec{X}$, $\vec{Z}$, $\vec{G}$, α_0, α_1 and α_2, the second-stage model (3.10) has conditional survival function

$$\begin{aligned}
S(T \mid X, Z, G) &= P(Y > T \mid X, Z, G) \\
&= P(\beta_0 + \beta_1 X + \beta_2{}'Z + \xi_2 > T) \\
&= P(\xi_2 > T - \beta_0 - \beta_1 X - \beta_2{}'Z) \\
&= 1 - \Phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2{}'Z - \frac{\sigma_2}{\sigma_1}\rho(X - \alpha_0 - \alpha_1{}'G - \alpha_2{}'Z)}{\sqrt{(1-\rho^2)\sigma_2^2}}\right)
\end{aligned}$$

and conditional density function

$$\begin{aligned} f_1(T \mid X, Z, G) &= -\frac{\partial}{\partial t} S(T \mid X, Z, G) \\ &= \phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2{}' Z - \frac{\sigma_2}{\sigma_1}\rho(X - \alpha_0 - \alpha_1{}' G - \alpha_2{}' Z)}{\sqrt{(1-\rho^2)\sigma_2^2}}\right) \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. From (.1), we have

$$P(\vec{T}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \theta) = \prod_{i=1}^{n} f_1(T_i \mid X_i, Z_i, G_i, \theta)^{\delta_i} \, S(T_i \mid X_i, Z_i, G_i, \theta)^{1-\delta_i} \tag{.3}$$

For the second part: the marginal distribution of ξ_{1i} is:

$$\xi_{1i} \sim N(0, \sigma_1^2)$$

which gives the marginal density function for the first-stage model (3.9):

$$f_2(X, Z, G) = \phi\left(\frac{X - \alpha_0 - \alpha_1{}' G - \alpha_2{}' Z}{\sqrt{\sigma_1^2}}\right)$$

Therefore, the likelihood of observing $\vec{X}$, $\vec{Z}$ and $\vec{G}$ is:

$$P(\vec{X}, \vec{Z}, \vec{G} \mid \theta) = \prod_{i=1}^{n} f_2(X_i, Z_i, G_i) \tag{.4}$$

From (.2), (.3) and (.4), we have the joint likelihood function (3.14).

A2. MCMC algorithm for the Parametric Bayesian Approach

We use the MCMC sampling method to generate samples from the posterior distribution of the parameters. In each iteration, a random walk Metropolis-Hasting algorithm is used to update the parameters in $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \rho)$ one by one, while other parameters are fixed at their current state. It updates the current state of the parameter, denoted as z, by candidate z^* generated from a proposal distribution $h(z^*|z)$, with acceptance probability:

$$a(z, z^*) = \min\left(1, \frac{h(z|z^*)\pi(z^*)}{h(z^*|z)\pi(z)}\right)$$

where $\pi(\cdot)$ is a probability density function, in our case, the unstandardized posterior distribution function (i.e. the product of prior distribution and likelihood function) of the parameter. The following is the detailed procedure to update z:

1. Generate a candidate sample z^* from proposal distribution $h(z^*|z)$.
2. Calculate

$$\begin{aligned}\log(a(z, z^*)) &= \log\left(\frac{h(z|z^*)\pi(z^*)}{h(z^*|z)\pi(z)}\right) \\ &= [\log(h(z|z^*)) - \log(h(z^*|z))] + [\log(\mathcal{P}(z)) - \log(\mathcal{P}(z^*))] \\ &\quad +[\log(\mathcal{L}(z)) - \log(\mathcal{L}(z^*))]\end{aligned}$$

   where $\mathcal{P}(\cdot)$ is the prior distribution density function and $\mathcal{L}(\cdot)$ is the likelihood function. In our case, $\mathcal{L}(\cdot)$ is the joint likelihood function (3.14).
3. Generate a random number r from Unif$(0, 1)$.
4. If $\log(a(z, z^*)) > \log(r)$, we accept z^* as the next sample of the parameter; else we keep z as the next sample.
5. Repeat from step 1.

Independent diffuse priors are used for the parameters: a normal distribution $N(\mu, \varsigma^2)$ with large variance (e.g. $\mu = 0$, $\varsigma^2 = 100^2$) for each element in α_0, α_1, α_2, β_0, β_1 and β_2; an inverse-gamma distribution Inv-Gamma(γ_1, γ_2) with small shape parameter and small scale parameter (e.g. $\gamma_1 = \gamma_2 = 0.001$) for σ_1^2 and σ_2^2; and a uniform distribution Unif$(-1, 1)$ for ρ. Uniform proposal distributions are used for the random walk: Unif$(z - \omega, z + \omega)$ for

each element in α_0, α_1, α_2, β_0, β_1 and β_2; Unif$(\max(z-\omega,0), z+\omega)$ for σ_1^2 and σ_2^2; and Unif$(\max(z-\omega,-1), \min(z+\omega,1))$ for ρ. Different positive ω is chosen for each parameter to obtain an appropriate acceptance rate (e.g. $20\% \sim 40\%$ depending on the sample size).

The detailed derivation of the log of acceptance probability for parameters in θ is as follows:

- α_0: Denote the current state and candidate sample as α_0 and α_0^*, respectively. With prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\alpha_0-\omega,\ \alpha_0+\omega)$, the log of acceptance probability:

$$\begin{aligned}\log(a(\alpha_0,\alpha_0^*)) =& \frac{1}{2\varsigma^2}\left((\alpha_0-\mu)^2-(\alpha_0^*-\mu)^2\right)+\sum_{i=1}^{n}\left[\delta_i\left(\frac{1}{2}(q_i^2-{q_i^*}^2)\right)\right.\\ &\left.+(1-\delta_i)\Big(\log(1-\Phi(q_i^*))-\log(1-\Phi(q_i))\Big)+\frac{1}{2}\Big(\nu_i^2-{\nu_i^*}^2\Big)\right]\end{aligned} \tag{.5}$$

  where

$$q_i = \frac{1}{\sqrt{(1-\rho^2)\sigma_2^2}}\left[\log(T_i)-(\beta_0+\beta_1 X_i+{\beta_2}'Z_i)-\frac{\sigma_2}{\sigma_1}\rho(X_i-\alpha_0-{\alpha_1}'G_i-{\alpha_2}'Z_i)\right] \tag{.6}$$

$$\nu_i = \frac{1}{\sqrt{\sigma_1^2}}(X_i-\alpha_0-{\alpha_1}'G_i-{\alpha_2}'Z_i) \tag{.7}$$

  q_i^* and ν_i^* are similar to q_i and ν_i, respectively, both equations with all α_0 replaced by α_0^*.

