Bayesian Mixture Modeling Approaches for Intermediate Variables and Causal Inference

by

Scott Lee Schwartz

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Fan Li, Co-Supervisor

_____
Jerome P. Reiter, Co-Supervisor

_____
Alan E. Gelfand

_____
Marie Lynn Miranda

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2010

ABSTRACT
(Statistics)

Bayesian Mixture Modeling Approaches for Intermediate Variables and Causal inference

by

Scott Lee Schwartz

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Fan Li, Co-Supervisor

_____
Jerome P. Reiter, Co-Supervisor

_____
Alan E. Gelfand

_____
Marie Lynn Miranda

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2010

Abstract

This thesis examines causal inference related topics involving intermediate variables, and uses Bayesian methodologies to advance analysis capabilities in these areas. First, joint modeling of outcome variables with intermediate variables is considered in the context of birthweight and censored gestational age analyses. The proposed methodology provides improved inference capabilities for birthweight and gestational age, avoids post-treatment selection bias problems associated with conditional on gestational age analyses, and appropriately assesses the uncertainty associated with censored gestational age. Second, principal stratification methodology for settings where causal inference analysis requires appropriate adjustment of intermediate variables is extended to observational settings with binary treatments and binary intermediate variables. This is done by uncovering the structural pathways of unmeasured confounding affecting principal stratification analysis and directly incorporating them into a model based sensitivity analysis methodology. Demonstration focuses on a study of the efficacy of influenza vaccination in elderly populations. Third, flexibility, interpretability, and capability of principal stratification analyses for continuous intermediate variables are improved by replacing the current fully parametric methodologies with semi-parametric Bayesian alternatives. This presentation is one of the first uses of nonparametric techniques in causal inference analysis, and opens a connection between these two fields. Demonstration focuses on two studies, one involving a cholesterol reduction drug, and one examine the effect of physical activity on cardiovascular disease as it relates to body mass index.

Dedicated to my role models, Kenneth and Cheryl Schwartz –

You are admired.

Contents

Abstract **iv**

List of Figures **x**

List of Tables **xiii**

Abbreviations and Notations **xv**

Acknowledgements **xviii**

1 Introduction **1**

1.1 Birthweight and Censored Gestational Age: Joint Modeling of Intermediate and Outcome Variables 2

1.2 Causal Inference and Principal Stratification: Adjusting for Intermediate Variables in Observational Settings 6

1.2.1 Causal Inference 8

1.2.2 Principal Stratification Sensitivity Analysis 14

1.3 Flexible Principal Stratification for Continuous Intermediate Variables: Bayesian Semi-parametric Modeling 16

2 Inferential Benefits of Joint Modeling of Birthweight and Gestational Age **23**

2.1 Introduction 23

2.1.1 Modeling Approaches for Birthweight and Gestational Age . . . 24

2.1.2 Data Application: NCDBR 26

2.2 Joint Birthweight and Gestational Age Model 27

2.2.1 Likelihood Specification 27

2.2.2 Additional Specification 29

2.2.3 Censored Continuous Gestational Age 30

2.3 Bivariate Modeling Vs. Conditional Modeling 31

2.4 Identifiability 36

2.4.1 Alternative Non-Identified Parameterization 36

2.4.2 More Identifiability and Number of Components 36

2.5 Model Demonstration 37

2.5.1 Bivariate Regression 38

2.5.2 Mixture Sub-Populations 40

2.5.3 Prediction 42

2.5.4 Bivariate Distribution 44

2.6 Discussion 45

3 Extension of Binary Principal Stratification into Observational Settings 47

3.1 Introduction 47

3.2 Confounding in PS 49

3.2.1 Characterizing Unmeasured Confounding 50

3.2.2 Implications of Unmeasured Confounding and a False Exclusion Restriction 51

3.2.3 Illustration of Confounding 55

3.3 Parametric Sensitivity Analysis 56

3.3.1 Estimation of CACE with Sensitivity Parameters 59

3.3.2 Sensitivity Demonstration Using Introduced Confounding . . . 61

3.4 Applying Sensitivity Analysis in Practice 65

3.5 Discussion 69

4 Flexible Bayesian Semi-Parametric PS Modeling for Continuous Intermediate Variables 71

4.1 Introduction 71

4.2 Models and Computation 75

4.2.1 Overview of Principal Stratification 75

4.2.2 Bayesian semi-parametric model 78

4.2.3 Posterior inference 81

4.3 Application to randomized trial with partial compliance 84

4.3.1 Data and Models 84

4.3.2 Results 87

4.4 Application to the Swedish National March Cohort 90

4.4.1 Data and Models 90

4.4.2 Results 92

4.5 Discussion 95

5 Conclusion and Future Direction **99**

5.1 Extensions to Joint Modeling Analyses 99

5.2 Extension to Sensitivity Analyses 101

5.3 Extensions to PS for Continuous Intermediate Variables 102

Bibliography **112**

Biography **113**

List of Figures

1.1 This DAG represents gestational age D as an intermediate variable: The risk factor, smoking T, may affect the birthweight outcome Y though gestational age, or perhaps via some other pathway. 4

1.2 DAGs (a) through (c) are representations of the three key conditional independence rules: (a) Intermediate Variable Case: $T \perp\!\!\!\perp Y|D$; (b) Confounder Case: $T \perp\!\!\!\perp Y|X$; (c) Ancestor Case: $T \perp\!\!\!\perp X$ but $T \not\perp\!\!\!\perp X|D$ or A. DAG (d) demonstrates the results of the backdoor criterion violation of conditioning on the intermediate variable, D. Namely, spurious correlation between the treatment T and the outcome Y is induced via X as a result of conditioning on the intermediate variable D. 9

1.3 This figure demonstrates the multi-treatment causal inference setting. There are multiple treatments T_m over time that may affect a final outcome of interest Y as well as future variables D_m, but may also be affected by previous D_m and Y_m. 10

2.1 Histograms of birthweight by gestational age (g, 24 to 42) for the data subset described in 2.1.2. 27

2.2 A log-scale heatmap version of a bivariate histogram of birthweight and gestational ages for the data subset described in 2.1.2. 27

2.3 (a) shows a treatment or risk factor T affecting the joint variable of birthweight and gestational age. (b) adjusts the setting to reflect the complication of mis-measured gestational age. 33

2.4 (a) demonstrates a setting with potential for a back-door criterion violation. (b) The back-door criterion violation is realized if g or $g^c + u$ are conditioned on. 34

2.5 Conditioning on the residuals $\hat{\epsilon}_{g|T}$ does not result in a back-door criterion violation. 35

2.6 Posterior point estimate of the component configurations for individuals $\mathcal{A}$ and $\mathcal{H}$. The ellipses correspond to contours containing $\approx 86.5\%$ of component mass. The thickness conveys the relative proportions in the mixture distribution of Table 2.5. 40

2.7 Conditional predictions of the small for gestational age cutpoint $SGA(g)$ for individuals $\mathcal{A}$ and $\mathcal{H}$. The single gestational age axis is separated into 3 plots so that the 95% credible intervals may be examined. The predictions were generated from the conditional distributions implied by the joint distributions represented in Figures 2.6. Table 2.9 provides related results for other individuals. 43

2.8 Point estimate of the surface of the mixture distribution for birthweight and gestational age for the referent individual $\mathcal{H}$. The orientation of this plot is a nonstandard $\approx 180^\circ$ rotational form. As a result, birthweight increases from top to bottom and gestational age decreases from left to right. Posterior 95% credible intervals of the surface tightly fit this curve, and so were not included in this image. 45

3.1 Possible sources of confounding: (a) None, (b) None, but direct effect of T on Y, (c) O (outcome), (d) Unmeasured, (e) S-confounding, and (f) Y-confounding. 52

3.2 Illustrations of sensitivity contour plots for manipulated McDonald data. The top plot shows MLE contours for $\hat{\theta}_c$ across the possible combinations of ξ_s with $\eta_c = \delta_{\{a,n\}} = 0$. The bottom plot shows the same when $\eta_c = \delta_{\{a,n\}} = 0.25$, which are the true values. The dashed cross-hairs are at the approximately correct S-confounding sensitivity parameter values, $\exp(\xi_a) = 2$ and $\exp(\xi_n) = 1/2$. The dashed curve in the bottom plot indicates where $\hat{\theta}_c$ equals θ_c; this curve does not appear in the top plot because it is off the graph. The plots show that standard PS estimates of θ_c are biased in the presence of unmeasured confounding, and it is possible to recover the true θ_c when correct sensitivity parameters are used. 64

3.3 Posterior probability of assignment of zero positive flu outcomes to the complier groups in the treatment and control arms from the original McDonald data as a function of δ_a and δ_n, assuming no S-confounding. Vertically oriented lines show probabilities for compliers in the treatment group, and horizontally oriented lines show probabilities for compliers in the control group. These graphs can be used to determine ranges of implausible values of sensitivity parameters, e.g., where the probabilities often equal zero (here, when $\delta_a > 0$ and $\delta_n < -.1$). . . . 67

3.4 Posterior medians and 95% intervals for θ_c for original McDonald data as a function of δ_a and δ_n, assuming no S-confounding. In the bottom panel, the dotted lines represent upper limits and the solid lines represent lower limits of the intervals. The 95% intervals always contain zero, indicating that the conclusions from the standard PS estimation are not overly sensitive to unmeasured confounding in this range of sensitivity paramters. 68

4.1 A single posterior imputation for $(D(0), D(1))$ based on the DPM S-model. The three predominant clusters appear to be continuous analogues of always-takers, never-takers, and defiers from binary PS. . . . 89

4.2 Median PCE over the entire $(D(0), D(1))$ space, as estimated by the DPM S-model. 90

4.3 (a) is the histogram of age across T. (b) is the histogram of BMI across T. (c) is the qqplot of BMI of $T_i = 0$ versus $T_i = 1$ group: 98% of the points lie above the diagonal. (d) is the scatterplot of age versus CVD incidence, displaying a clear positive correlation. (e) is the scatterplot of age versus BMI. (f) is the scatterplot of BMI versus CVD incidence 97

4.4 A representative posterior draw of principal strata S_i under the DPM S-model. Each component is labeled with a number and a light dot representing its mass contribution, e.g., component 1 contributes 80%. The solid line is the 45^o line. 98

4.5 Median surface and point-wise 95% credible intervals for the PCE, over the relevant space of $(D(0), D(1))$ for individuals 10 years above the median age (60 years old). The green surface is the reference surface of PCE = 0. 98

List of Tables

2.1 Birthweight (standard) regressions with and without Gestational Age included. 95% confidence intervals are included. 32

2.2 Initial values and prior distribution specifications for the results of a three ($H = 3$) component model used through out Section 2.5. 38

2.3 Individuals $\mathcal{A}$ through $\mathcal{H}$ provide 8 risk factor sets for mothers used for demonstration in Section 2.5 and tables 7 through 10. All mothers are 25-30 years old and at the high school education level with a male infant. 39

2.4 Birthweight and gestational age regression coefficients with 95% credible intervals for in each mixture model component. 39

2.5 Location and shape parameter estimates with 95% credible intervals for each mixture model component. 40

2.6 The compositional makeup, given in percentages (except for the first row, which is a count) of each mixture model component as observed in posterior sampling. Additionally, the final column labeled 'Overall' shows the characteristics of the original population. 41

2.7 Conditional expectation of birthweight given gestational age along with 95% credible interval for individuals $\mathcal{A}$ through $\mathcal{H}$ at 34, 37, 39, and 40 weeks. 44

2.8 Conditional expectation of gestational age given birthweight along with 95% credible interval for individuals $\mathcal{A}$ through $\mathcal{H}$ at 1500, 2500, 3500, and 4000 grams. 44

2.9 SGA cutpoint predictions and 95% credible interval for individuals $\mathcal{A}$ through $\mathcal{H}$ at gestational ages 34, 37, 39, and 40. 44

2.10 Probability estimates and 95% credible intervals for simultaneous LBW and PTB, $AI(35)$, and $AI(37+)$ for individuals $\mathcal{A}$ through $\mathcal{H}$ at gestational ages 34, 37, 39, and 40 weeks. 46

3.1 Observed proportions in the McDonald (1992) study. 55

3.2 Example of population probabilities with no S-confounding and no Y-confounding that are consistent with observed data in McDonald study. 56

3.3 Example of population probabilities with S-confounding but no Y-confounding that are consistent with observed data in McDonald study. 56

3.4 Example of population probabilities with Y-confounding but no S-confounding that are consistent with observed data in McDonald study. 57

4.1 Coefficients in the outcome Y-model (4.5) as estimated using the Bayesian DPM S-model, where posterior medians and 95% credible intervals are shown; and the frequentist copula S-model of BG, where MLE and SEs are shown. 88

4.2 Estimated PCE for selected principal stratum (D_0, D_1) using the DPM approach, the fully parametric approach of JR, and the copula approach of BG. 89

4.3 Posterior medians and 95% credible intervals for the coefficients in the Y-model and S-model. 93

4.4 Posterior medians and 95% credible intervals for the percent PCE, $E(Y(1) - Y(0)|S = (d_0, d_1)) \times 100$, for selected principal strata S at age of 50 and 60 years. 94

Abbreviations and Notations

BIC	Bayesian information criterion
BMI	Body mass index
CACE	Complier average causal effect θ_c
COPD	Chronic obstructive pulmonary disease
CVD	Cardiovascular disease
DAG	Directed acyclic graph
DP	Dirichlet process
DPM	Dirichlet process mixture model
ER	Exclusion restriction
EM	Expectation-Maximization
HD	Heart Disease
HS	Hispanic
LBW	Low birthweight (less than 2500 grams)
LMP	Last menstural period
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
NCDBR	North Carolina Detailed Birth Record
NHW	Non-Hispanic white
NHB	Non-Hispanic black
NMC	National March cohort
PA	Physical activity
PCE	Principal causal effect
PS	Principal stratification

PSDE	Principal strata direct effect
PSIE	Principal strata indirect effect
PTB	Preterm birth (less than 37 weeks gestational age)
RCM	Rubin causal model
SB	Stick breaking
SGA	Small for gestational age (10% birthweight quantile for gestational age)
SUTVA	Stable unit treatment value assumptions

$1_{[x']}(x)$	Indicator function equal to 1 if $x = x'$, 0 otherwise; (x) may be omitted
a	Always-takers
A	A space
b	Birthweight
c	Compliers
d	Defiers
D	Intermediate variable
D^{obs}	Observed potential intermediate variable $D(T)$
$D(\cdot)$	Potential intermediate variable
g	Gestational age
g^c	Censored gestational age
G	Random probability measure
G	DP base measure
h	Component h in finite mixture model
H	Total number of components in finite mixture model
i	subject
K	Kernel, or density
F	Density or distribution
n	Never-takers
N	Sample size

p	Probability
Pr	Probability, density, or distribution
S	Principal strata $(D(1), D(0))$
T	Treatment or risk factor
u	Unknown and imputed decimal part of censored gestational age
U	Unmeasured confounder
w_h	Probability of component h in finite mixture model
X	Covariates
Y	Outcome variable
Y^{obs}	Observed potential outcome $Y(T)$
$Y(\cdot)$	Potential outcome
Z	Latent component membership indicator

α	DP strength parameter
β	Linear model coefficients
δ_a	Y-confounding and direct effect for always-takers
δ_n	Y-confounding and direct effect for never-takers
η_a	Y-confounding for always-takers
η_c	Y-confounding for compliers
η_n	Y-confounding for never-takers
σ^2	Variances
Σ	Covariance matix
τ_a	Direct effect for always-takers
τ_n	Direct effect for never-takers
θ	Generic parameter
θ_c	Complier average causal effect
ξ_n	S-confounding for never-takers
ξ_a	S-confounding for always-takers

Acknowledgements

My graduate school years have been like no other. The personal and intellectual growth I've experienced at Duke is unmatched by any other period in my life. I am more aware than I have ever been of who I am, who I want to be, and who I will become – I am filling up the space that was meant for me. The research presented here is a proud achievement of substance and merit, but to me, it represents only a fraction of what I have learned from my time at Duke.