- α_1: We update α_1 by updating its elements one-by-one. To update the j-th element α_{1j} with candidate sample α_{1j}^*, and with prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\alpha_{1j}-\omega,\ \alpha_{1j}+\omega)$, the log of acceptance probability is similar to (.5), with the first term replaced by $\frac{1}{2\varsigma^2}((\alpha_{1j}-\mu)^2-(\alpha_{1j}^*-\mu)^2)$. q_i and ν_i stay the same as (.6) and (.7). q_i^* and ν_i^* are similar to q_i and ν_i, respectively: Both equations have all α_1 replaced by α_1^*, where α_1^* is α_1 with the j-th element replaced by α_{1j}^*.

- α_2: We update α_2 by updating its elements one-by-one. To update the j-th element α_{2j} with candidate sample α_{2j}^*, and with prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\alpha_{2j}-\omega,\ \alpha_{2j}+\omega)$, the log of acceptance probability is similar to (.5), with the first term replaced by $\frac{1}{2\varsigma^2}((\alpha_{2j}-\mu)^2-(\alpha_{2j}^*-\mu)^2)$. q_i and ν_i stay the same as (.6) and (.7). q_i^* and ν_i^* are similar to q_i and ν_i, respectively: Both equations have all α_2 replaced by α_2^*, where α_2^* is α_2 with the j-th element replaced by α_{2j}^*.

- β_0: We update the current state β_0 with candidate sample β_0^*. With prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif($\beta_0-\omega$, $\beta_0+\omega$), the log of acceptance probability:

$$\begin{aligned}\log(a(\beta_0, \beta_0^*)) =& \frac{1}{2\varsigma^2}\left((\beta_0-\mu)^2 - (\beta_0^*-\mu)^2\right) + \sum_{i=1}^{n}\left[\delta_i\left(\frac{1}{2}(q_i^2 - {q_i^*}^2)\right)\right. \\ &\left. + (1-\delta_i)\Big(\log(1-\Phi(q_i^*)) - \log(1-\Phi(q_i))\Big)\right]\end{aligned} \tag{.8}$$

  where q_i stays the same as (.6). q_i^* is similar to q_i, with β_0 replaced by β_0^*.

- β_1: We update the current state β_1 with candidate sample β_1^*. With prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif($\beta_1-\omega$, $\beta_1+\omega$), the log of acceptance probability is similar to (.8), with the first term replaced by $\frac{1}{2\varsigma^2}((\beta_1-\mu)^2 - (\beta_1^*-\mu)^2)$. q_i stays the same as (.6). q_i^* is similar to q_i, with β_1 replaced by β_1^*.

- β_2: We update β_2 by updating its elements one-by-one. To update the j-th element β_{2j} with candidate sample β_{2j}^*, and with prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif($\beta_{2j}-\omega$, $\beta_{2j}+\omega$), the log of acceptance probability is similar to (.8), with the first term replaced by $\frac{1}{2\varsigma^2}((\beta_{2j}-\mu)^2 - (\beta_{2j}^*-\mu)^2)$. q_i stays the same as (.6). q_i^* is similar to q_i, with β_2 replaced by β_2^*, where β_2^* is β_2 with the j-th element replaced by β_{2j}^*.

- σ_1^2: We update the current state σ_1^2 with candidate sample ${\sigma_1^2}^*$. With prior distribution Inv-Gamma(γ_1, γ_2) and proposal distribution Unif($\max(\sigma_1^2-\omega, 0), \sigma_1^2+\omega$), the log of acceptance probability:

$$\begin{aligned}\log(a(\sigma_1^2, {\sigma_1^2}^*)) =& \Big[\log(\sigma_1^2+\omega-\max(0,\sigma_1^2-\omega)) - \log({\sigma_1^2}^*+\omega-\max(0,{\sigma_1^2}^*-\omega))\Big] \\ &+ \left[(\gamma_1+1)(\log\sigma_1^2 - \log{\sigma_1^2}^*) + \gamma_2\left(\frac{1}{\sigma_1^2} - \frac{1}{{\sigma_1^2}^*}\right)\right] \\ &+ \sum_{i=1}^{n}\left[\delta_i\left(\frac{1}{2}(q_i^2 - {q_i^*}^2)\right) + (1-\delta_i)\Big(\log(1-\Phi(q_i^*)) - \log(1-\Phi(q_i))\Big)\right. \\ &\left. + \frac{1}{2}\Big((\log\sigma_1^2 - \log{\sigma_1^2}^*) + (\nu_i^2 - {\nu_i^*}^2)\Big)\right]\end{aligned}$$

  where q_i and ν_i stay the same as (.6) and (.7). q_i^* and ν_i^* are similar to q_i and ν_i, respectively: Both equations have all σ_1^2 replaced by ${\sigma_1^2}^*$.

- σ_2^2: We update the current state σ_2^2 with candidate sample $\sigma_2^{2^*}$. With prior distribution Inv-Gamma(γ_1, γ_2) and proposal distribution Unif$(\max(\sigma_2^2 - \omega, 0), \sigma_2^2 + \omega)$, the log of acceptance probability:

$$\begin{aligned}\log(a(\sigma_2^2, \sigma_2^{2^*})) = &\Bigg[\log(\sigma_2^2 + \omega - \max(0, \sigma_2^2 - \omega)) - \log(\sigma_2^{2^*} + \omega - \max(0, \sigma_2^{2^*} - \omega))\Bigg] \\ &+ \left[(\gamma_1 + 1)(\log\sigma_2^2 - \log\sigma_2^{2^*}) + \gamma_2 \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_2^{2^*}}\right)\right] \\ &+ \sum_{i=1}^{n} \Bigg[\delta_i \left(\frac{1}{2}(\log\sigma_2^2 - \log\sigma_2^{2^*} + q_i^2 - q_i^{*2})\right) \\ &+ (1 - \delta_i)\Bigg(\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))\Bigg)\Bigg]\end{aligned}$$

where q_i stays the same as (.6). q_i^* is similar to q_i, with all σ_2^2 replaced by $\sigma_2^{2^*}$.