For that, I thank my family – Mommy, Daddy, Daniel, Kendra, Jommy, and Katie – for always being there and believing in me; Emily for reminding me why I work; Fan Li, for her continued support, patience, understanding, and willing attention, Jerry Reiter for his generosity with his time, and his example; Alan Gelfand, for always wanting the best for me; Marie Lynn Miranda, for providing me the opportunity to discover my passion, and putting up with me while I did; Dalene Stangl, for guiding my growth as a teacher; and Hao and Avi, and all the others who gave me a sense of esprit de core as I made my way through the more trying times of graduate school. I thank you all.Special thanks to Fabrizia Mealli for valuable suggestions influencing the direction of Chapter 4.

Chapter 1

Introduction

This thesis focuses on statistical inference for intermediate variables – post-treatment variables affected by a treatment and affecting an outcome. Intermediate variables are important for many statistical analyses, particularly causal analyses, but they may not be dealt with in the same way as standard pretreatment covariates. In this thesis, three methodologies related to intermediate variables are introduced: (1) joint modeling of birthweight and gestational age, (2) sensitivity analysis for unmeasured confounding in observational studies while appropriately adjusting for intermediate variables, and (3) using a Dirichlet process mixture model (DPM) for intermediate variable submodels in causal inference analyses. Although these topics comprise three separate chapters, a theme underpinning the methodology in each is the advantage of Bayesian modeling approaches. In addition, all three methodologies have strong connections to mixture modeling and causal inference techniques and ideas. The flow, intentions, and relationships between remaining chapters of this thesis are as follows.

Chapter 2 introduces finite mixture models in the setting of joint birthweight and censored gestational age modeling. This chapter advances modeling capabilities currently available in the reproductive epidemiology literature, and clearly demonstrates the benefits of improved modeling perspectives. The methodology is this section is appropriate for general settings involving treatment of intermediate variables as joint variables, both with an with-

out the presence of censored or missing data. In addition, this chapter builds connections with causal inference in the presence of intermediate variables.

Chapter 3 builds on the notions of intermediate variables discussed in Chapter 2, and examines the causal inference methodology of principal stratification (PS) to adjust for intermediate variables. The assumptions of PS analysis are thoroughly and carefully studied for the case of binary treatments and binary intermediate variables in order to understand the implications of confounding in the binary setting. Based on these developments, binary PS is extended into the observational setting through a general sensitivity analysis methodology that explores the effects of possible unmeasured confounding in a given PS analyses.

Chapter 4 continues the focus on PS from Chapter 3, and considers the extension of PS to continuous intermediate variables. Using DPM methodology – a nonparametric form of the mixture modeling techniques of chapter 2 – the flexibility, interpretability, and capabilities of continuous PS analysis are markedly improved. In addition, as one of the first examples of the use of nonparametric Bayesian modeling techniques in causal inference analyses, the chapter also provides a general connection between these two important areas, and opens the possibility for future connections to be made.

To begin, Chapter 1 gives a high level overview of the work as it relates to the relevant associated literature. Following this introduction, the remaining chapters reconsider each topic in depth in the context of an applied data example. Chapter 5 concludes with future research opportunities arising from the methodologies developed in this thesis.

1.1 Birthweight and Censored Gestational Age: Joint Modeling of Intermediate and Outcome Variables

In the United States, particularly Durham, North Carolina and the South, African American mothers experience more adverse birth outcomes than white mothers. Parinatal (fetal, neonatal) and infant mortality and morbidity rates are higher in African American pop-

ulations. And, late arriving manifestations of birth complications such as cerebral palsy, asthma, low IQ, hypertension, heart diseases, diabetes, and impairments of hearing and vision are more prevalent in African American populations. These disparities hold even after controlling for relevant personal characteristics and socio-economic indicators. It is unclear what pathways drive the disparities and what must be done to correct them. Further, there is no evidence from other countries suggesting that such disparities should be expected, so their existence is indeed unsettling. Because of this lack of understanding, a major push in developmental epidemiology research is to uncover the sources of birth outcome disparities and to develop effective intervention programs to reduce them.

Birthweight and low birthweight (LBW, birthweight less than 2500 grams), and gestational age and preterm birth (PTB, gestational age less than 37 weeks) are frequently studied for this purpose. This is because they clearly demonstrate the disparity between African American and white mothers, and because they associate with adverse birth outcomes. Of course, birthweight and gestational age are highly correlated themselves and, from a medical viewpoint, are together the relevant joint outcome. Thus, denoting birthweight by Y and gestational age by D, interest lies in the joint distribution

$$\Pr(Y, D|T, X) = \Pr(Y|D, T, X) \Pr(D|T, X) \tag{1.1}$$

where throughout this thesis T denotes some treatment or risk factor of interest and X denotes additional covariates of interest.

The factorization in (1.1) provides a model formulation for outcome variables and intermediate variables (hereafter denoted by Y and D, respectively) that is heavily featured throughout this thesis. As this suggests, gestational age may be viewed as an intermediate variable for birthweight. For example, babies may be born earlier because of smoking or nutrition, and early birth affects birthweight. Of course, smoking or nutrition may also affect birthweight through pathways other than gestational age. The directed acyclic graph (DAG) in Figure 1.1 displays a graphical representation of gestational age as an intermediate variable: The risk factor, smoking, may affect the birthweight outcome through gestational

age, or perhaps through some other pathway.

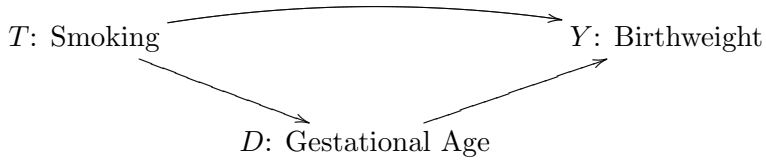

Figure 1.1: This DAG represents gestational age D as an intermediate variable: The risk factor, smoking T, may affect the birthweight outcome Y though gestational age, or perhaps via some other pathway.

Since the primary datasets used to study risk factors and potential treatments affecting birthweight and gestational age are observational in nature, linear or generalized linear models are typically used to adjust for X in the hopes of controlling confounding. A key factor influencing such analysis is that gestational age is typically censored and so is not precisely observed. For example, gestational age may be reported as the number of weeks since last menstrual cycle or a clinical estimate of the number of completed weeks of gestation. Thus, typical analyses fall into two categories:

1. Traditional regression analyses: Actual birthweight is modeled conditional on censored gestational age and other covariates, e.g., smoking, race, maternal age, infant sex, and maternal smoking status. In these analyses, conditioning on gestational age has been deemed favorable since (a) within each week of gestational age the observed distribution of birthweight is approximately normal, and (b) the predictive power of gestational age dwarfs that of the remaining typical predictors. Researchers generally do not attempt to use censored gestational age as the outcome variable in a regression model because of and resulting violations of the linear regression assumptions.

2. Traditional logistic regression analyses: LBW is modeled conditional on PTB and other covariates, and vice-versa. This approach allows for a simplified analysis and, in the case of PTB as the outcome, allows practitioners to consider gestational age as an outcome variable while avoiding more advanced modeling approaches such as multinomial regression or censored data analyses.

Unfortunately, both of these approaches entail three statistical modeling and analysis

mistakes.

1. Using gestational age or PTB as a covariate when considering birthweight does not treat birthweight and gestational age as a joint variable. The fact that birthweight and gestational age are so strongly correlated has no doubt contributed to analyses such as birthweight conditional on gestational age, but this relationship has been interpreted the wrong way. Arguably, the medically relevant treatment of birthweight and gestational age is as a joint variable.

2. Conditioning on intermediate variables results in post-treatment selection bias. This is further detailed beginning in Section 1.2. Posttreatment selection bias does not negatively affect predictive performance – predictive models for birthweight using gestational age perform better – but it results in a loss of causal interpretation for estimated treatment and outcome variable relationships. This detracts from understanding birthweight and gestational age disparities.

3. The choice to coarsen data, e.g., use LBW rather than actual birthweight, results in unnecessary information loss. With appropriate analyses approaches, high level summaries such as LBW may be recovered after the fact without sacrificing information in the model fitting.

Avoiding joint modeling, conditioning on intermediate variables, and sacrificing information through data coarsening are no doubt a result of the lack of availability and dissemination of alternative and preferable modeling techniques into the relevant scientific circles. It is important that alternative methods be developed and popularized so that such inefficient and insufficient analyses be avoided in the future. Chapter 2 details the development of such a model simultaneously addressing all three concerns. The key features of this model are

1. joint treatment of birthweight and gestational age, which avoids intermediate variables as a matter of course, and

2. treatment of gestational age as a continuous, but censored variable requiring imputation.

Gage (2003) and Ananth and Platt (2004) have previously examined and advocated the use of joint models for birthweight and gestational age. Joseph *et al.* (Joseph *et al.*, 2004b; Platt *et al.*, 2003; Joseph, 2007; Joseph *et al.*, 2004a) have considered models that utilize gestational age as a time axis for birthweight outcomes. The methodology presented here builds on the approach of Gage (2003) by using finite mixtures of bivariate regressions to flexibly model the nonstandard joint distribution of birthweight and gestational age conditional on covariates. The methodology improves on that of Gage (2003) by further clarifying the benefits available from joint modeling and addressing the issue of censored gestational age. When gestational age is reported as a censored variable, bivariate mixtures of regressions are not immediately appropriate. As continuous models, they ignore the uncertainty associated with the censorship of gestational age, which introduces bias into the modeling. Chapter 2 describes how this uncertainty may be appropriately incorporated into a mixture of bivariate regressions model. Under this methodology, the common censorship of gestational age is not prohibitive to joint modeling of birthweight and gestational age, and thus the benefits of doing so may be realized in censored data settings.

The methodology presented in Chapter 2 is not restricted to the joint birthweight and censored gestational age setting. It is generally instructive in the treatment, usage, and benefits of joint variable modeling, as well as the proper treatment of censored data. In addition, Chapter 2 also contributes the continued effort to raise awareness of the correct usage of intermediate variables.

1.2 Causal Inference and Principal Stratification: Adjusting for Intermediate Variables in Observational Settings

Unless a model is to be used for prediction alone, post-treatment (intermediate) variables should not be used as pretreatment (covariate) variables. As noted in Section 1.1, doing so

results in post-treatment selection bias and thus causes parameter estimates to lose causal interpretation. To give an illustrative example, suppose we design a randomized study to examine the efficacy of a nutritional supplement to reduce LBW outcomes. Now, suppose there is a disadvantaged group of individuals whose babies will always be LBW and PTB, regardless of whether they receive the nutrition supplement or not. Then, if we consider birthweight outcomes for individuals taking the nutrition supplement within the PTB strata (i.e., condition on an intermediate variable PTB), this will include all the disadvantaged members of LBW/PTB group, but will not include any individuals who are not PTB as a result of the nutrition supplement. On the other hand, no such post-treatment selection will have had the opportunity to occur in the non-treated arm since there was no nutrition supplement to potentially induce individuals to no longer be PTB. This means that, within the PTB strata comparison, the balance achieved by the original randomization may no longer be satisfied across treatments.

Continuing the example, If there is a group of birth outcomes that respond advantageously to the nutritional supplement – i.e. LBW/PTB without the nutritional supplement, but not LBW/PTB with the supplement – they will be selected out of the treatment arm (but not the control arm) within the PTB strata. This upweights the prevalence of LBW in the PTB strata of the treatment arm since non LBW individuals have been removed. Thus, within the PTB strata, the nutrition supplement group may appear to have a higher percentage of LBW than the non treatment arm for the very reason that the nutritional supplement has a positive effect on the outcome. Thus, conditioning on PTB can result in the opposite conclusion compared to not conditioning on PTB at all. If treatment is randomized, and there is an overall positive effect of a nutrition supplement on birth outcomes, it is indeed surprising to find the opposite conclusion upon conditioning on PTB.

This toy example illustrates the post-treatment selection bias resulting from conditioning on an intermediate variable. As will be discussed, this pitfall may be formally represented using conditional independence properties on DAGs (Pearl, 2000), conditional expectation (Rosenbaum, 1984), and potential outcomes (Frangakis and Rubin, 2002). Indeed, famil-

iarity with these concerns dates back to Pearson *et al.* (1898) and Yule (1903) who were the first to describe Simpson's (Blyth, 1972) and Berkson's (Berkson, 1946) paradoxes. The best known example of the paradoxes of intermediate variables is the venerable Berkeley admissions study (Bickel *et al.*, 1975) where evidence for, or against, gender bias in admissions into the University of California at Berkeley depended on whether analysis stratified on department (an intermediate variable), or not A recent incarnation of this issue has been the highly debated birthweight paradox where, within the PTB strata, smoking appears to reduce LBW outcomes (Hernández-Díaz *et al.*, 2006). One explanation for this puzzling result follows directly from the nutritional supplement example of the previous paragraph. Unfortunately, as with the birthweight paradox, it is not clear that the effects of intermediate variables are fully appreciated in many applied areas. The work in this thesis augments efforts to raise awareness of this issue and provide alternative methods of analysis in situations where intermediate variables can cause confusion by inducing post-treatment selection bias.

The vast majority of the work examining and advancing the understanding of intermediate variables is found in the causal inference literature. Causal inference is a relatively young and extremely active area. There are several prominent schools of thought, and substantial discussion (and bantering) between the available ideologies.

1.2.1 Causal Inference

Causal inference notions have already been tentatively introduced in the form of Figure 1.1. Much of causal inference is concerned with understanding causal pathways and structures using graphical representations, such as DAGs. Pearl (2001, 2000), and several others (Robins, 2003; Robins and Greenland, 1992; Dawid and Didelez, 2005; Didelez *et al.*, 2006; Didelez and Sheehan, 2005) have used graphical representations to initiate discussions of causality, experiments, interventions, direct effects, indirect effects, and so on. Indeed, there is apparently much room for debate as to how and what exactly graphical representations may and should be used for. In this regard, Dawid *et al.* (Dawid, 2000; Dawid and Didelez,

2005; Didelez *et al.*, 2006) has played a moderating role, carefully cautioning against overly simplistic and enthusiastic interpretations of graphical causal structure representations.

One success of graphical approaches is the identification of the so called back door criterion violation (Pearl, 2000, 1995; Greenland *et al.*, 1999), which explains and diagnoses the post-treatment selection bias using conditional independence properties of DAGs (Pearl, 2000; Dawid, 1980), such as those in Figure 1.2. As Figure 1.2d demonstrates, within the strata of the intermediate variable, a (spurious) correlation structure exists between ancestors of the intermediate variable (i.e., those effecting the intermediate variable), even if the ancestors are *a priori* marginally independent. An intuitive example of this is a sidewalk that becomes wet from two independent sources, rain, and broken water pipes. If the sidewalk is wet, and it didn't rain, then the water pipes have broken. That is, by virtue of conditioning on the downstream (intermediate) variable (the sidewalk is wet), the ancestor variables suddenly contain information about each other and are associated.

The induced spurious relationships may result in a loss of causal interpretation. For example, consider Figure 1.2d in the setting of Section 1.2.1: Let T be the nutritional supplement, D be PTB, X to be an unobserved variable resulting in the disadvantaged group of always PTB/LBW, and Y to be LBW. Conditioning on the intermediate variable PTB opens a spurious relationship between nutrition and LBW through X. The relationship between nutrition and LBW is now mixed up with the spurious relationship between the two through X induced by conditioning on PTB, and so it cannot be treated in a causal manner.

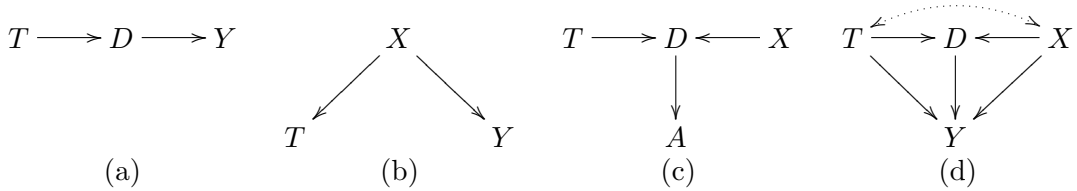

Figure 1.2: DAGs (a) through (c) are representations of the three key conditional independence rules: (a) Intermediate Variable Case: $T \perp\!\!\!\perp Y|D$; (b) Confounder Case: $T \perp\!\!\!\perp Y|X$; (c) Ancestor Case: $T \perp\!\!\!\perp X$ but $T \not\perp\!\!\!\perp X|D$ or A. DAG (d) demonstrates the results of the backdoor criterion violation of conditioning on the intermediate variable, D. Namely, spurious correlation between the treatment T and the outcome Y is induced via X as a result of conditioning on the intermediate variable D.

Discovery of causal pathways and structures between a set of observed variables has been attempted with some success by the TEDTRAD project out of Carnegie Mellon University (See, `http://www.phil.cmu.edu/projects/tetrad/publications.html`). The major difficulties with such methodologies is that while correlation structure is easily obtained, causal structure is more difficult to elucidate. For example, in Figure 1.2, observational data cannot distinguish between graphs (a) and (b). In many situations, it appears that expert generated guidance is required to augment the information in the data to allow for causal structure discovery.

Extensive, but known, causal structures have been studied in the multi-treatment setting by Robins et al. (Robins and Greenland, 1992; Robins *et al.*, 1999, 2000; Robins, 2001; Hernán *et al.*, 2000). Figure 1.3 provides a simple illustration of this setting: There are a series of consecutive treatments T and covariates D that vary over time and influence some final outcome Y. Individual treatments and covariates may vary both independently or dependently of previous treatments and covariates, effectively resulting in a temporal sequence of several intermediate variables. Because of the possible dependencies, intermediate variables must be adjusted for in these contexts.

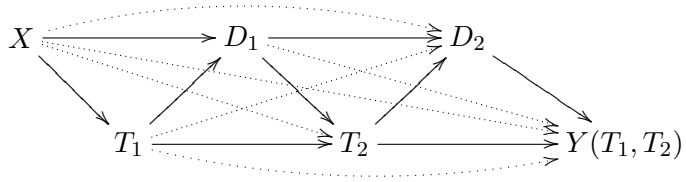

Figure 1.3: This figure demonstrates the multi-treatment causal inference setting. There are multiple treatments T_m over time that may affect a final outcome of interest Y as well as future variables D_m, but may also be affected by previous D_m and Y_m.

Appropriate adjustment for multiple intermediate variables may be accomplished with the g-computation formula, which is simply a set of distributional properties following from ignorability-like assumptions. For instance, for the case of two treatments in Figure 1.3,

the g-computation formula is

$$\begin{aligned} E[Y(T_1,T_2)] &= \int E[Y(T_1,T_2)|T_1,X]dF(X) \\ &= \int E[Y(T_1,T_2)|D_1,T_1,X]dF(D_1|X)dF(X) \\ &= \int E[Y(T_1,T_2)|D_2,T_2,D_1,T_1,X] \\ &\quad dF(D_2|D_1,X)dF(D_1|X)dF(X), \end{aligned}$$

where the equalities follow for $T_1 \perp\!\!\!\perp Y(T_1,T_2)|X$ and $T_2 \perp\!\!\!\perp Y(T_1,T_2)|D_1,T_1,X$, respectively. In general, for M treatments, the g-computation formula is

$$\begin{aligned} E[Y(T_1,\cdots T_M)] &= \int E[Y(T_1,\cdots T_M)|T_1,\cdots T_M,D_1,\cdots D_M,X] \\ &\times \quad dF(X)\prod_{m=1}^{M} dF(D_M|T_1,\cdots T_m,D_1,\cdots D_{m-1},X). \end{aligned}$$

Unfortunately, it is generally not feasible to simply substitute parametric forms for the models into the g-computation formula because they often cannot functionally provide the necessary conditional independence properties (Wasserman, 1999). Specialized models specifically designed to satisfy the necessary assumptions have thus been developed, including the classes of Structural Nested Models (Robins and Greenland, 1992) and Marginal Structural Models (Hernán *et al.*, 2000). Marginal Structural Models are increasingly popular for categorical variable settings where they can be fit using the so called inverse-probability-of-treatment weighting, which re-expresses the g-computation formula as

$$\begin{aligned} E[Y(T_1,\cdots T_M)] &= \sum E[Y(T_1,\cdots T_M)|T_1,\cdots T_M,D_1,\cdots D_M,X] \\ &\times \quad \frac{\Pr(X)\prod_{m=1}^{M}\Pr(T_m|T_1,\cdots T_{m-1},X)}{\prod_{m=1}^{M}\Pr(T_m|T_1,\cdots T_{m-1},D_1,\cdots D_{m-1},X)}. \end{aligned}$$

It is interesting to note the similarity between the inverse-probability-of-treatment weighting scheme and the propensity score, which has long had a prominent role in causal inference. Recall that a propensity score $\mathrm{e}(X) = \Pr(T=1|X)$ has the property that if $Y \perp\!\!\!\perp T|X$ then

$Y \perp\!\!\!\perp T|\mathrm{e}(X)$ (Rosenbaum and Rubin, 1983b). Thus, matching or sub-classification based on the propensity score has long been viewed as a viable alternative to covariance adjustment and other forms of matching and sub-classification aiming to adjust for imbalance in covariates or reduce variability in parameter estimation. Inverse-probability-of-treatment weighting is yet another use of the propensity score in the context of treatment effect estimation.

As of yet, this discussion has managed to avoid the fundamental building block of all causal inference analyses, namely, the potential outcome. Potential outcomes are a series of 'what if?' questions. Returning to the nutrition setting for an example, imagine that mothers either receive, $T = 1$, or do not receive, $T = 0$, a nutrition supplement that may influence the birthweight Y of their child. For each mother i, there are two a priori existing hypothetical outcomes, $Y_i(T = 1)$ and $Y_i(T = 0)$, which represent what the birthweight would have been had mother i received, and not received the supplement, respectively. Similarly, for each mother i gestational age too has potential outcomes $D_i(1)$ and $D_i(0)$ if it is realized after the treatment.

The potential outcomes framework dates back to Fisher (1918, 1925) and Neyman (1923), but has been popularized as the so called Rubin Causal Model (RCM, Holland, 1986), advocated by Rubin (1978). Potential outcomes are simple because the causal effect of the treatment for say, birthweight, is based on the appealing and intuitive comparison between $Y_i(1)$ and $Y_i(0)$. Potential outcomes are difficult because outcomes $Y_i(1 - T_i)$ and intermediate variables $D_i(1 - T_i)$ are not observed. That is, $Y_i^{obs} = Y_i(1)$ for individuals i such that $T_i = 1$, $Y_i^{obs} = Y_i(0)$ for individuals i such that $T_i = 1$. Nonetheless, conceptualization of potential outcomes provides basis for many causal inference analyses. Often, causal estamands may be identified despite the unobserved potential outcomes (e.g., see Angrist *et al.*, 1996). Other times, causal inference analyses require imputation of missing potential outcomes (e.g., see Hill and McCulloch, 2007; Hirano *et al.*, 2000). The primary benefit of potential outcomes is to provide clear delineation of the necessary assumptions required to justify causal statements in a particular setting. Potential outcomes thus provide

a straightforward and open groundwork where the merits of causal claims may be judged. Because potential outcome considerations focus attention on the underlying assumptions and structures of causal inference, they often leads to constructive formulations of causality. For example, post-treatment selection bias may be explained in two ways using potential outcomes. Rosenbaum (1984) demonstrates that conditioning on an intermediate variable results in a loss of causal estimation, since

$$\begin{aligned} E[Y(1)-Y(0)] &= E[Y(1)-Y(0)] \\ &\overset{Y \perp\!\!\!\perp T}{=} E[Y(1)|T=1]] - E[Y(0)|T=0] \\ &= E_{D(1)}[E[Y(1)|T=1, D(T)]] \\ &\quad -E_{D(0)}[E[Y(0)|T=0, D(T)]] \\ &\neq E_{D^{obs}}[E[Y(1)|T=1, D(T)]] \\ &\quad -E_{D^{obs}}[E[Y(0)|T=0, D(T)]], \end{aligned}$$

where in the last line expectation is taken with respect to the observed distribution D^{obs}, which is the mixture distribution consisting of $D(1)$ and $D(0)$ and not distinguishing between the two. Frangakis and Rubin (2002) explain post-treatment selection bias by noting that direct comparison between

$$\{Y^{obs} : D^{obs} = D_0, T_i = 1\} \qquad \text{and} \qquad \{Y^{obs} : D^{obs} = D_0, T_i = 0\}$$

or, equivalently

$$\{Y_i(1) : D_i(1) = D_0\} \qquad \text{and} \qquad \{Y_i(0) : D_i(0) = D_0\},$$

does not compare exchangeable individuals and so cannot be causal since – even with randomization of T_i – the sets

$$\{i : D_i(1) = D_0, T_i = 1\} \quad \text{and} \quad \{i : D_i(0) = D_0, T_i = 0\}$$

are not the same if D is indeed an intermediate variable affected by the treatment T_i That

is, if $D_i(0) \neq D_i(1)$, i cannot be in both sets simultaneously.

1.2.2 Principal Stratification Sensitivity Analysis

The previous demonstration of post-treatment selection bias from Frangakis and Rubin (2002) also provides the motivation for their proposal of PS. PS returns the focus back to comparable sets of individuals by, instead of conditioning on the intermediate variable itself, conditioning on the so called principal strata, as with the comparison between

$$\{Y_i(1) : (D_i(0), D_i(1)) = (D_0, D_1)\} \quad \text{and} \quad \{Y_i(0) : (D_i(0), D_i(1)) = (D_0, D_1)\}.$$

The principal stratum $(D_i(0), D_i(1))$ is the joint potential outcome of the intermediate variable, and it is invariant under treatment assignment since it exists prior to treatment. Thus conditioning on the principal strata of the intermediate variable is just like conditioning on a pretreatment variable. Specifically, conditioning on the principal strata does not induce post-treatment selection bias. PS is distinct from the g-computation formula previously discussed, and results in different causal estimands than the g-computation formula. Nonetheless, it has been used with much success in many settings. This is partly because PS may be used to disentangle and richly interpret causal effects, as will be emphasized in Chapter 4.

The formal presentation of PS (Frangakis and Rubin, 2002) was somewhat late in arriving, as many of the practical benefits of PS had already been harnessed in the causal inference literature (e.g., see Imbens and Angrist, 1996; Angrist *et al.*, 1996; Hirano *et al.*, 2000). In fact, for binary intermediate variables and treatments, PS simply recovers the instrumental variables estimator (Angrist *et al.*, 1996; Pearl, 2000) dating back to the structural equations models of Wright and Haavelmo (Wright, 1928, 1934; Haavelmo, 1943, 1944; Goldberger, 1972; Morgan, 1990). A nice example of the use of instrumental variables is given in McClellan *et al.* (1994). It is worth noting here that the econometrics literature, of which instrumental variables is a part, has also produced its own take on causal inference. For further discussion of this branch of causal inference, the interested reader may begin

with Heckman (2008).

Instrumental variables and PS methodology are useful because they provides a way to uncover causal relationship between a variable D and an outcome Y, even when D and Y are possibly confounded. To do so, however, instrumental variables requires an instrument T that is not confounded with D and may only be correlated with Y through its effect on D. PS, on the other hand, requires essentially the same thing but describes its assumptions in terms of ignorability and monotonicity (see Chapter 3). Angrist *et al.* (1996) argues that the assumptions implied by the PS derivation are different and perhaps more interpretable than those resulting from a structural equations models derivation. Nonetheless, Instrumental variables and PS methodology are generally both restricted to natural randomization or partially controlled experiment settings as a result of their necessary assumptions. However, numerous questions of significant importance are not available for study under these idealized settings, and so many important questions remain inaccessible with currently available analyses.

Chapter 3 of this thesis provides a methodology to perform PS analyses for binary intermediate variables and treatments under relaxed ignorability assumptions. This extends PS into the completely observation setting without partial randomization or natural experiment assumptions, and thus allows for the undertaking of analyses previously unavailable due to lack of experimental control or appropriate instrument availability. The extension uses a model based sensitivity analysis approach that complements current sensitivity analysis approaches to unmeasured confounders.

Sensitivity analyses define parameters denoting the amount of unobserved confounding that may be present in a study, such as the maximum odds ratio of treatment assignment

$$\frac{1}{\Gamma} \leq \frac{\Pr(T_i = 1)\Pr(T_j = 0)}{\Pr(T_i = 0)\Pr(T_j = 1)} \leq \Gamma,$$

for any i and j such that $X_i = X_j$ (Rosenbaum, 2002). By computing the null distribution of a test statistic of interest under the most extreme settings for Γ, the sensitivity of results to unmeasured confounding may be determined. Small and Rosenbaum (2008) have used this

permutation testing approach to begin considering the extension of instrumental variables to observational settings. They have not addressed all forms of confounding, however.

Another approach to sensitivity analysis was presented by Rosenbaum and Rubin (1983a), who directly modeled the unmeasured confounding, as in

$$\Pr(Y_i(T_i), T_i, U_i, X_i) \quad = \quad \Pr(Y_i(T_i)|T_i, U_i, X_i)\Pr(T_i|U_i, X_i)\Pr(U_i|X_i)\Pr(X_i).$$

For binary treatments, outcomes, and confounders, for example,

$$\begin{aligned}
\Pr(X_i = s) &= \phi_s, \left(\sum_s \phi_s = 1\right) \\
\Pr(U_i = 0|X_i = s) &= \pi_s \\
\Pr(T_i = 0|U_i, X_i) &= (1 + \exp(\gamma_s + U_i\alpha_s))^{-1} \\
\Pr(Y(T_i) = 0|U_i, X_i) &= (1 + \exp(\beta_{st} + U_i\delta_{st}))^{-1}.
\end{aligned}$$

Under model based sensitivity analysis, all plausible confounding scenarios are translated into the sensitivity parameters, here, π_s, α_s, and δ_s, and the resulting model fits show the sensitivity of results to unmeasured confounding. Sjölander *et al.* (2008) provide a model based sensitivity approach in the PS context, but their methodology again does not consider all potential forms of confounding. Chapter 3 of this thesis extends their proposal to appropriately deal with any form of confounding present in PS analysis.

1.3 Flexible Principal Stratification for Continuous Intermediate Variables: Bayesian Semi-parametric Modeling

Chapter 2 of this thesis uses finite mixture modeling approaches to capture the distributional shape of joint birthweight and gestational age. This is because the distribution is very nonstandard and there are no adequate parametric models. This situation is not atypical. For many data settings, parametric models are simply to stringent too capture unique distributional shapes. For a full review of finite mixture models, see McLachlan and Peel

(2000) and Dey and Rao (2005). An abbreviated introduction is given here.