- ρ: We update the current state ρ with candidate sample ρ^*. With prior distribution Unif$(-1, 1)$ and proposal distribution Unif$(\max(\rho - \omega, -1), \min(\rho + \omega, 1))$, the log of acceptance probability:

$$\begin{aligned}\log(a(\rho^2, \rho^*)) = &\Bigg[\log(\min(\rho + \omega, 1) - \max(\rho - \omega, -1)) - \log(\min(\rho^* + \omega, 1) - \max(\rho^* - \omega, -1))\Bigg] \\ &+ \sum_{i=1}^{n} \Bigg[\delta_i \left(\frac{1}{2}(\log(1 - \rho^2) - \log(1 - \rho^{*2}) + q_i^2 - q_i^{*2})\right) \\ &+ (1 - \delta_i)\Bigg(\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))\Bigg)\Bigg]\end{aligned}$$

where q_i stays the same as (.6). q_i^* is similar to q_i, with all ρ replaced by ρ^*.

A3. Likelihood Derivation for the Semiparametric Bayesian Approach with Arbitrary Censoring

We follow notations defined in Sections 4.2.1 and 4.2.2:

For time-to-event data subject to arbitrary-censoring, the likelihood of observing $(\vec{L}, \vec{R}, \vec{\delta})$ is:

$$\mathcal{L} = \prod_{i=1}^{n} (1 - S(L_i))^{I\{\delta_i=1\}} (S(L_i) - S(R_i))^{I\{\delta_i=2\}} S(R_i)^{I\{\delta_i=3\}} f(L_i)^{I\{\delta_i=4\}} \tag{.9}$$

where $S(y) = Pr(Y > y)$ is the survival distribution function and $f(y) = -\frac{d}{dy}S(y)$ is the survival time density function.

Based on the two-stage IV model (4.9) and (4.10), the likelihood function of observing $(\vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$ can be written as:

$$\mathcal{L}(\Theta \mid \vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) \quad = \quad P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \cdot P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) \tag{.10}$$

where the first part is the marginal likelihood of the first-stage model (4.9), and the second part is the conditional likelihood of the second-stage model (4.10).

For the first part: From the bivariate normality assumption of ξ_1 and ξ_2 given by (4.11), the marginal distribution of ξ_{1i} is:

$$\xi_{1i} \sim N(\mu_{1i}, \sigma_{1i}^2)$$

which gives the marginal density function for the first-stage model (4.9):

$$f_{1i}(X, Z, G) \quad = \quad \phi\left(\frac{X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z}{\sqrt{\sigma_{1i}^2}}\right)$$

$i = 1, \ldots n$. Therefore, the likelihood of observing $\vec{X}$, $\vec{Z}$ and $\vec{G}$ is:

$$P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \quad = \quad \prod_{i=1}^{n} f_{1i}(X_i, Z_i, G_i) \tag{.11}$$

For the second part: From the bivariate normality assumption of ξ_1 and ξ_2 given by (4.11), the conditional distribution of ξ_{2i} given ξ_{1i} is:

$$\xi_{2i} \mid \xi_{1i} \sim N\left(\mu_{2i} + \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(\xi_{1i} - \mu_{1i}),\, (1 - \rho_i^2)\sigma_{2i}^2\right)$$

$i = 1, \ldots n$. Since $\xi_{1i} = X_i - \alpha_1{}'G_i - \alpha_2{}'Z_i$ from the first-stage model (4.9), the conditional distribution becomes:

$$\xi_{2i} \mid X_i, Z_i, G_i \sim N\left(\mu_{2i} + \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X_i - \mu_{1i} - \alpha_1{}'G_i - \alpha_2{}'Z_i),\ \ (1-\rho_i^2)\sigma_{2i}^2\right)$$

Therefore, given $\vec{X}$, $\vec{Z}$, $\vec{G}$ and Θ, the second-stage model (4.10) has conditional survival function

$$\begin{aligned} S_i(T \mid X, Z, G) &= P(Y > T \mid X, Z, G) \\ &= P(\beta_1 X + \beta_2{}'Z + \xi_2 > T) \\ &= P(\xi_2 > T - \beta_1 X - \beta_2{}'Z) \\ &= 1 - \Phi\left(\frac{T - \mu_{2i} - \beta_1 X - \beta_2{}'Z - \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z)}{\sqrt{(1-\rho_i^2)\sigma_{2i}^2}}\right) \end{aligned}$$

and conditional density function

$$\begin{aligned} f_{2i}(T \mid X, Z, G) &= -\frac{\partial}{\partial t} S_i(T \mid X, Z, G) \\ &= \phi\left(\frac{T - \mu_{2i} - \beta_1 X - \beta_2{}'Z - \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z)}{\sqrt{(1-\rho_i^2)\sigma_{2i}^2}}\right) \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. From (.9), we have

$$\begin{aligned} P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) &= \prod_{i=1}^{n} [1 - S_i(L_i \mid X_i, Z_i, G_i)]^{I\{\delta_i = 1\}} \\ &\quad \cdot [S_i(L_i \mid X_i, Z_i, G_i) - S_i(R_i \mid X_i, Z_i, G_i)]^{I\{\delta_i = 2\}} \\ &\quad \cdot S_i(R_i \mid X_i, Z_i, G_i)^{I\{\delta_i = 3\}} \cdot f_{2i}(L_i \mid X_i, Z_i, G_i)^{I\{\delta_i = 4\}} \end{aligned} \tag{.12}$$

From (.10), (.11) and (.12), we have the joint likelihood function (4.17).