In many cases, Y_i cannot be adequately represented by a single distribution K_{θ_1}. But often it may be reasonably represented by weighting and combining distributions, as in

$$\Pr(Y_i) = \sum_{h=1}^{H} w_h K_{\theta_h}(Y_i), \qquad \sum_{h=1}^{H} w_h = 1, w_h > 0 \quad \text{for } h \in \{1, \cdots H\}.$$

Indeed, such modeling approaches give extreme flexibility in creating unique distributional forms. In their most extreme form with $H = N$ (the number of data points), finite mixture models produce nonparametric kernel density estimation by setting $w_h = 1/N$, and $K_{\theta_h}(Y^*) = K_{\theta_h}((Y^* - Y_h)/r)/r$, where r is a bandwidth parameter. On the other end of the spectrum with $H = 1$ mixture models collapse back to standard parametric models. In this sense, finite mixture models can be viewed as a methodology that lies somewhere between parametric and nonparametric approaches.

Because the likelihood becomes intractable as the number of samples N grows, mixture models are implemented using a data augmentation approach that conceptualizes each mixture component, K_{θ_h}, as a subpopulation that contributes $w_h \times 100\%$ of the overall population. For each observation Y_i, a latent (unobserved) multinomial variable $Z_i \in \{1, \cdots H\}$ is introduced into the model indicating the subpopulation to which Y_i belongs. An augmented model is then specified as

$$f_\theta(Y_i) = \sum_{h=1}^{H} K_{\theta_h}(Y_i) 1_{[Z_i=h]}, \qquad Z_i \sim \text{MN}(w_1, \cdots w_h),$$

which reverts back to the original finite mixture model when the latent Z_i indicator variables are integrated out.

One stumbling block in the implementation of finite mixture modeling approaches is the so called label switching problem, which refers to the fact the labels $h \in \{1, \cdots, H\}$ of the mixture models are not identified. That is, the labels may be reordered and yet still produce the same likelihood. This issue is particularly noticeable in Bayesian analyses since posterior sampling chains for component parameters may frequently change places as the

parameters switch their roles in the mixture model. Pragmatic solutions involve putting order constraints on the component parameters, or performing post processing after model fitting. The second option is generally preferred (Jasra *et al.*, 2005).

Another problem is choosing the number of components H. When the components of a mixture model actually represent true underlying subpopulations in the overall population, and if the number of subpopulations is known, there is no problem choosing H. When the number of subpopulations is not known, or there is no true subpopulation interpretation available for a given data set, the selection of H becomes unclear. There are several pragmatic solutions available, but the problem remains an open debate (McLachlan and Peel, 2000).

An alternative to constructing the necessary distribution via mixture model approach is to specify a nonparametric model for the data. The Bayesian conceptualization is very natural and straightforward. Let Y_i follow some undefined distribution G, and define a prior distribution over G with support on the space of all possible distributions, as in

$$Y_i \sim G, \qquad G \sim \Pr(G).$$

The posterior distribution $\Pr(G|Y)$ represents what has been learned about G given the information in the data. Since the distribution G has support over the space of all possible distributions, such an approach is nonparametric in the sense that the model G is determined by the data Y, free of any *a priori* model family assumptions.

There are many nonparametric and semi-parametric procedures in the statistical literature including numerous distribution free tests, flexible functional relationship specifications such as basis expansions, trees, and wavelets, and nonparametric distribution estimation methodologies (Wasserman, 2005). The approach above – using a prior to induce a nonparametric distribution for data – is a nonparametric Bayes approach for nonparametric distribution estimation. Nonparametric Bayes is an extensive and active research area with many implementation approaches and variants for density estimation, such as those based on DPs (Ferguson, 1973; Escobar and West, 1995a; Ishwaran and James, 2001), Polya urn

schemes (Lavine, 1992b,a), and numerious others involving, for example, Poisson processes, Beta processes, and Lévy processes. Many of the recent advances in nonparametric Bayes involve the development of dependent DPs (Müller *et al.*, 2004), such as the hierarchical DP (Teh *et al.*, 2003), the nested DP (Rodrguez *et al.*, 2008), the local DP (Chung and DB, 2009) and similarly behaving ordered DP (Griffin and Steel, 2004). This thesis however, focuses on the independent DP (Ferguson, 1973), which is now described.

A set of random probability measures G_j defined on a space A are sampled from a DP with strength parameter α and base measure G_0 (also defined on A) – denoted by $G_j \sim DP(\alpha G_0)$ – if for any partition of the space $A, \{A_1, \cdots A_k\}$, the distribution of the set of probabilities defined by the sampled random measures G_j on the partition follows the Dirichlet distribution with parameters determined by α times the probabilities defined on the partition by G_0. That is,

$$\begin{aligned} G_j &\sim DP(\alpha G_0) \\ &\text{if} \\ (\Pr_{G_1}(A_1), \cdots \Pr_{G_k}(A_k)) &\sim \text{Dirichlet}\left(\alpha \Pr_{G_0}(A_1), \cdots \alpha \Pr_{G_0}(A_k)\right). \end{aligned}$$

The DP is a distribution on distributions: Large values for the scalar strength parameter α imply less variation of a realized distribution G_j from the base measure G_0, where $E(G_j) = G_0$ in the sense that $E(G_j(A_1)) = G_0(A_1)$ for any $A_1 \subset A$. Despite being a distribution on distributions, the DP cannot be universally used to implement the nonparametric Bayes approach above because distributions G_j sampled from DPs are true distributions in the formal sense. This is seen from the stick breaking construction of the DP (Sethuraman,

1994), which shows that setting G_j equal to the point mass distribution

$$\begin{aligned} G_j(\cdot) &= \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot), \quad \text{where} \\ \theta_h &\overset{iid}{\sim} G_0, \\ w_h &= w'_h \prod_{k<h}(1-w'_k), \quad \text{and} \\ w'_h &\overset{iid}{\sim} \text{Beta}(1,\alpha) \end{aligned}$$

is equivalent to sampling $G_j \sim DP(\alpha G_0)$. In the stick breaking specification of the DP, θ_h are called atoms and w_h are probabilities that sum to 1. In addition to highlighting the discrete (point mass) nature of DP realizations, this representation shows that the DP encourages decreasing weights $\pi_i > \pi_j$ for $i > j$ since $E[w_h] = 1/(1+\alpha)$. Small α corresponds to sparser models, i.e., models having fewer nontrivial weights and hence, a coarser approximation to G_0.

Even though the DP cannot be directly used as a prior distribution for a nonparametric distribution G for continuous data Y_i, it may be used indirectly through a scale mixture approach. A continuous kernel K mixed across a distribution G sampled from a Dirichlet process will result in a continuous distribution

$$\begin{aligned} \Pr(Y_i) &= \int K_\theta(Y_i) dG(\theta) \\ &= \sum_{h=1}^{\infty} w_h K_{\theta_h}(Y_i) \end{aligned}$$

that takes on an infinite mixture model specification and can be used as a model for continuous data Y_i. This approach is called the DPM.

In the DPM, the tendency for decreasing weights w_h implied by the stick breaking representation is further facilitated by the natural Bayesian Occam's razor that follows

since the conditional marginal likelihood

$$\int \sum_{h=1}^{\infty} w_h K_{\theta_h}(Y_i) dG_0(\theta_1), \cdots dG_0(\theta_H), \cdots$$

is generally larger when many $w_h \approx 0$. This sparsity property of the DPM effectively provides an automatic selection mechanism for the number of active components $H < \infty$ in the MDP, i.e., the number of nontrivial w_h. Thus, when the data size is fixed, and only a small number of w_h are nonzero, the nonparametric behavior of the DPM may be approximated with a finite mixture model that truncates the infinite mixture at some large H (Ishwaran and James, 2001). As more data arrives, however, the number of w_h that significantly contribute to the mixture model will increase, and the truncation will result in a loss of the nonparametric behavior. That is, the model will no longer be able to nonparametrically respond to data and will instead function exactly like a finite mixture model. The data examined in this thesis has fixed sample size, and thus the truncation

$$\begin{aligned} G(\cdot) &= \sum_{h=1}^{H} w_h \delta_{\theta_h}(\cdot), \\ \theta_h &\overset{iid}{\sim} G_0, \\ w_h &= w_h \prod_{k<h} (1 - w_k), \\ w_h &\overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } h < H, w_H \equiv 1. \end{aligned}$$

with very large H is used to approximate the DPM. This approximation may be avoided if desired (Escobar and West, 1995a).

Returning to causal inference, recall that the PS approach described in the previous section requires conditioning on the principal strata $(D(1), D(0))$ rather than on the observed intermediate variable D^{obs}. Since the principal strata are not jointly observed, the unobserved marginal distribution of the principal strata must be imputed. This may be

facilitated by modeling

$$\begin{aligned} & \Pr(Y_i(1), Y_i(0), D_i(1), D_i(0)|T_i, X_i) \\ = \; & \Pr(Y_i(1), Y_i(0)|D_i(1), D_i(0), T_i, X_i) \Pr(D_i(1), D_i(0)|T_i, X_i), \end{aligned}$$

and using the information contained in (Y_i^{obs}, T_i, X_i) to inform the unobserved principal strata $D_i(T_i - 1)$ through the principal strata and outcome models.

When the intermediate variable D_i^{obs} is categorical, the principal stratum are also categorical (Imbens and Angrist, 1996; Angrist *et al.*, 1996) and may be modeled (conditional on the treatment and covariates) using a multinomial regression (Hirano *et al.*, 2000), as will be discussed in detail in Chapter 3. When the intermediate variable D is continuous, the principal stratum is a continuous two-dimensional variable. PS analysis in the context continuous intermediate variables affords many new interpretation opportunities over the binary context, but also many new challenges. Continuous principal strata models were proposed by Jin and Rubin (2008), but to date, all available modeling techniques are fully parametric, and little flexibility in modeling. When the models cannot adequately capture the true distributional shape of the principal strata, the the necessary imputation of the principal strata will suffer.

To address this concerns, Chapter 4 proposes semi-parametric method that uses a nonparametric DPM model for the principal strata, and a fully parametric model for the outcome. The DPM approach provides completely flexible nonparametric model and capable of handling nonstandard principal strata distributions. An additional potential benefit of the approach is the clustering properties that naturally follow from DPMs. In the data analyzed in this thesis, clustering via the DPM provides meaningful interpretations of results. This use of the DPM is one of the first examples nonparametric modeling methods being applied in causal inference settings. This thesis thus provides a path between two significant and relevant areas of research, and opens the doors for further improvement in causal inference analyses via more flexible nonparametric Bayes approaches.

Chapter 2

Inferential Benefits of Joint Modeling of Birthweight and Gestational Age

2.1 Introduction

Birthweight and gestational age are closely related and represent important indicators of a the health of a newborn. Customary modeling for birthweight is conditional on gestational age. However, joint modeling directly addresses the relationship between gestational age and birthweight, provides increased flexibility and interpretation (Gage, 2003; Ananth and Platt, 2004), and a strategy to avoid using gestational age as an intermediate variable. Chapter 2 clarifies the advantages of bivariate analyses over birthweight conditional on gestational age analyses, and illuminates the inferential (prognostic) uses and benefits of joint modeling (similarly argued Tassone *et al.* (2010)). In addition, Chapter 2 extends the currently available finite mixtures of bivariate regression models of Gage (2003) through a latent specification that accounts for interval censored gestational age. The methodology is described and implemented in a Bayesian framework that enables inference beyond customary parameter estimation as well as exact assessment of uncertainty. The model is

demonstrated using the North Carolina Detailed Birth Record (NCDBR) database available through the Children's Environmental Health Initiative.

Section 2.1.1 briefly reviews the relevance and progress of birthweight and gestational age analyses, and positions the work of this thesis in this literature. Section 2.1.2 describes the NCDBR. Section 2.2 introduces the finite mixture of bivariate regressions model specification for the joint variable birthweight and gestational age (Gage, 2003; Fang *et al.*, 2007; Gage *et al.*, 2008a) and extends the model to allow for the common interval censored form of gestational age. In section 2.3, the negative consequences of analyses using intermediate variables, such as birthweight conditional on gestational age analyses, are highlighted in contrast to joint analyses. Section 2.4 addresses model identifiability concerns. Finally, in Section 2.5 the inferential benefits of the bivariate model are demonstrated, e.g., in examination of disparities within the general population and recovery of conditional results.

2.1.1 Modeling Approaches for Birthweight and Gestational Age

LBW and PTB have long been associated with many adverse birth and developmental outcomes (e.g. Ylppö, 1919; Karn and Penrose, 1951). However, the joint role of birthweight and gestational age, while recognized, is not well understood. Often, Small for Gestational Age (SGA, less than the 10% birthweight quantile for a given gestational age) is used as a proxy for the joint information of birthweight and gestational age. While LBW, PTB, and SGA are used prospectively as indicators of potential birth complications, their physiological importance is not so clear cut. As Grimes (1998) relates, these classifications achieve relevant sensitivity to adverse birth outcomes (Type I error) at the cost of specificity (Type II error), often not corresponding to medical signs of abnormalities. For instance, Wilcox (2001) notes that interventions aimed at reducing LBW have not yet met with success despite widespread interpretation of LBW as a cause of adverse birth outcomes (e.g., Paneth, 1995).

Work seeking to understand variables like LBW, PTB, and SGA has thus far proven very productive, though has perhaps not yet made its way into common practice. For

instance, Wilcox and Russell (1990) have brought attention to the varied relevance of LBW by sub-population, partially as a byproduct of arbitrary specification of the LBW cutoff. And Platt *et al.* (2003) and Hernández-Díaz *et al.* (2006) have provided constructive advice concerning the once puzzling birthweight paradox. Platt *et al.* (2003) (and see also Joseph *et al.*, 2004b,a; Joseph, 2007) clarifies the difference between treating gestational age as a time axis verses a covariate that does not capitalize on the temporal nature of gestational age. Namely, covariate strategies imply comparisons within gestational age week strata and are prognostic in nature, whereas time axis strategies compare among the at risk population and are more traditionally causal in nature. The emphasis here, however, relates more to Hernández-Díaz *et al.* (2006) since it shows that use of intermediate variables may introduce bias and thus provides an impetus for joint modeling.

One area of research frequently pursued is the exploration of LBW and PTB as adverse birth outcomes themselves, and some effort has been spent carefully modeling in these contexts (Wilcox, 2001; Wilcox and Skjaerven, 1992; Oja *et al.*, 1991; Gage and Therriault, 1998; Gage, 2000, 2002; Gage *et al.*, 2004). The proposal to study birthweight and gestational age as a joint variable soon followed as the natural course of this tradition Gage (2003); Ananth and Platt (2004). Models for the joint birthweight and gestational age variable have subsequently been incorporated as sub-models in analyses of further adverse birth outcomes, e.g., fetal death), as in Fang *et al.* (2007) and Gage *et al.* (2008a). These models introduce a logistic regression conditional birthweight and gestational age in order to model a tri-variate outcome. Since gestational age is again used as a covariate rather than a time axis, these models are prognostic in nature as indicated by the discussion from Platt *et al.* (2003).

This work pursues the original proposal to study birthweight and gestational age jointly and re-emphasizes that they are intimately related and thus natural candidates for a joint outcome. Further, jointly modeling birthweight and gestational age provides a means to bypass the potential difficulties associated with conditional modeling while at the same time facilitating understanding and interpretation of these important indicators of pregnancy

health.

2.1.2 Data Application: NCDBR

Through a negotiated data sharing agreement with the North Carolina state center for health statistics, the Children's Environmental Health Initiative (CEHI) at Duke University has access to the NCDBR. These data include birth certificate information for all NC births from 1990-2007 (N = 1,862,405 births). The current study is limited to birth records from 2004-2006, (N = 371,924). The data set is further restricted to women who self-declare as non-Hispanic white (NHW), non-Hispanic black (NHB), and Hispanic (HS) mothers, aged 15-44, who report no alcohol use during pregnancy. Only singleton births with no congenital anomalies, birthweight greater than 399 grams, and gestational age 24 to 42 weeks are considered. Finally, results are based on complete case analysis using the variables birthweight, gestational age, reported smoking, infant sex, reported marital status, maternal race, maternal age (15-19, 20-24, 30-34, 35-39, 40-44, and the referent 25-29), maternal education (middle-school or less, some high school, some college, at least college, and the referent high school), and first birth infant. Thus, the final data set has N = 336,129 observations. The population characteristics of this data set are given in Table 2.6 in the final column labeled 'Overall'. This research was conducted according to a human subjects research protocol approved by the University's institutional review board.

Birthweight is reported in pounds and ounces and converted to grams for analysis. Gestational age is reported as a clinical estimate of the number of weeks gestation completed. Gestational age is thus a censored integer valued response. Figure 2.1 displays histograms of birthweight for each gestational age from 24 to 42. Figure 2.2 displays the same data in bivariate form. Both figures reveal the strong dependence between the birthweight and gestational age with the latter revealing that a simple bivariate Gaussian specification may not suffice.

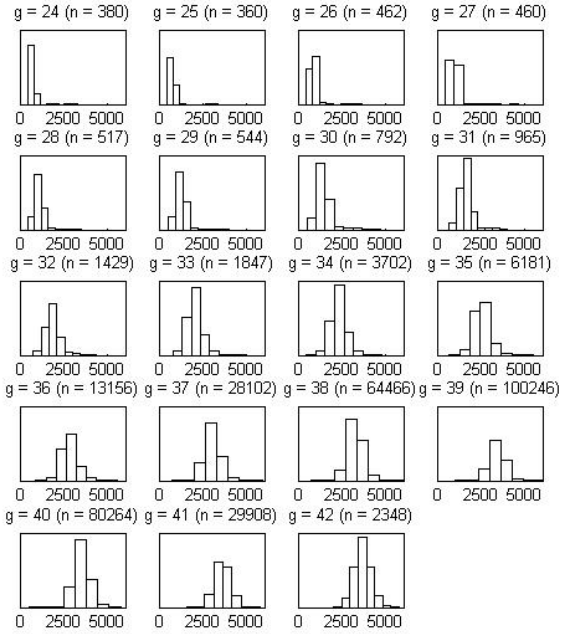

Figure 2.1: Histograms of birthweight by gestational age (g, 24 to 42) for the data subset described in 2.1.2.

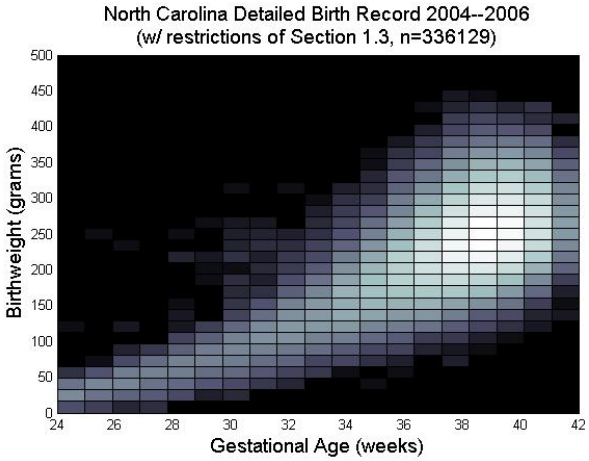

Figure 2.2: A log-scale heatmap version of a bivariate histogram of birthweight and gestational ages for the data subset described in 2.1.2.

2.2 Joint Birthweight and Gestational Age Model

2.2.1 Likelihood Specification

The unique shape of the joint birthweight and gestational age distribution shown in Figure 2.2 can be flexibly modeled using finite-mixture models (McLachlan and Peel, 2000; Dey

and Rao, 2005) as discussed in Gage (2003), Fang *et al.* (2007), and Gage *et al.* (2008a). For each individual i, denote birthweight by b_i, uncensored gestational age by g_i, and covariates by x_i. A marginal times conditional form for the H-component mixture-model,

$$\begin{aligned}(b_i, g_i)' \sim \sum_{h=1}^{H} w_h \mathrm{N}(g_i|\mu_{g,h} + x_i'\beta_{g,h}, \sigma^2_{g,h}) \times \\ \mathrm{N}(b_i|\mu_{b,h} + x_i'\beta_{b,h} + (g_i - (\mu_{g,h} + x_i'\beta_{g,h}))\beta_{*h}, \sigma^2_{b|g,h}),\end{aligned} \tag{2.1}$$

provides a natural birthweight conditional on gestational age interpretation. The mixing weights, w_h, sum up to 1. As shown in Section 2.5.1, there is sufficient justification to allow coefficient parameters to differ by component.

The centering (see Gelfand and Sahu, 1999) of g_i in (2.1) results in the equivalent bivariate regression mixture model specification

$$(g_i, b_i)' \sim \sum_{h=1}^{H} w_h \mathrm{N}\left(M_h, \Sigma_h\right),$$

with

$$M_h = \begin{bmatrix} \mu_{b,h} + x_i'\beta_{b,h} \\ \mu_{g,h} + x_i'\beta_{g,h} \end{bmatrix}, \Sigma_h = \begin{bmatrix} \dfrac{\sigma^2_{b|g,h}}{1-\rho_k^2} & \rho_k \dfrac{\sigma_{b|g,h}}{\sqrt{1-\rho_h^2}}\sigma_{g,h} \\ \rho_h \dfrac{\sigma_{b|g,h}}{\sqrt{1-\rho_h^2}}\sigma_{g,h} & \sigma^2_{g,h} \end{bmatrix},$$

where

$$\rho_h = \beta_{*h}\sqrt{\left(\beta^2_{*h} + \frac{\sigma^2_{b|g,h}}{\sigma^2_{g,h}}\right)^{-1}}.$$

Model (2.1) provides the framework to treat birthweight and gestational age as a continuous joint variable. The bivariate regression structure incorporates covariates x_i' into the component means, though not in the mixing proportions as proposed in the univariate case in Gage *et al.* (2008b). The mixture portion of the model provides a flexible structure to model the resulting residuals for b_i and g_i given x_i', i.e. $(b_i, g_i)' \sim \sum_{h=1}^{H} w_h M_h + \sum_{h=1}^{H} w_h \mathrm{N}\left(0, \Sigma_h\right)$.

The mixture structure for the residuals provides aggregated bivariate structure for birthweight and gestational age. Local-scale structure within each component is modeled by ρ_h, which depends on β_{*h}, $\sigma_{b|g,h}$, and $\sigma_{g,h}$. Both covariate coefficients and resulting birthweight and gestational age residuals are component dependent due to the component-varying parameters. The covariance structure Σ_h also varies by component. Finally, conditional models may be recovered from the joint specification; e.g., the conditional distribution $b_i|g_i$ can be derived from (2.1) and is $\frac{\sum_{h=1}^{H} q_h(g_i) \Pr_h(b_i|g_i)}{\sum_{h=1}^{H} q_h(g_i)}$ where $q_h(g_i) = w_h \Pr_h(g_i)$.

2.2.2 Additional Specification

In contrast to Gage (2003), Fang *et al.* (2007), and Gage *et al.* (2008a), which use direct maximum likelihood (ML) estimation, the approach here employs the data augmented form for finite mixture models. Latent indicators, $z_i \sim \text{MN}(\pi_1, \cdots w_h), z_i \in \{1, \cdots, H\}$, denoting the component to which $(g_i, b_i)'$ belongs are introduced into the model. The resulting model,

$$(g_i, b_i)' \sim \sum_{h=1}^{H} \text{N}\left(M_h, \Sigma_h\right) 1_{[z_i=h]}, \tag{2.2}$$

is marginally equivalent to the original specification. Under this specification, ML estimation of model parameters proceeds through the Expectation-Maximization (EM) algorithm, and full Bayesian posterior inference proceeds by specifying prior distributions and utilizing MCMC methodology. The details can be found in McLachlan and Peel (2000) and Dey and Rao (2005). Whereas Gage (2003) uses a bootstrapping approach to estimate parameter uncertainty, the approach here pursues full Bayesian inference via a Gibbs sampling algorithm to directly provide parameter estimates and associated uncertainty (Gelfand and Smith, 1990; Diebolt and Robert, 1994). To complete the Bayesian specification, the following conjugate and assumed mutually independent prior distributions for the weights, w, the component coefficients, β_k (now including μ_k), and the component variances, σ_h, are

used here:

$$\begin{aligned} w &\sim \text{Dirichlet}(\omega), \\ \beta_h &\sim \text{N}(\beta_{h0}, \Sigma_{h0}), \\ \sigma_h^{-2} &\sim \text{G}(k_h, r_h), \end{aligned} \tag{2.3}$$

This specification avoids the use of Inverse-Wishart prior specifications for the covariance matrix of birthweight and gestational age.

2.2.3 Censored Continuous Gestational Age

The proposed framework readily deals with the often ignored issue of interval censorship of gestational age. Gestational age is reported in many ways, though all are typically interval censored. A standard reporting measure of gestational age is the Last Menstrual Period (LMP), which is reported as days since LMP. On the other hand, the gestational age data here are reported as integers representing the clinical estimate of the number of completed weeks of gestation (no uniform definition exists and the meaning of 'clinically estimated gestational age' varies by state). The approach suggested here is to imagine g_i is the true gestational age that is not observable, and take the observed g_i^c as an interval censored version of g_i. For the NCDBR data, the number of complete weeks is observed, so that $g_i^c \equiv \lfloor g_i \rfloor$. Defining $g_i \equiv g_i^c + u_i$, let $u_i \in [0, 1)$ to take the role of an unknown parameter. If g_i^c is interpreted differently, the specification may be modified accordingly. For instance, for LMP gestational age a Berkson measurement error model could be introduced, centering true g_i around the observed gestational age in days.

Upon specification of a prior, u_i may be seamlessly incorporated into the posterior sampling scheme. The simple prior used here is $u_i \sim \text{Unif}[0, 1)$. However, it may be argued that, given g_i^c, the distribution for g_i is likely to put more mass on days later in the week, i.e., the probability of birth increases on a daily basis, particularly for preterm and early term gestational ages. Thus, a more general beta prior for u_i is an alternate choice. Using

$u_i \sim \text{Beta}(a_i, r_i)$ specifies a non-conjugate prior for this model, requiring a Metropolis-Hastings or importance sampling step in the model fitting. The truncated conjugate prior $u_i \sim \text{N}(\theta_i, \tau_i^2)1_{[0,1)}$ may also be considered.

Recognizing the censored nature of reported gestational age measurements allows: (1) treatment of gestational age as a continuous parameter; (2) appropriate assessment of the uncertainty associated with censorship of gestational age; and (3) learning about the actual effect of the censorship (u_i) from the data.

Clinically estimated gestational age and LMP measurements are known to have error, with certain sub-populations possibly having more or less accurate reporting of g_i^c than others. The model presented here assumes that reported clinical estimates of gestational age are accurate. For the data, clinical estimates of gestational age for many sub-populations are considered to be relatively reliable after the year 2000, while for the remaining sub-populations this may not be so. The nature, effect, and size of such bias in the model is unclear. This consideration, in part, influenced the data restriction to the years 2004-2006. Alternative measures of gestational age such as ultrasound are more precise, but LMP and clinical estimations of gestational age remain much more prevalent. As such, models that can account for measurement error are still needed.

2.3 Bivariate Modeling Vs. Conditional Modeling

A wide literature cautions against the "fallacy of controlling for an intermediate outcome" (Gelman, July 18 2008; Delbaere *et al.*, 2007; Hernández-Díaz *et al.*, 2006; Rosenbaum, 1984; Pearl, 1995; Greenland *et al.*, 1999; Pearl, 2000; Robins *et al.*, 1999, 2000; Rubin, 2004). As is now understood, adjusting for an intermediate variable can result in the other observed covariate effects being wrongly boosted, attenuated, or even reversed. This happens for two reasons: (1) indirect effects of covariates that were mediated through the intermediate variable are no longer attributed to the covariates, and (2) spurious associations are artificially induced by back-door criterion violations caused by conditioning on an intermediate variable. These issues were noted by Gage (2003), and the above citations connect this

rightful concern to additional literature.

Using the data subset described in Section 2.1.2, the extent of change brought about by these issues is demonstrated in regression coefficients using ordinary least squares. Table 2.1 shows the coefficients resulting from birthweight regressions with and without gestational age as a covariate. Nearly all coefficients change between regressions. For example, smoking and NHB mother are attenuated, and infant sex and HS mother are boosted. Some coefficients even have sign changes. Since the contribution of lost mediated effects cannot be separated from that of the back-door criterion violations, intermediate variables should not be used as covariates if coefficients are to retain meaningful interpretations.

Covariate	Birthweight Regression Coefficients			
Intercept	-3578.8	(-3606.9, -3550.7)	3385.5	(3379.7, 3391.4)
Reported Smoking	-187.8	(-192.5, -183.0)	-227.2	(-233.4, -221.0)
Male Infant	126.2	(123.4, 129.0)	114.1	(110.3, 117.8)
Not Married	-36.2	(-39.8, -32.5)	-39.6	(-44.4, -34.7)
NHB Mother	-176.5	(-180.3, -172.7)	-233.7	(-238.7, -228.6)
Hispanic Mother	-70.2	(-75.1, -65.2)	-24.3	(-30.8, -17.9)
Complete MS	-30.1	(-36.9, -23.3)	-25.9	(-34.7, -17.0)
Some HS	-30.2	(-34.9, -25.5)	-39.2	(-45.3, -33.0)
Some College	26.5	(22.3, 30.6)	27.3	(21.8, 32.7)
Completed College	28.5	(23.9, 33.0)	65.4	(59.4, 71.3)
Maternal Age 15-19	-35.4	(-41.2, -29.5)	-26.9	(-34.5, -19.2)
Maternal Age 20-24	-27.0	(-31.0, -22.9)	-14.0	(-19.4, -8.7)
Maternal Age 30-34	18.1	(13.9, 22.2)	0.8	(-4.6, 6.3)
Maternal Age 35-40	21.9	(16.5, 27.2)	-15.3	(-22.3, -8.3)
Maternal Age 41-45	-0.4	(-11.2, 10.3)	-58.7	(-72.8, -44.6)
First Birth Infant	-120.1	(-123.3, -116.9)	-93.9	(-98.1, -89.7)
Gestational Age	180.2	(179.5, 180.9)		

Table 2.1: Birthweight (standard) regressions with and without Gestational Age included. 95% confidence intervals are included.