A4. MCMC algorithm for the Semiparametric Bayesian Approach with Arbitrary Censoring

We follow notations defined in Sections 4.2.1 and 4.2.2. We develop an MCMC procedure to generate posterior samples of α_1, α_2, β_1, β_2, $\theta_i = \{\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i\}$, $\vec{C} = \{c_1, \ldots, c_n\}$, $\theta_c = \{\mu_{1c}, \mu_{2c}, \sigma_{1c}^2, \sigma_{2c}^2, \rho_c\}$, and ν, where $i = 1, \ldots, n$, cluster indicators $\{c_1, \ldots, c_n\}$ are coded as values in $\{1, 2, \ldots, k\}$, k is the total number of clusters, $c = 1, \ldots, k$. In each iteration, we generate a new sample for each of the parameters listed above using the following algorithm.

- For α_1: We update vector α_1 by updating its elements one-by-one using the random walk Metropolis-Hasting (M-H) algorithm described in Section A2. For the j-th element α_{1j}, we propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ with large variance (e.g. $\mu_p = 0$, $\varsigma_p^2 = 100^2$), and a uniform proposal distribution Unif$(\alpha_{1j} - \omega_p, \alpha_{1j} + \omega_p)$ for the random walk, where ω_p is a positive number chosen to give an appropriate acceptance rate (e.g. $20\% \sim 40\%$). A candidate sample α_{1j}^* is generated from the proposal distribution, and accepted as the current state of α_{1j} with probability $a(\alpha_{1j}, \alpha_{1j}^*)$. The log of acceptance probability is given by:

$$\log(a(\alpha_{1j}, \alpha_{1j}^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\alpha_{1j} - \mu_p)^2 - (\alpha_{1j}^* - \mu_p)^2\right)$$

  where Θ^* is Θ with α_{1j} replaced by α_{1j}^*, $\ell(\cdot)$ is the log-likelihood function given by $\ell(\Theta) = log(\mathcal{L}(\Theta))$, and $\mathcal{L}(\Theta)$ is the likelihood function given by equation (4.17).

- For α_2: Similar procedure as for α_1 is used. Elements in vector α_2 is updated one-by-one using the M-H sampling algorithm. For the j-th element α_{2j}, we propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ and a uniform proposal distribution Unif$(\alpha_{2j} - \omega_p, \alpha_{2j} + \omega_p)$ with appropriate width ω_p. A candidate sample α_{2j}^* is generated from the proposal distribution, and accepted as the current state of α_{2j} with probability $a(\alpha_{2j}, \alpha_{2j}^*)$. Similarly, the log of acceptance probability is given by:

$$\log(a(\alpha_{2j}, \alpha_{2j}^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\alpha_{2j} - \mu_p)^2 - (\alpha_{2j}^* - \mu_p)^2\right)$$

  where Θ^* is Θ with α_{2j} replaced by α_{2j}^*.

- For β_1: We update β_1 using the M-H sampling algorithm, similar to the procedure for α_{1j}. We propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ and a uniform

proposal distribution Unif($\beta_1 - \omega_p$, $\beta_1 + \omega_p$) with appropriate width ω_p. A candidate sample β_1^* is generated from the proposal distribution, and accepted as the current state of β_1 with probability $a(\beta_1, \beta_1^*)$. Similarly, the log of acceptance probability is given by:

$$\log(a(\beta_1, \beta_1^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\beta_1 - \mu_p)^2 - (\beta_1^* - \mu_p)^2\right)$$

where Θ^* is Θ with β_1 replaced by β_1^*.

- For β_2: Similar procedure as for α_1 is used. Elements in vector β_2 is updated one-by-one using the M-H sampling algorithm. For the j-th element β_{2j}, we propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ and a uniform proposal distribution Unif($\beta_{2j} - \omega_p$, $\beta_{2j} + \omega_p$) with appropriate width ω_p. A candidate sample β_{2j}^* is generated from the proposal distribution, and accepted as the current state of β_{2j} with probability $a(\beta_{2j}, \beta_{2j}^*)$. Similarly, the log of acceptance probability is given by:

$$\log(a(\beta_{2j}, \beta_{2j}^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\beta_{2j} - \mu_p)^2 - (\beta_{2j}^* - \mu_p)^2\right)$$

where Θ^* is Θ with β_{2j} replaced by β_{2j}^*.

- For $\vec{C}$: We update the cluster indicators $c_1, \ldots, c_n$, one-by-one. Let m be a prefixed number of auxiliary parameters. We use $m = 10$ in our simulation studies and real data examples in Chapter 4. For the base distribution H_0 of the Dirichlet process prior, we propose to use independent slightly informative priors $H_0 = \pi(\mu_{1i})\pi(\mu_{2i})\pi(\sigma_{1i}^2)\pi(\sigma_{2i}^2)\pi(\rho_i)$. Here 'slightly informative' means that the chosen priors spread out and properly cover the reasonable values for the parameters. We propose to use normal distributions for $\pi(\mu_{1i})$ and $\pi(\mu_{2i})$, inverse-gamma distributions for $\pi(\sigma_{1i}^2)$ and $\pi(\sigma_{2i}^2)$, and a uniform distribution Unif$(-1, 1)$ for $\pi(\rho_i)$. The following procedure is used to update cluster indicator c_i:

    1. For subject i: Let k^- be the number of distinct c_j for $j \neq i$. Let $h = k^- + m$, and $c^{-i} = \{c_j : j \neq i\}$.

    2. If $c_i = c_j$ for some $j \neq i$ (i.e. subject i is not a 'singleton'), draw m samples independently from H_0 as $\{\theta_{k^-+1}, ..., \theta_h\}$ (i.e. draw m independent samples from $\pi(\mu_{1i})$ as $\{\mu_{1,k^-+1}, ..., \mu_{1h}\}$, draw m independent samples from $\pi(\mu_{2i})$ as $\{\mu_{2,k^-+1}, ..., \mu_{2h}\}$, draw m independent samples from $\pi(\sigma_{1i}^2)$ as $\{\sigma_{1,k^-+1}^2, ..., \sigma_{1h}^2\}$, draw m independent samples from $\pi(\sigma_{2i}^2)$ as $\{\sigma_{2,k^-+1}^2, ..., \sigma_{2h}^2\}$, draw m independent samples from $\pi(\rho_i)$ as $\{\rho_k^- + 1, ..., \rho_h\}$).

3. If $c_i \neq c_j$ for all $j \neq i$ (i.e. subject i is a 'singleton'), relabel these c_j with values in $\{1, ..., k^-\}$, and label c_i as $k^- + 1$. Draw $m - 1$ samples independently from H_0 as $\{\theta_{k^-+2}, ..., \theta_h\}$.

4. Draw a new value for c_i from $\{1, ..., h\}$ with probabilities:
$$P(c_i = c|c^{-i}, \theta_1, \ldots, \theta_h) = \begin{cases} b \cdot n_{-i,c} \cdot L_i(\theta_c) \ , & 1 \leq c \leq k^- \\ b \cdot \frac{\nu}{m} \cdot L_i(\theta_c) \ , & k^- \leq c \leq h \end{cases}$$
where $n_{-i,c}$ is the number of subjects that are in $\{j : j \neq i, c_j = c\}$, and $L_i(\theta_c)$ is the likelihood of subject i with parameter θ_c:
$$L_i(\theta_c) = \mathcal{L}(\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_c \mid L_i, R_i, \delta_i, X_i, Z_i, G_i)$$
and b is a normalizing constant.

5. Update the total number of clusters k accordingly.

- For θ_c: We update cluster parameters θ_c, $c = 1, \ldots, k$, one-by-one. For each $c \in \{1, \ldots, k\}$, we update $\{\mu_{1c}, \mu_{2c}, \sigma^2_{1c}, \sigma^2_{2c}, \rho_c\}$ one-by-one, while keeping the other parameters at their current state, using the M-H sampling algorithm. We propose to use independent vague priors for the parameters: a normal distribution $N(\mu, \varsigma^2)$ with large variance (e.g. $\mu = 0$, $\varsigma^2 = 100^2$) for μ_{1c} and μ_{2c}; an inverse-gamma distribution Inv-Gamma(γ_1, γ_2) with small shape parameter and small scale parameter (e.g. $\gamma_1 = \gamma_2 = 0.001$) for σ^2_{1c} and σ^2_{2c}; and a uniform distribution Unif$(-1, 1)$ for ρ_c.

  - For μ_{1c}: We use a uniform proposal distribution Unif$(\mu_{1c} - \omega_p, \mu_{1c} + \omega_p)$ with appropriate width ω_p. A candidate sample μ^*_{1c} is generated from the proposal distribution, and accepted as the current state of μ_{1c} with probability $a(\mu_{1c}, \mu^*_{1c})$. The log of acceptance probability is given by:
$$\log(a(\mu_{1c}, \mu^*_{1c})) = \ell_c(\Theta^*) - \ell_c(\Theta) + \frac{1}{2\varsigma_p^2}\left((\mu_{1c} - \mu_p)^2 - (\mu^*_{1c} - \mu_p)^2\right)$$
where Θ^* is Θ with μ_{1c} replaced by μ^*_{1c}, and $\ell_c(\cdot)$ is the log-likelihood function with subjects in cluster c only,
$$\ell_c(\Theta) = log(\mathcal{L}(\Theta \mid L_i, R_i, \delta_i, X_i, Z_i, G_i, i \in \{j : c_j = c\}))$$

  - For σ^2_{1c}: We use a uniform proposal distribution Unif$(\max(\sigma^2_{1c} - \omega_p, 0), \sigma^2_{1c} + \omega_p)$ with appropriate width ω_p. A candidate sample ${\sigma^2_{1c}}^*$ is generated from the proposal

distribution, and accepted as the current state of σ_{1c}^2 with probability $a(\sigma_{1c}^2, {\sigma_{1c}^2}^*)$. The log of acceptance probability is given by:

$$\begin{aligned}\log(a(\sigma_{1c}^2, {\sigma_{1c}^2}^*)) =& \ell_c(\Theta^*) - \ell_c(\Theta) \\ &+ \log(\min(2\omega_p, \sigma_{1c}^2 + \omega_p)) - \log(\min(2\omega_p, {\sigma_{1c}^2}^* + \omega_p)) \\ &+ (\gamma_1 + 1)\left[\log(\sigma_{1c}^2) - \log({\sigma_{1c}^2}^*)\right] + \gamma_2\left(\frac{1}{\sigma_{1c}^2} - \frac{1}{{\sigma_{1c}^2}^*}\right)\end{aligned}$$

where Θ^* is Θ with σ_{1c}^2 replaced by ${\sigma_{1c}^2}^*$.

– For μ_{2c}: Similar to μ_{1c}, we use a uniform proposal distribution Unif($\mu_{2c}-\omega_p$, $\mu_{2c}+\omega_p$) with appropriate width ω_p. A candidate sample μ_{2c}^* is generated from the proposal distribution, and accepted as the current state of μ_{2c} with probability $a(\mu_{2c}, \mu_{2c}^*)$. The log of acceptance probability is given by:

$$\log(a(\mu_{2c}, \mu_{2c}^*)) = \ell_c(\Theta^*) - \ell_c(\Theta) + \frac{1}{2\varsigma_p^2}\left((\mu_{2c} - \mu_p)^2 - (\mu_{2c}^* - \mu_p)^2\right)$$

where Θ^* is Θ with μ_{2c} replaced by μ_{2c}^*.

– For σ_{2c}^2: Similar to σ_{1c}^2, we use a uniform proposal distribution Unif($\max(\sigma_{2c}^2 - \omega_p, 0), \sigma_{2c}^2 + \omega_p$) with appropriate width ω_p. A candidate sample ${\sigma_{2c}^2}^*$ is generated from the proposal distribution, and accepted as the current state of σ_{2c}^2 with probability $a(\sigma_{2c}^2, {\sigma_{2c}^2}^*)$. The log of acceptance probability is given by:

$$\begin{aligned}\log(a(\sigma_{2c}^2, {\sigma_{2c}^2}^*)) =& \ell_c(\Theta^*) - \ell_c(\Theta) \\ &+ \log(\min(2\omega_p, \sigma_{2c}^2 + \omega_p)) - \log(\min(2\omega_p, {\sigma_{2c}^2}^* + \omega_p)) \\ &+ (\gamma_1 + 1)\left[\log(\sigma_{2c}^2) - \log({\sigma_{2c}^2}^*)\right] + \gamma_2\left(\frac{1}{\sigma_{2c}^2} - \frac{1}{{\sigma_{2c}^2}^*}\right)\end{aligned}$$

where Θ^* is Θ with σ_{2c}^2 replaced by ${\sigma_{2c}^2}^*$.