However, excluding intermediate variables from analyses does not seem reasonable in the birthweight and gestational age context. Ignoring gestational age, or birthweight, entails a large loss of information (Wilcox and Skjaerven, 1992). This is not necessary, however, if one employs appropriate joint modeling techniques. A correctly modeled bivariate relationship of birthweight and gestational age may replace the use of either as an intermediate variable

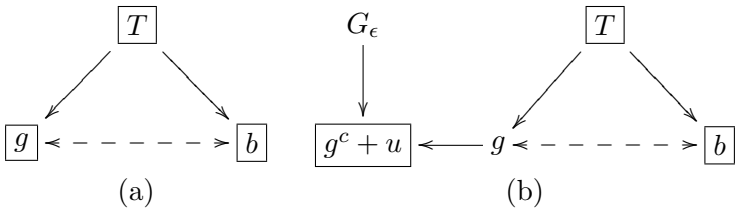

Figure 2.3: (a) shows a treatment or risk factor T affecting the joint variable of birthweight and gestational age. (b) adjusts the setting to reflect the complication of mis-measured gestational age.

in a conditional model. To demonstrate this, consider the situation of Figure 2.3a, and suppose there is interest in the effect of T on b.

Under suitable circumstances, regressing b on T estimates β_T^c, which captures a causal effect that includes both direct effects from T and indirect effects from T mediated through g. Adding g as a covariate results in a new estimate, β_T^d, which only includes the direct effect of T on b. Just as g blocks the indirect path from T to b in Figure 2.3a, adding g as a covariate blocks the indirect effects of T. Because the relationship between covariates is not uncovered through regression, β_T^c cannot be recovered from β_T^d.

Nonetheless, if β_T^d is the desired estimand, then g must have no measurement error when used as a covariate. In Figure 2.3b, g is measured with error, and denoted by $g^c + u$. Adding $g^c + u$ as a covariate rather than g results in a new estimate, β_T^ϵ, which comprises the direct effect and some of the indirect effect of T, as determined by the amount of measurement error in $g^c + u$. This is a result of mis-estimating the true direct effect of g on b, β_g^d, with β_g^a, which is an attenuated estimate as a result of the mis-measured $g^c + u$. Using the full direct effect of g on b, β_g^d, blocks the indirect path, but the attenuated β_g^a allows the path to be partially open, just as the indirect path is partially open in Figure 2.3b depending on how different $g^c + u$ is from g.

The situation is worse yet if there is an unmeasured confounder U that affects both g and b, as in Figure 2.4a. This is likely the case for birthweight and gestational age, where unknown genetic factors certainly play the role of U. As always, unmeasured confounders pose difficulties, but U in 2.4a is not even a confounder with T in the traditional sense

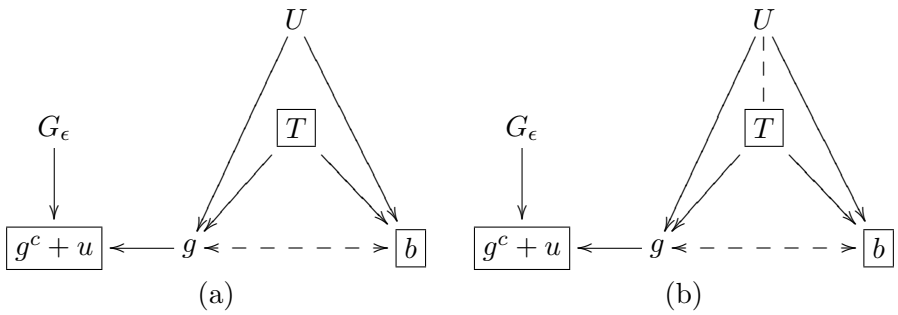

Figure 2.4: (a) demonstrates a setting with potential for a back-door criterion violation. (b) The back-door criterion violation is realized if g or $g^c + u$ are conditioned on.

and only influences g and b. Regardless, conditioning on the downstream variable $g^c + u$, results in the spurious relationship between U and T depicted in Figure 2.4b. This opens a back-door path into T from b through U resulting in a new estimate, $\beta_T^?$, which includes the direct effect of T on b, some of the indirect effect of T on b mediated through g due to mis-measured $g^c + u$, and now a spurious relationship between T and b carried through U because of conditioning on on the intermediate variable $g^c + u$.

A proposal that (1) avoids the artificially produced association between T and U, and (2) focusses on the causal rather than direct estimand for T is to condition on the residuals of g (actually read $g^c + u$, throughout this paragraph) from a previous regression on T, $\hat{\epsilon}_{g|T}$, rather than g itself. This approach essentially attempts to model the relationship between the intermediate variable and T, and remove the effects of T on g before estimating the effects of T on b. Then, (1) follows because by definition $\hat{\epsilon}_{g|T}$ is not related to T and so is not an intermediate variable, as is demonstrated in Figure 2.5, and (2) follows since correctly adjusted $\hat{\epsilon}_{g|T}$ represents g before being changed by T, so any changes in b are attributable to T only. This approach cannot avoid the effects of mis-measured $g^c + u$, but if u is is appropriately estimated, then $g^c + u \approx g$.

The bivariate model can be seen to take advantage of the full structure of Figure 2.5 in order to follow the above proposal and explicitly avoid intermediate variables. For each

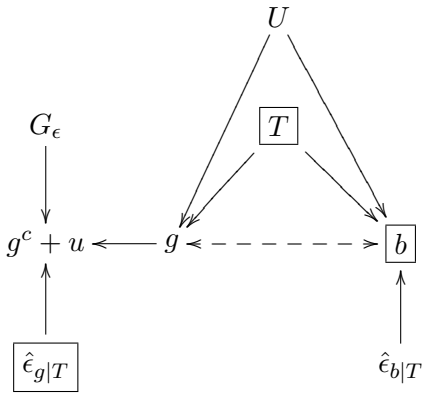

Figure 2.5: Conditioning on the residuals $\hat{\epsilon}_{g|T}$ does not result in a back-door criterion violation.

component, the two conditional distributions of the model are

$$\begin{aligned}
\Pr(b|g,x) &= N(b_i|\mu_{b,h} + x'\beta_{b,h} + (g^c + u - (\mu_{g,h} + x'\beta_{g,h}))\beta_{*,h}, \sigma_{b|g,h}) \\
&= N(b_i|\mu_{b,h} + x'\beta_{b,h} + \hat{\epsilon}_{g|x}\beta_{*,h}, \sigma_{b|g,h}) \\
\Pr(g^c + u|b,x) &= N\left(\mu_{g,h} + x_i'\beta_{g,h} + \frac{\beta_*}{\beta_{*,h}^2 + \left(\frac{\sigma_{b|g,h}}{\sigma_{g,h}}\right)^2}(b - (\mu_{b,h} + x_i'\beta_{b,h})), \sqrt{\sigma_{g,h}^2(1-\rho_h^2)}\right) \\
&= N\left(\mu_{g,h} + x_i'\beta_{g,h} + \frac{\beta_{*,h}}{\beta_{*,h}^2 + \left(\frac{\sigma_{b|g,h}}{\sigma_{g,h}}\right)^2}\hat{\epsilon}_{b|x}, \sqrt{\sigma_{g,h}^2(1-\rho_h^2)}\right)
\end{aligned}$$

so that each conditional distribution is based on the residuals of the conditioned margin. This type of conditional distribution is implied by the bivariate normal distribution. Thus, the bivariate normal model replaces the use of either b or g as intermediate variables with a modeled bivariate relationship. This avoids the use of either as an intermediate variable, which avoids potential bias of estimates from back door criterion violations. Further, coefficients estimated from this bivariate approach attempt to include both direct effects and indirect effects of covariates, which are generally more relevant than estimating only direct effects. If T has a beneficial direct effect but an overall detrimental effect, the direct effect is of little importance.

2.4 Identifiability

2.4.1 Alternative Non-Identified Parameterization

Correlation between MCMC posterior draws of parameters can render attempted posterior sampling useless (Gelfand and Sahu, 1999; Gelfand *et al.*, 1995). The centering of g in model (2.1) curbs such unattractive circumstances. Without centering, i.e., replacing $g_i - (\mu_{g,h} + z_i'\beta_{g,h})$ with only g_i, $E[b_i|z_i = 1] = \mu_{b,h} + \mu_{g,h}\beta_{*,h} + z_i'(\beta_{b,h} + \beta_{g,h}\beta_{*h})$. It follows that $\mu_{b,h}, \beta_{*,h}$ and $\beta_{b,h}$ will tend to drift as only the sums they are involved in are identified.

2.4.2 More Identifiability and Number of Components

Model (2.2) is invariant under re-ordering of the labels h, i.e., the $h!$ different parameterizations result in identical models. This well known conundrum for mixture models, known as label switching, is discussed by Jasra *et al.* (2005). Often, order constraints on parameters (e.g. $\mu_i < \mu_j$ for $i < j$) are utilized to identify components. This was not necessary under usual specifications of the model as no label-switching was observed in the mixing. This appears to be the result of the mixing relative to the high-dimensional nature of the proposed mixture model. In essence, for label switching to occur, a components parameters (e.g., two intercept parameters, one slope parameter, two variance parameters) must be exchanged with their counterparts in another component.

When $H = 4$ or more components are specified, the posterior distributions become multimodal (within the symmetric multimodality induced by label switching), and mixing across posterior modes becomes poor. With $H = 4$ components, observed parallel posterior chains with different initial value specifications did not meet. Instead, each exhibited intermittent periods of apparent stability punctuated by sporadic re-configurations that often did not improve log-likelihoods and rarely returned to the original configurations. The re-configurations amounted to slight changes in component location and almost no detectable difference in covariate coefficients. Thus, the mixing issue appears to be primarily one of location of the residual components. Regardless, the posterior chains showed several different

plausible models, none of which appeared preeminent, and across which mixing was poor. Indeed, label switching indicating good mixing did not occur. It is possible that a Metropolis step within the Gibbs sampler could improve mixing, but this was not attempted. This same circumstance of numerous adequate models no doubt exists under ML estimation, but is more easily uncovered through Bayesian analysis since in ML estimation only a single model is returned once the maximization algorithm has converged to some mode.

Model selection involving competing unconverged chains is a difficult issue. One pragmatic though somewhat ad hoc approach is to use an EM algorithm to find the best initial values as judged by largest likelihood, and proceed with full Bayesian inference using the stable part of the chain. Various competing models then may be pragmatically chosen using minimum posterior predictive loss in cross-validation (Gelfand and Ghosh, 1998), or naive Bayesian information criterion (BIC). Although it is not theoretically appropriate to use BIC in the finite mixture model setting even for converged chains, it has seen some application and success (McLachlan and Peel, 2000), and so this criterion is used in this work. For both three component ($H = 3$) and two component ($H = 2$) models, the mixing issues described above were not observed. Indeed, proper identification of a two component ($H = 2$) model is shown by Frimpong *et al.* (2009). Thus, these chains were assumed to have converged, and models were compared using the BIC. The data set of Section 2.1.2 strongly suggested the superiority of three component ($H = 3$) model to the two component ($H = 2$) model. The choice to avoid comparison to any four component ($H = 4$) models was driven by the mixing issues described above, and is thus an artifact of the operational fitting of the model rather than a judgement of clinical significance or a model choice criterion.

2.5 Model Demonstration

This section demonstrates the three component ($H = 3$) model using thedata described in Section 2.1.2. A wide range of alternative prior distributions and initial value specifications produced only slightly varied results in the three component ($H = 3$) case, and thus the demonstration is restricted to specifications of Table 2.2. Burn in was set at 5,000, and the

results of this section were generated from the subsequent 100,000 MCMC draws provided by the Gibbs sampler directly available under the specification (2.1). The mixing of individual chains did not show lack of convergence.

	Comp. 1	Comp. 2	Comp. 3
	Initial Values		
w	.34	.33	.33
μ_b	3000	2500	1500
μ_g	40	37	33
σ_b^2	250000	250000	250000
σ_g^2	2	2	2
β	$\vec{0}$	$\vec{0}$	$\vec{0}$
	Prior Hyperparamter Values		
p	1	1	1
μ_b	3000	2500	1500
μ_g	40	37	33
β_0	μ_b, μ_g, $\vec{0}$	μ_b, μ_g, $\vec{0}$	μ_b, μ_g, $\vec{0}$
Σ_0	$1000I$	$1000I$	$1000I$
k	1	1	1
r	1	1	1

Table 2.2: Initial values and prior distribution specifications for the results of a three ($H = 3$) component model used through out Section 2.5.

Where useful, inference under the model is illustrated through a series of prototypical individuals, $\mathcal{A}$ through $\mathcal{H}$. $\mathcal{A}$ through $\mathcal{H}$ represent the possible configurations of NHW/NHB, reported smoking, and reported marital status, for a 25-30 year old mother at the high school education level with a male infant. The covariate configurations of $\mathcal{A}$ through $\mathcal{H}$ are given in Table 2.3.

2.5.1 Bivariate Regression

One benefit of using a bivariate regression is that a single model produces estimates of the relationship of birthweight and gestational age to covariates, and to each other, simultaneously. Further, the mixture model framework provides $H = 3$ regressions – not just one – with each component supporting a separate regression. In addition to allowing improved flexibility in the distributional shapes that may be captured by the model, it also provides

Individual	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{H}$
Mother Reported Not Married	1	0	1	0	1	0	1	0
Non-Hispanic Black Mother	1	1	0	0	1	1	0	0
Reported Maternal Smoking	1	1	1	1	0	0	0	0

Table 2.3: Individuals $\mathcal{A}$ through $\mathcal{H}$ provide 8 risk factor sets for mothers used for demonstration in Section 2.5 and tables 7 through 10. All mothers are 25-30 years old and at the high school education level with a male infant.

the potential to uncover the differential strength of covariate effects across components. In contrast to Table 2.1, Table 2.4 shows the ability to explicitly model and detect how relationships differ by component sub-populations.