– For ρ_c: We use a uniform proposal distribution Unif($\max(\rho_c - \omega_p, -1), \min(\rho_c + \omega_p, 1)$) with appropriate width ω_p. A candidate sample ρ_c^* is generated from the proposal distribution, and accepted as the current state of ρ_c with probability $a(\rho_c, \rho_c^*)$. The log of acceptance probability is given by:

$$\begin{aligned}\log(a(\rho_c, \rho_c^*)) =& \ell_c(\Theta^*) - \ell_c(\Theta) + \log(\min(\rho_c + \omega_p, 1)) - \log(\max(\rho_c - \omega_p, -1)) \\ &- \log(\min(\rho_c^* + \omega_p, 1)) + \log(\max(\rho_c^* - \omega_p, -1))\end{aligned}$$

where Θ^* is Θ with ρ_c replaced by ρ_c^*.

- For θ_i: After updating $\vec{C}$ and θ_c, $c = 1, \ldots, k$, the individual parameters $\theta_i = \{\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i\}$, $i = 1, \ldots, n$, can be derived.

- For ν: We update the strength parameter ν of the Dirichlet process prior using the M-H sampling algorithm. We propose to use prior distribution

$$P(\nu) \propto \left(\frac{\overline{\nu} - \nu}{\overline{\nu} - \underline{\nu}}\right)^{\omega} \cdot I(\underline{\nu} < \nu < \overline{\nu})$$

where $\underline{\nu}$ and $\overline{\nu}$ are chosen to give small k (e.g. mode of $k = 1$) and large k (e.g. mode of $k = 15$), respectively. ω is a constant chosen to control the shape of the prior (e.g. $\omega = 0.8$). We use a uniform proposal distribution $\text{Unif}(\max(\underline{\nu}, \nu - \omega_p), \min(\overline{\nu}, \nu + \omega_p))$. A candidate sample ν^* is generated from the proposal distribution, and accepted as the current state of ν with probability $a(\nu, \nu^*)$. The log of acceptance probability is given by:

$$\begin{aligned}
\log(a(\nu, \nu^*)) =& \log(\min(\overline{\nu}, \nu + \omega_p) - \max(\underline{\nu}, \nu - \omega_p)) \\
& - \log(\min(\overline{\nu}, \nu^* + \omega_p) - \max(\underline{\nu}, \nu^* - \omega_p)) \\
& + \omega_p \left[\log(\overline{\nu} - \nu^*) - \log(\overline{\nu} - \nu)\right] \\
& + k(\log\nu^* - \log\nu) + \log(\Gamma(\nu^*)) - \log(\Gamma(\nu^* + n)) \\
& - \log(\Gamma(\nu)) + \log(\Gamma(\nu + n))
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function.

Bibliography

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):pp. 431–442, 1995. ISSN 01621459.

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. ISSN 01621459.

Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

Govert E. Bijwaard. Instrumental variable estimation for duration data. Tinbergen Institute Discussion Papers 08-032/4, Tinbergen Institute, March 2008. URL `http://ideas.repec.org/p/dgr/uvatin/20080032.html`.

David Blackwell and James B. Macqueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

Roger J. Bowden and Darrell A. Turkington. *Instrumental Variables*, volume 8. Cambridge University Press, 1984.

Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, December 1998.

Eric J. Brunner, Mika Kivimäki, Daniel R. Witte, Debbie A. Lawlor, George Davey Smith, Jackie A. Cooper, Michelle Miller, Gordon D. Lowe, Ann Rumley, Juan P. Casas, Tina Shah, Steve E. Humphries, Aroon D. Hingorani, Michael G. Marmot, Nicholas J. Timpson, and Meena Kumari. Inflammation, insulin resistance, and diabetes–Mendelian randomization using CRP haplotypes points upstream. *PLoS medicine*, 5(8), August 2008.

Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3): 429–436, December 1979.

Christopher A. Bush and Steven N. MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

Jeffrey S. Buzas and Leonard A. Stefanski. Instrumental variables estimation in generalized linear measurement error models. *Journal of the American Statistical Association*, 91:999–1006, 1996.

Raymond J. Carroll and Leonard A. Stefanski. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13 (12):1265–82, 1994.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall/CRC, 2 edition, June 2006. ISBN 1584886331.

K.H. Chan, K. Brennan, N.C. You, X. Lu, Y. Song, Y.H. Hsu, G. Chaudhuri, L. Nathan, L. Tinker, and S. Liu. Common variations in the genes encoding c-reactive protein, tumor necrosis factor-alpha, and interleukin-6, and the risk of clinical diabetes in the women's health initiative observational study. *Clinical Chemistry*, 57(2):317–325, 2011.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995. URL `http://dx.doi.org/10.2307/2684568`.

M.A. Chmielewski. Elliptically symmetric distributions: a review and bibliography. *International Statistical Review*, 49:67–74, 1981.

Timothy G. Conley, Christian B. Hansen, Robert E. McCulloch, and Peter E. Rossi. A semiparametric bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, May 2008.

David R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34(2):187–220, 1972.

D.R. Cox and D. Oakes. *Analysis of survival data*. Chapman & Hall, 1984.

George Davey Smith and Shah Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.*, 32 (1):1–22, February 2003.

George Davey Smith and Shah Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol*, 33(1):30–42, 2004.

Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4):309–330, August 2007.

Eric L. Ding, Yiqing Song, JoAnn E. Manson, David J. Hunter, Cathy C. Lee, Nader Rifai, Julie E. Buring, J. Michael Gaziano, and Simin Liu. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *The New England journal of medicine*, 361(12): 1152–1163, September 2009.

James Durbin. Errors in variables. *Review of the International Statistical Institute*, 22:23–32, 1954.

Theo Stefan Eicher, Alex Lenkoski, and Adrian Raftery. Bayesian model averaging and endogeneity under model uncertainty: An application to development determinants. Working Papers UWEC-2009-19-FC, University of Washington, Department of Economics, 2009.

Paul Elliott, John C. Chambers, Weihua Zhang, Robert Clarke, Jemma C. Hopewell, John F. Peden, Jeanette Erdmann, Peter Braund, James C. Engert, Derrick Bennett, Lachlan Coin, Deborah Ashby, Ioanna Tzoulaki, Ian J. Brown, Shahrul Mt-Isa, Mark I. McCarthy, Leena Peltonen, Nelson B. Freimer, Martin Farrall, Aimo Ruokonen, Anders Hamsten, Noha Lim, Philippe Froguel, Dawn M. Waterworth, Peter Vollenweider, Gerard Waeber, Marjo-Riitta R. Jarvelin, Vincent Mooser, James Scott, Alistair S. Hall, Heribert Schunkert, Sonia S. Anand, Rory Collins, Nilesh J. Samani, Hugh Watkins, and Jaspal S. Kooner. Genetic loci associated with c-reactive protein levels and risk of coronary heart disease. *JAMA : the journal of the American Medical Association*, 302(1):37–48, July 2009.

Michael D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, June 1995.

Kai-Tai Fang and T.W. Anderson. *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Press, New York, 1990.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

D.J. Freeman, J. Norrie, M.J. Caslake, A. Gaw, I. Ford, G.D. Lowe, D.S. O'Reilly, C.J. Packard, and N.C. Sattar. C-reactive protein is an independent predictor of risk for the development of diabetes in the west of scotland coronary prevention study. *Diabetes*, 51:1596–1600, 2002.

Wayne A. Fuller. *Measurement Error Models*. Wiley, New York, 1987.

Els Goetghebeur and Stijn Vansteelandt. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research*, 14:397–415, 2005.

Arthur S. Goldberger. Structural equation methods in the social sciences. *Econometrica*, 40 (6):979–1001, 1972.

Richard Gray and Keith Wheatley. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant*, 7, Suppl 3:9–12, 1991.

Paul Gustafson. Measurement error modelling with an approximate instrumental variable. *Journal of the Royal Statistical Society: Series B*, 69(5):797–815, 2007.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

Trygve Haavelmo. The probability approach in econometrics. *Econometrica*, 12, Suppl.:iii–115, 1944.

Alastair R. Hall. *Generalized Method of Moments*. New York: Oxford University Press, 2005.

T.S. Han, N. Sattar, K. Williams, C. Gonzalez-Villalpando, M.E. Lean, and S.M. Haffner. Prospective study of c-reactive protein in relation to the development of diabetes and metabolic syndrome in the mexico city diabetes study. *Diabetes Care*, 25:2016–2021, 2002.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–54, July 1982.

James W. Hardin. The robust variance estimator for two-stage models. *Stata Journal*, 2(3): 253–266(14), 2002.

James W. Hardin and Raymond J. Carroll. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *Stata Journal*, 3(4):342–350(9), 2003.

James W. Hardin, Henrik Schmiediche, and Raymond J. Carroll. Instrumental variables, bootstrapping, and generalized linear models. *Stata Journal*, 3(4):351–360(10), 2003.

Frank E. Harrell. *Design: Design Package*, 2009. URL `http://CRAN.R-project.org/package=Design`. R package version 2.3-0.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

James J. Heckman. Econometric causality. Working Paper 13934, National Bureau of Economic Research, April 2008.

James J. Heckman and V. Joseph Hotz. Choosing among alternative non-experimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, 84:862–880, 1989.

James J. Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1-2):239–267, 1985.

Arne Risa Hole. Calculating murphy-topel variance estimates in stata: A simplified procedure. *Stata Journal*, 6(4):521–529, 2006.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Lennart Hoogerheide, Frank Kleibergen, and Herman K. van Dijk. Natural conjugate priors for the instrumental variables regression model applied to the AngristKrueger data. *Journal of Econometrics*, 138:63–103, May 2007.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the berkeley symposium on mathematical statistics and Probability*, 1967.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for Stick-Breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the dirichlet process. *Can J Statistics*, 30(2):269–283, 2002.

Sonia Jain and Radford M. Neal. Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.

Pia R. Kamstrup, Anne Tybjaerg-Hansen, Rolf Steffensen, and Børge G. Nordestgaard. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *The Journal of the American Medical Association*, 301(22):2331–2339, June 2009.

E.L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Göran Kauermann and Raymond J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

Bernard Keavney. More evidence against a causal association between c-reactive protein and diabetes. *PLoS Med*, 5(8):e174, 08 2008.

Mika Kivimäki, Costan G. Magnussen, Markus Juonala, Mika Kähönen, Johannes Kettunen, Britt-Marie Loo, Terho Lehtimäki, Jorma Viikari, and Olli T. Raitakari. Conventional and mendelian randomization analyses suggest no association between lipoprotein(a) and early atherosclerosis: the young finns study. *International Journal of Epidemiology*, 40(2):470–478, April 2011.

Frank Kleibergen and Herman K. Van Dijk. Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14:701–743, 1998.

Frank Kleibergen and Eric Zivot. Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114(1):29–72, 2003.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, second edition, February 2003. ISBN 038795399X.

Helmut Kuchenhoff. The identification of logistic regression models with errors in the variables. *Statistical Paper*, 36:41–48, 1995.

Tze L. Lai and Zhiliang Ying. Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics*, 19(3):1370–1402, 1991.

Tony Lancaster. *An introduction to modern Bayesian econometrics. Chapter 8, Instrumental Variables*. Wiley-Blackwell, 2004.

Zinoviy M. Landsman and Emiliano A. Valdez. Tail conditional expectations for elliptical distributions. *North American Actuarial Journal*, 7:55–71, 2003.

Kenneth L. Lange, Roderick J.A. Little, and Jeremy M.G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, December 1989.

Debbie A. Lawlor, Roger M. Harbord, Jonathan A. C. Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–63, 2008.