	Covariate	Component $h = 1$		Component $h = 2$		Component $h = 3$	
BW	Smoking	-205.0	(-211.6, -198.4)	-227.8	(-241.3, -214.6)	-85.1	(-116.3, -53.9)
	Male Infant	137.2	(133.2, 141.2)	80.7	(72.4, 88.9)	52.6	(29.7, 75.7)
	Not Married	-26.4	(-31.5, -21.2)	-40.3	(-51.0, -29.7)	-83.6	(-109.8, -57.3)
	NHB Mother	-188.4	(-193.7, -183.0)	-231.0	(-241.8, -220.0)	-318.0	(-344.9, -291.2)
	Hispanic Mother	-49.1	(-56.1, -42.1)	44.0	(29.6, 58.5)	111.3	(76.5, 145.9)
	Complete MS	-26.3	(-35.8, -16.9)	-8.8	(-27.7, 10.1)	23.5	(-18.5, 65.1)
	Some HS	-35.9	(-42.5, -29.4)	-24.9	(-38.0, -11.7)	10.8	(-20.6, 42.0)
	Some College	20.1	(14.4, 25.8)	47.2	(35.4, 59.2)	49.6	(20.4, 78.7)
	College	27.8	(21.5, 34.2)	125.9	(112.9, 139.0)	201.5	(169.6, 233.1)
	Maternal Age 15-19	-51.1	(-59.1, -43.0)	8.5	(-7.5, 24.4)	43.8	(8.0, 79.3)
	Maternal Age 20-24	-29.6	(-35.3, -24.0)	12.2	(0.7, 23.7)	71.0	(42.3, 100.0)
	Maternal Age 30-34	18.6	(12.8, 24.5)	7.6	(-4.6, 19.9)	17.1	(-12.9, 46.9)
	Maternal Age 35-40	25.2	(17.7, 32.8)	-41.6	(-57.1, -25.9)	-15.2	(-50.2, 19.4)
	Maternal Age 41-45	2.3	(-12.6, 17.1)	-88.4	(-116.5, -60.1)	-35.4	(-83.2, 12.8)
	First Birth	-55.1	(-59.6, -50.6)	-116.9	(-126.3, -107.6)	-162.0	(-186.5, -137.7)
	Residuals GA	104.7	(102.0, 107.5)	146.7	(143.0, 150.4)	146.2	(143.1, 149.2)
GA	Smoking	-0.06	(-0.08, -0.04)	-0.36	(-0.41, -0.31)	-0.30	(-0.50, -0.11)
	Male Infant	-0.01	(-0.02, -0.00)	-0.12	(-0.16, -0.09)	-0.09	(-0.22, 0.05)
	Not Married	0.07	(0.06, 0.08)	-0.07	(-0.11, -0.03)	-0.47	(-0.63, -0.31)
	NHB Mother	-0.03	(-0.05, -0.02)	-0.43	(-0.47, -0.39)	-1.75	(-1.91, -1.59)
	Hispanic Mother	0.19	(0.17, 0.21)	0.42	(0.37, 0.47)	0.52	(0.30, 0.73)
	Complete MS	0.08	(0.06, 0.11)	-0.07	(-0.14, 0.01)	-0.04	(-0.33, 0.24)
	Some HS	0.01	(-0.00, 0.03)	-0.11	(-0.16, -0.06)	-0.04	(-0.23, 0.16)
	Some College	-0.04	(-0.05, -0.02)	0.07	(0.03, 0.12)	0.20	(0.02, 0.38)
	College	0.05	(0.04, 0.07)	0.39	(0.34, 0.44)	0.94	(0.74, 1.13)
	Maternal Age 15-19	-0.01	(-0.03, 0.01)	0.09	(0.03, 0.15)	0.08	(-0.15, 0.31)
	Maternal Age 20-24	0.02	(0.01, 0.03)	0.11	(0.06, 0.15)	0.39	(0.21, 0.57)
	Maternal Age 30-34	-0.04	(-0.06, -0.03)	-0.04	(-0.09, 0.00)	0.12	(-0.06, 0.31)
	Maternal Age 35-40	-0.09	(-0.11, -0.07)	-0.21	(-0.26, -0.15)	-0.02	(-0.24, 0.20)
	Maternal Age 41-45	-0.10	(-0.13, -0.06)	-0.39	(-0.51, -0.28)	-0.22	(-0.59, 0.14)
	First Birth	0.40	(0.39, 0.42)	-0.19	(-0.23, -0.16)	-0.60	(-0.75, -0.46)

Table 2.4: Birthweight and gestational age regression coefficients with 95% credible intervals for in each mixture model component.

2.5.2 Mixture Sub-Populations

As emphasized by Gage (2003), a benefit of the mixture model approach is that the components provide a natural classification mechanism. In finite mixtures of regressions, this classification is an augmentation of the covariate set because the mixture feature of the model is defined on the residuals. After covariance adjustment, the leftover structure defines the components and the corresponding memberships. Posterior estimates of the location and shape parameters for the three components are given in Table 2.5. The component configuration determines distribution location and shape, and is governed by the covariates, which creates flexibility in modeling. For example, Figure 2.6 displays a general lowering in birthweight and lengthening of gestational age towards shorter ages for individual $\mathcal{A}$ relative to individual $\mathcal{H}$.

	Component $h = 1$		Component $h = 2$		Component $h = 3$	
w_k	0.716	(0.708, 0.724)	0.249	(0.241, 0.257)	0.035	(0.034, 0.036)
$\sigma^2_{b,k}$	175073	(173808, 176345)	127073	(123911, 130318)	131820	(124370, 139481)
$\sigma^2_{g,k}$	0.96	(0.95, 0.97)	2.48	(2.42, 2.54)	13.23	(12.78, 13.67)
$\mu_{b,k}$	3514	(3507, 3521)	3103	(3088, 3118)	1899	(1864, 1934)
$\mu_{b,k}$	39.59	(39.58, 39.61)	38.26	(38.20, 38.32)	33.29	(33.07, 33.51)
ρ_k	0.238	(0.232, 0.2425)	0.544	(0.533, 0.555)	0.826	(0.816, 0.835)

Table 2.5: Location and shape parameter estimates with 95% credible intervals for each mixture model component.

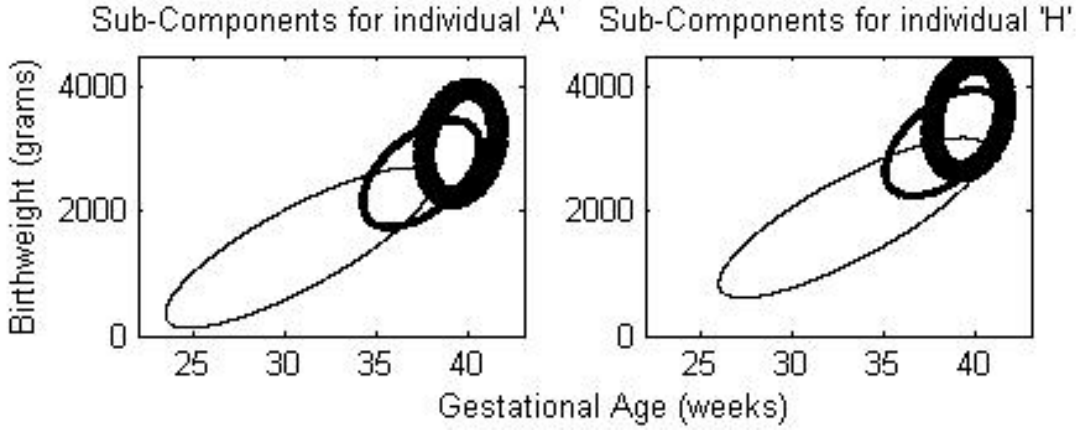

Figure 2.6: Posterior point estimate of the component configurations for individuals $\mathcal{A}$ and $\mathcal{H}$. The ellipses correspond to contours containing $\approx 86.5\%$ of component mass. The thickness conveys the relative proportions in the mixture distribution of Table 2.5.

Component Composition	Component 1	Component 2	Component 3	Overall
Subcomponent Size	240679.77 (2401180, 241120)	83702.95 (83301, 84277)	11746.28 (11600, 11905)	336129
Maternal Smoking	11.6 (11.6, 11.7)	12.0 (11.8, 12.2)	14.3 (13.9, 14.7)	11.8
Male Infant	51.0 (50.9, 51.1)	51.1 (50.9, 51.4)	53.1 (52.4, 53.8)	51.1
Not Married	38.1 (38.0, 38.2)	38.8 (38.5, 39.1)	44.5 (44.0, 45.2)	38.5
NHB Mother	23.3 (23.2, 23.3)	23.9 (23.7, 24.1)	30.8 (30.3, 31.3)	23.7
Hispanic Mother	16.4 (16.3, 16.5)	16.5 (16.3, 16.8)	15.0 (14.6, 15.6)	16.4
Completed MS	7.2 (7.1, 7.2)	7.3 (7.2, 7.5)	6.8 (6.5, 7.2)	7.2
Some HS	15.9 (15.8, 16.0)	16.3 (16.1, 16.6)	18.0 (17.5, 18.4)	16.1
Some College	22.1 (22.0, 22.2)	22.1 (21.8, 22.3)	22.3 (21.8, 22.9)	22.1
College	26.1 (26.0, 26.2)	25.6 (25.4, 25.8)	23.0 (22.4, 23.4)	25.9
Maternal Age 15-19	11.4 (11.4, 11.5)	11.7 (11.5, 12.0)	13.2 (12.8, 13.5)	11.6
Maternal Age 20-24	27.1 (27.0, 27.2)	27.4 (27.1, 27.6)	26.8 (26.4, 27.5)	27.1
Maternal Age 30-44	22.1 (22.0, 22.2)	21.8 (21.6, 22.2)	21.8 (21.3, 22.3)	22.0
Maternal Age 35-39	10.1 (10.0, 10.2)	10.0 (9.8, 10.2)	10.9 (10.6, 11.2)	10.1
Maternal Age 40-44	1.9 (1.8, 1.9)	1.9 (1.8, 2.0)	2.4 (2.2, 2.6)	1.9
First Birth Infant	40.8 (40.7, 40.8)	41.3 (41.0, 41.5)	45.0 (44.3, 45.8)	41.0

Table 2.6: The compositional makeup, given in percentages (except for the first row, which is a count) of each mixture model component as observed in posterior sampling. Additionally, the final column labeled 'Overall' shows the characteristics of the original population.

Under the latent indicator specification (2.2), the components are formed by repeatedly stochastically assigning every individual to membership in one of the components. For any posterior iteration t, let $\theta^{(t)}$ denote the current regression coefficients and configurations of the components. Given $\theta^{(t)}$, every individual i is randomly assigned to a component $h_i^{(t)}$, $z_i^{(t)} = h_i^{(t)}$, according to the relative probability that the residuals resulting from b_i, g_i, and x_i' under $\theta^{(t)}$ belong to component h. The membership configuration then forms the basis for updating $\theta^{(t+1)}$, and the process is repeated. The posterior distribution of z_i expresses the propensity for individual i to join component h and allows learning about the propensities of individual i, or perhaps the propensities of a collection of individuals. For instance, the overall composition of covariates across components is given in Table 2.6.

Table 2.6 was generated from 1000 random assignments of every individual i to a component according to the posterior distribution of z_i. In each one of the 1000 complete assignments, the covariate distribution was calculated, and from these 1000 samples, the mean and the 95% credible intervals for the covariate distribution were determined. Table 2.6 shows that the distribution of the covariates is relatively uniform among components. Thus, there seems to be no combination of the specified covariates that strongly interact to inform component membership, i.e., membership is driven by a factor that has not been identified. To the extent that covariates are balanced between the three components, there would seem to be no benefit in incorporating covariates to influence the mixing proportions

since the covariates do not provide further information beyond the overall proportions. However, Gage *et al.* (2008b) found that covariates did affect the mixing proportions in a univariate mixture model for birthweight. Despite the inability to predict component membership from the specified covariates, component 3 is associated with elevated vulnerability to adverse birth outcomes and, so, is the natural sub-population to focus on for exploration of risk.

2.5.3 Prediction

Bivariate predictions can be made from the model, as well as predictions from the induced distributions of $g_i|b_i, x_i$ and $b_i|g_i, x_i$. Bivariate predictions are given by:

$$E(b_i, g_i|x_i) = \sum_{h=1}^{H} w_h \begin{bmatrix} \mu_{b,h} + x_i'\beta_{b,h} \\ \mu_{g,h} + x_i'\beta_{g,h} \end{bmatrix}. \tag{2.4}$$

Tables 2.4 and 2.5 give some indication of bivariate predictions, but they provide estimates and credible intervals for parameters rather than predictions. Calculating (2.4) at each posterior iteration t provides the correct estimates and uncertainties.

Predictions of birthweight given gestational age (or vice-versa) may be conditional on any continuous value, e.g., birthweight conditional on the true gestational age, not only censored integer gestational age, as given by

$$E(b_i|g_i, x_i) = \sum_{h=1}^{H} \frac{w_h N(g_i|\bar{x}_i'\beta_{g,h}, \sigma^2_{g,h})}{\sum_{j=1}^{H} w_j N(g_i|\bar{x}_i'\beta_{g,j}, \sigma^2_{g,j})} \left(\bar{x}_i'\beta_{b,h} + (g_i - \bar{x}_i'\beta_{g,h})\beta_{*,h}\right) \tag{2.5}$$

$$E(g_i|b_i, x_i) = \sum_{h=1}^{H} \frac{w_h N\left(b_i \middle| \bar{x}_i'\beta_{b,h}, \frac{\sigma^2_{b|g,h}}{1-\rho_h^2}\right)}{\sum_{j=1}^{H} w_j N\left(b_i \middle| \bar{x}_i'\beta_{b,j}, \frac{\sigma^2_{b|g,j}}{1-\rho_j^2}\right)} \left(\bar{x}_i'\beta_{g,h} + \tilde{B}_{*,h}(b_i - \bar{x}_i'\beta_{b,h})\right) \tag{2.6}$$

where $\tilde{B}_{*,h} = \beta_{*,h}/(\beta^2_{*,h} + (\sigma_{b|g,h}/\sigma_{g,h})^2)$ and μ has been incorporated into β for compactness, which has generated the byproduct $\bar{x}$. While the conditional prediction or distribution of gestational age given birthweight is not a standard consideration, it may be useful, e.g., in

imputation of missing values and detection of mis-measured gestational ages.

A related conditional prediction is the small for gestational age cutpoint, $SGA(g_i)$, which is found through area prediction in the conditional model of birthweight given gestational age. For a given covariate level, x_i, estimation of $SGA(g_i)$ is made using

$$0.1 = \int_{-\infty}^{SGA(g_i)} \sum_{h=1}^{H} \frac{w_h N(g_i|\mu_{g,h} + z_i'\beta_{g,h}, \sigma^2_{g,h})}{\sum_{j=1}^{H} \pi_j N(g_i|\mu_{g,j} + z_i'\beta_{g,j}, \sigma^2_{g,j})} \times \tag{2.7}$$
$$N(b_i|\mu_{b,h} + z_i'\beta_{b,h} + (g_i - (\mu_{g,h} + z_i'\beta_{g,h}))\beta_{*,h}, \sigma_{b|g,h})db$$

In Tables 2.7, 2.8, and 2.9, conditional predictions of birthweight given gestational age, gestational age given birthweight, and the SGA cutpoint are given for individuals $\mathcal{A}$ through $\mathcal{H}$ from Table 2.3. Prediction and interval curves are available for the three conditional predictions described above, but are only demonstrated for the SGA cutpoint in Figure 2.7, which contrasts SGA for individuals $\mathcal{A}$ and $\mathcal{H}$. The differences in predictions seen in Tables 2.7, 2.8, and 2.9 are due to the different covariate configurations of individual $\mathcal{A}$ through $\mathcal{H}$ that result in different joint birthweight gestational age distributions, as seen in Figure 2.6.

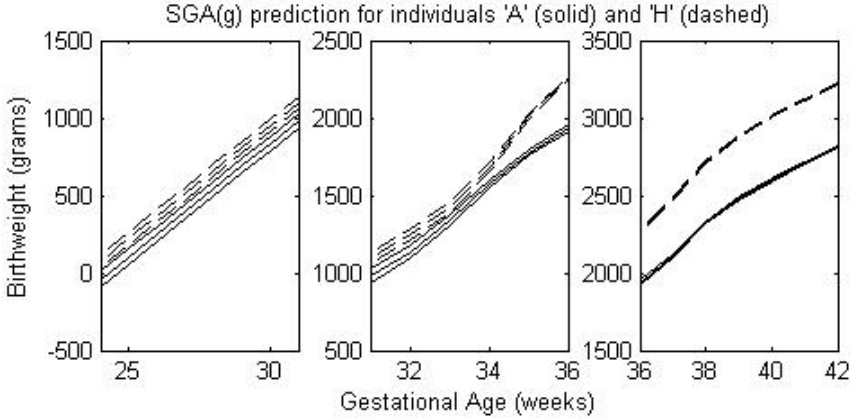

Figure 2.7: Conditional predictions of the small for gestational age cutpoint $SGA(g)$ for individuals $\mathcal{A}$ and $\mathcal{H}$. The single gestational age axis is separated into 3 plots so that the 95% credible intervals may be examined. The predictions were generated from the conditional distributions implied by the joint distributions represented in Figures 2.6. Table 2.9 provides related results for other individuals.