Cathy C. Lee, Nai-Chieh Y. You, Yiqing Song, Yi-Hsiang Hsu, Joann Manson, Lauren Nathan, Lesley Tinker, and Simin Liu. Relation of Genetic Variation in the Gene Coding for C-Reactive Protein with Its Plasma Protein Concentrations: Findings from the Women's Health Initiative Observational Cohort. *Clinical Chemistry*, 55(2):351–360, February 2009. doi: 10.1373/clinchem.2008.117176.

Simin Liu, Lesley Tinker, Yiqing Song, Nader Rifai, Denise E. Bonds, Nancy R. Cook, Gerardo Heiss, Barbara V. Howard, Gokhan S. Hotamisligil, Frank B. Hu, Lewis H. Kuller, and JoAnn E. Manson. A prospective study of inflammatory cytokines and diabetes mellitus in a multiethnic cohort of postmenopausal women. *Archives of internal medicine*, 167(15): 1676–1685, 2007.

Steven N. Maceachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications In Statistics - Simulation and Computation*, 23(3):727–741, 1994.

Steven N. Maceachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, June 1998.

David B. Marden and Dimitris G. Manolakis. Using elliptically contoured distributions to model hyperspectral imaging data and generate statistically similar synthetic data. *Proceedings of SPIE*, 5425:558–572, 2004.

MATLAB. *version 7.1.* The MathWorks Inc., Natick, MA, 2005.

Paul M. McKeigue, Harry Campbell, Sarah Wild, Veronique Vitart, Caroline Hayward, Igor Rudan, Alan F. Wright, and James F. Wilson. Bayesian methods for instrumental variables analysis with genetic instruments ("mendelian randomization"): example with urate transporter slc2a9 as instrumental variable for effect of urate levels on metabolic syndrom. *Journal of Computational and Graphical Statistics*, 739:907–918, 2010.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Mary S. Morgan. *The History of Econometric Ideas.* Cambridge University Press, 1991.

Kevin M. Murphy and Robert H. Topel. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 3(4):370–379, 1985.

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, June 2000.

Dorothea Nitsch, Mariam Molokhia, Liam Smeeth, Bianca L. DeStavola, John C. Whittaker, and David A. Leon. Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *American journal of epidemiology*, 163(5):397–403, March 2006.

Ikechukwu U. Ogbuanu, Hongmei Zhang, and Wilfried Karmaus. Can we apply the mendelian randomization methodology without considering epigenetic effects? *Emerging themes in epidemiology*, 6:3+, May 2009.

Tom M. Palmer, John R. Thompson, Martin D. Tobin, Nuala A. Sheehan, and Paul R. Burton. Adjusting for bias and unmeasured confounding in mendelian randomization studies with binary responses. *International journal of epidemiology*, 37(5):1161–1168, 2008.

Judea Pearl. *Causality: models, reasoning, and inference*, chapter 7. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Ya'acov Ritov. Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18(1):303–328, 1990.

D. Roodman. Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal*, 11(2):159–206, 2011.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B. Rubin. Bayesian inference for causal effects. *The Annals of Statistics*, 6:34–58, 1978.

SAS Institute Inc. *SAS/STAT® 9.2 User's Guide*. SAS Institute Inc., Cary, NC, 2008.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Janez Stare, Frank E. Harrell, and Harald Heinzl. Bj: an s-plus program to fit linear regression models to censored data using the buckley-james method. *Computer Methods and Programs in Biomedicine*, 64(1):45–52, 2001.

StataCorp. *Stata 12 Base Reference Manual*. Stata Press, College Station, TX, 2011.

George Thanassoulis, Catherine Y. Campbell, David S. Owens, J. Gustav Smith, Albert V. Smith, Gina M. Peloso, Kathleen F. Kerr, Sonali Pechlivanis, Matthew J. Budoff, Tamara B. Harris, Rajeev Malhotra, Kevin D. O'Brien, Pia R. Kamstrup, Brge G. Nordestgaard, Anne Tybjaerg-Hansen, Matthew A. Allison, Thor Aspelund, Michael H. Criqui, Susan R. Heckbert, Shih-Jen Hwang, Yongmei Liu, Marketa Sjogren, Jesper van der Pals, Hagen Klsch,

Thomas W. Mhleisen, Markus M. Nthen, L. Adrienne Cupples, Muriel Caslake, Emanuele Di Angelantonio, John Danesh, Jerome I. Rotter, Sigurdur Sigurdsson, Quenna Wong, Raimund Erbel, Sekar Kathiresan, Olle Melander, Vilmundur Gudnason, Christopher J. O'Donnell, and Wendy S. Post. Genetic associations with valvular calcification and aortic stenosis. *New England Journal of Medicine*, 368(6):503–512, 2013. PMID: 23388002.

The ARIC Investigators. The atherosclerosis risk in communities (aric) study: design and objectives. *American Journal of Epidemiology*, 129:687–702, 1989.

Henry Theil. *Economic forecasts and policy*. North-Holland, Amsterdam, 1958.

Terry M Therneau. *A Package for Survival Analysis in S*, 2013. URL `http://CRAN.R-project.org/package=survival`. R package version 2.37-4.

Duncan Thomas and David Conti. Commentary: the concept of 'mendelian randomization'. *International journal of epidemiology*, 33(1):21–25, 2004. URL `http://dx.doi.org/10.1093/ije/dyh048`.

George L. Wehby, Robert L. Ohsfeldt, and Jeffrey C. Murray. 'mendelian randomization' equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, 27: 2745–2749, 2008.

Mike West, Peter Muller, and Michael D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P. Freeman and A. Smith, editors, *Aspects of Uncertainty*, pages 363–386. John Wiley, 1994.

Halbert White. A Heteroskedasticity-Consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, May 1998.

Philip G. Wright. *Appendix to The Tariff on Animal and Vegetable Oils*. MacMillan, New York, NY, USA, 1928.

Guosheng Yin, Yanyuan Ma, Faming Liang, and Ying Yuan. Stochastic generalized method of moments. *Journal of Computational and Graphical Statistics*, 20(3):714–727, 2011.