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
34	2039.2 (2018.3, 2059.7)	2.0547 (2.0308, 2.0779)	2103.8 (2076.6, 2130.3)	2113.3 (2085.9, 2140.1)
37	2579.5 (2564.9, 2594.2)	2.6160 (2.6005, 2.6316)	2743.8 (2729.8, 2757.7)	2780.0 (2767.2, 2792.9)
39	3008.7 (3001.0, 3016.5)	3.0415 (3.0333, 3.0498)	3183.5 (3176.3, 3190.6)	3216.3 (3209.6, 3223.2)
40	3130.9 (3122.8, 3139.1)	3.1632 (3.1545, 3.1719)	3309.8 (3302.3, 3317.4)	3341.5 (3334.4, 3348.6)
	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{H}$
34	2126.6 (2103.6, 2149.0)	2131.9 (2106.8, 2156.2)	2142.0 (2111.9, 2171.6)	2146.5 (2119.5, 2173.2)
37	2752.5 (2740.7, 2764.3)	2788.9 (2777.1, 2800.9)	2914.3 (2901.7, 2926.7)	2950.3 (2940.1, 2960.6)
39	3197.5 (3191.4, 3203.5)	3230.5 (3224.2, 3236.8)	3371.4 (3365.2, 3377.6)	3404.6 (3399.3, 3410.0)
40	3324.6 (3318.2, 3330.8)	3356.4 (3349.9, 3363.0)	3501.9 (3495.4, 3508.3)	3533.2 (3527.8, 3538.6)

Table 2.7: Conditional expectation of birthweight given gestational age along with 95% credible interval for individuals $\mathcal{A}$ through $\mathcal{H}$ at 34, 37, 39, and 40 weeks.

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
1500	32.41 (32.21, 32.62)	32.26 (32.03, 32.48)	31.77 (31.54, 31.99)	31.78 (31.55, 32.00)
2500	38.26 (38.22, 38.29)	38.17 (38.13, 38.21)	37.95 (37.91, 37.99)	37.88 (37.84, 37.92)
3500	39.76 (39.73, 39.78)	39.67 (39.65, 39.69)	39.66 (39.64, 39.68)	39.58 (39.56, 39.60)
4000	40.06 (40.04, 40.08)	39.98 (39.96, 40.01)	40.00 (39.98, 40.02)	39.92 (39.90, 39.94)
	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{H}$
1500	31.41 (31.22, 31.60)	31.41 (31.20, 31.61)	31.41 31.19 31.63	31.47 (31.28, 31.67)
2500	37.90 (37.87, 37.94)	37.83 (37.79, 37.87)	37.60 37.56 37.65	37.54 (37.50, 37.58)
3500	39.67 (39.66, 39.69)	39.59 (39.57, 39.61)	39.56 39.55 39.58	39.49 (39.47, 39.50)
4000	40.01 (39.99, 40.03)	39.93 (39.91, 39.95)	39.95 39.93 39.96	39.87 (39.86, 39.89)

Table 2.8: Conditional expectation of gestational age given birthweight along with 95% credible interval for individuals $\mathcal{A}$ through $\mathcal{H}$ at 1500, 2500, 3500, and 4000 grams.

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
34	1579.9 (1557.9, 1601.2)	1594.9 (1570.2, 1619.0)	1642.5 (1613.8, 1670.4)	1651.8 (1622.7, 1679.9)
37	2112.1 (2096.8, 2127.2)	2146.6 (2130.6, 2162.5)	2276.2 (2261.6, 2290.7)	2310.5 (2297.1, 2323.8)
39	2483.1 (2475.2, 2491.1)	2516.0 (2507.6, 2524.5)	2660.6 (2653.1, 2668.0)	2693.4 (2686.3, 2700.5)
40	2599.3 (2590.9, 2607.7)	2632.1 (2623.1, 2641.0)	2780.4 (2772.6, 2788.1)	2812.7 (2805.3, 2820.0)
	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{H}$
34	1666.0 (1641.8, 1689.7)	1670.9 (1644.3, 1696.7)	1679.5 (1647.4, 1711.0)	1683.8 (1654.6, 1712.7)
37	2285.8 (2273.2, 2298.1)	2320.3 (2307.7, 2332.8)	2446.8 (2433.8, 2459.9)	2480.8 (2469.9, 2491.7)
39	2674.3 (2667.9, 2680.6)	2707.2 (2700.6, 2713.8)	2850.6 (2844.1, 2857.1)	2883.6 (2878.0, 2889.3)
40	2794.6 (2788.0, 2801.1)	2827.0 (2820.2, 2833.8)	2974.2 (2967.5, 2980.8)	3006.1 (3000.4, 3011.8)

Table 2.9: SGA cutpoint predictions and 95% credible interval for individuals $\mathcal{A}$ through $\mathcal{H}$ at gestational ages 34, 37, 39, and 40.

2.5.4 Bivariate Distribution

Model (2.1) provides a bivariate distribution to capture the empirical joint distribution of birthweight and gestational age as seen in Figures 2.1 and 2.2. Such a model allows incorporation of covariates and provides a joint surface to use for inference, e.g., see Figure 2.8. Analysis is not limited to the previously discussed conditional inferences, as it can address joint inference associated with the joint distribution.

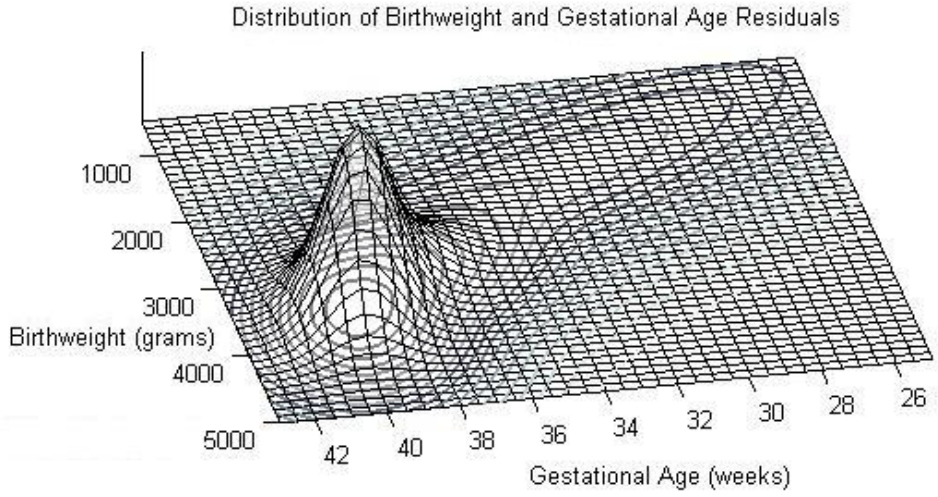

Figure 2.8: Point estimate of the surface of the mixture distribution for birthweight and gestational age for the referent individual $\mathcal{H}$. The orientation of this plot is a nonstandard $\approx 180^\circ$ rotational form. As a result, birthweight increases from top to bottom and gestational age decreases from left to right. Posterior 95% credible intervals of the surface tightly fit this curve, and so were not included in this image.

Table 2.10 provides estimates of the probability of both LBW and PTB for individuals $\mathcal{A}$ through $\mathcal{H}$, using

$$\Pr((b_i, g_i) \in LBW \times PTB | z_i) = \sum_{h=1}^{H} w_h \int_{LBW \times PTB} N\big(M_h, \Sigma_h\big). \tag{2.8}$$

Again for individuals $\mathcal{A}$ through $\mathcal{H}$, Table 2.10 provides probability estimates for two age inappropriate ($AI(g_i)$) birthweight classifications: $AI(35)$, less than 2000 grams for [35, 36) weeks gestational age, and $AI(37+)$, less than 2500 grams for greater or equal to 37 weeks gestational age. These probability estimates are provided using an expression similar to (2.8).

2.6 Discussion

This demonstration has highlighted the gradient of differences between individuals $\mathcal{A}$ through $\mathcal{H}$ with respect to the joint variable birthweight and gestational age. Specifically, a gradient

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
LBW+PTB	9.55% (9.27%, 9.83%)	8.97% (8.69%, 9.27%)	6.49% (6.29%, 6.69%)	6.01% (5.83%, 6.19%)
$AI(35)$	0.88% (0.83%, 0.93%)	0.78% (0.73%, 0.84%)	0.44% (0.41%, 0.47%)	0.39% (0.37%, 0.42%)
$AI(37+)$	1.52% (1.45%, 1.60%)	1.34% (1.27%, 1.41%)	0.64% (0.60%, 0.68%)	0.55% (0.52%, 0.59%)
	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{H}$
LBW+PTB	6.90% (6.73%, 7.07%)	6.44% (6.27%, 6.61%)	4.61% (4.49%, 4.74%)	4.26% (4.15%, 4.36%)
$AI(35)$	0.42% (0.39%, 0.45%)	0.37% (0.35%, 0.40%)	0.21% (0.20%, 0.23%)	0.20% (0.18%, 0.21%)
$AI(37+)$	0.59% (0.56%, 0.62%)	0.51% (0.48%, 0.53%)	0.23% (0.21%, 0.24%)	0.20% (0.18%, 0.21%)

Table 2.10: Probability estimates and 95% credible intervals for simultaneous LBW and PTB, $AI(35)$, and $AI(37+)$ for individuals $\mathcal{A}$ through $\mathcal{H}$ at gestational ages 34, 37, 39, and 40 weeks.

of impacts associated with the characteristics of individual $\mathcal{A}$ through the referent individual $\mathcal{H}$ have been quantified. For example, Figure 2.6 demonstrates how the overall joint distribution is less favorable for $\mathcal{A}$ than $\mathcal{H}$. As indicated in Table 2.4, race is the primary variable associated with distribution location difference of up to approximately -320 grams and approximately -1.75 weeks gestation, with the strongest differences appearing in the tail of the joint distribution. Smoking is also a major driver accounting for location difference of up to approximately -230 grams and approximately -0.35 weeks gestation and tends to affect birthweight in the main mass and gestational age in the tail of the distribution. Marital status contributes additional difference of up to approximately -80 grams and approximately -0.5 weeks gestation for unmarried women, primarily in the tail.

Because of the gradient of distributional differences from individuals $\mathcal{A}$ through $\mathcal{H}$, there is a resulting gradient of differences small for gestational age and expected birthweight conditional on gestational age, with the predictions separating by as much as approximately 400 grams in places. An analogous gradient occurs in the percentage of PTB and LBW infants (with up to an approximately 2 fold prevalence increase), and the percentage of age inappropriate births for gestational ages 35 and 37+ (with up to approximately 5 fold and approximately 8 fold prevalence increases, respectively).

Further detail of the varying impacts of individual covariates across the joint distribution is given in Table 2.4 and may be contrasted with Table 2.1. As discussed in Section 2.3, the joint variable framework coefficient estimates in Table 2.4 are free of the problem of treating birthweight or gestational age as intermediate variables.

Chapter 3

Extension of Binary Principal Stratification into Observational Settings

3.1 Introduction

Regardless of whether a study design is randomized or observational, intermediate variables are frequently present, e.g., in settings involving non-compliance, missing data, and surrogate endpoints. Under such circumstances, standard intention-to-treat or per-protocol analyses may not be sufficient to estimate treatment efficacy, and intermediate variables must be dealt with for causal inference. It is well documented that applying standard methods of pre-treatment variable adjustment to intermediate variables, such as regression on the intermediate variable, can result in post-treatment selection bias (Rosenbaum, 1984; Robins and Greenland, 1992). To illustrate, let $Y_i(T_i)$ and $D_i(T_i)$ be respectively the potential outcomes (Rubin, 1978) of the response of interest and the intermediate variable for unit i under an assigned binary treatment, $T_i = 0, 1$. In general, the comparison between $\{Y_i(0)\colon D_i(0) = d\}$ and $\{Y_i(1)\colon D_i(1) = d\}$ for all $i = 1, \ldots, n$ units in the study is not a causal effect when T_i affects D_i, because $\{i\colon D_i(0) = d\} \neq \{i\colon D_i(1) = d\}$.

A principled approach to handling intermediate variables in causal inference is principal stratification (PS, Frangakis and Rubin, 2002), in which one compares $\{Y_i(1) : S_i = s\}$ and $\{Y_i(0) : S_i = s\}$. Here $S_i = (D_i(1), D_i(0))$ is called a principal stratum. The key insight is that S_i is invariant under treatment assignment, so the principal strata may be used as pre-treatment variables. That is, comparisons within $S_i = s$, known as principal effects, are well-defined causal effects. The principal effects in $\{S_i \colon D_i(0) = D_i(1)\}$ and $\{S_i \colon D_i(0) \neq D_i(1)\}$ can be interpreted as direct and indirect effects of treatment on response, respectively (Rubin, 2004). Alternative definitions of direct and indirect effects are given in Robins and Greenland (1992) and Pearl (2001).

Since the principal strata S are not observed, the identifiability of principal effects relies on a set of substantive assumptions; see Section 3.2. PS is typically applied in controlled assignment settings where those assumptions are likely to hold (e.g., Barnard *et al.*, 2003; Gilbert *et al.*, 2003; Jin and Rubin, 2008). However, PS has been used in observational studies (e.g., Sjölander *et al.*, 2008; Flores and Flores-Lagunes, 2009), where the identifiability of PS estimands is challenging due to both the questionable assumption of no unmeasured confounding and the potential presence of direct effects of treatment on response. With minimal assumptions, unidentifiable PS estimands can be bounded (e.g., Balke and Pearl, 1997; Imai and Yamamoto, 2010); however, the resulting bounds can be too wide to be practically informative. Alternatively, one can explicitly model the identification assumptions as sensitivity parameters and examine the impacts on causal estimates for a range of plausible values of these parameters. For example, Sjölander *et al.* (2008) present such a sensitivity model for an observational study where direct effects are plausible. Their methodology does not explicitly deal with unmeasured confounding. Small and Rosenbaum (2008) provide a sensitivity analysis in an instrumental variables setting to identify the level of unmeasured confounding that would discount a significant treatment effect; their approach uses randomization permutation distributions rather than parametric models.

This article presents a sensitivity analysis methodology for PS that incorporates both unmeasured confounding and unknown direct effects. Two types of (simultaneous) unmea-

sured confounding are identified in PS: (1) S-confounding, which affects the estimation of principal strata, and (2) Y-confounding, which affects the estimation of effects on the response within principal strata. A model-based approach that can be used to examine the impacts on inferences of different variations of potential unmeasured confounding is presented. This strategy can also be used to assess sensitivity when direct effects are suspected. This is because in PS direct effects are operationally indistinguishable from Y-confounding.

The remainder of the article is organized as follows. Section 2 reviews the standard assumptions used in PS, clarifies the role of the assumption of no unmeasured confounding, and demonstrates the effects of unmeasured confounding on causal effect estimands. Section 3 develops a parametric approach to sensitivity analysis that addresses both unmeasured confounding and direct effects, and illustrates the ability of the approach to recover truth in the presence of confounding. Section 4 undertakes a sensitivity analyses for direct effects and unmeasured confounding in a medical example concerning influenza vaccination.

3.2 Confounding in PS

When T_i and D_i are binary, the possible principal strata are $S_i \in \{(0,0),(1,0),(1,1),(0,1)\}$. In the context of non-compliance, the S_i are often called, in the order shown above, never-takers ($S_i = n$), compliers ($S_i = c$), always-takers ($S_i = a$), and defiers ($S_i = d$), as in Angrist *et al.* (1996). Principal strata can be defined in settings other than non-compliance as well; for instance, smoking as a treatment, hypertension as a binary intermediate variable, and low birthweight as a response. The remainder of this article uses the familiar nomenclature of non-compliance to generically refer to S_i.

In order to identify the principal effects, the following assumptions are often made.

A1. *Stable unit treatment value assumption* (SUTVA) (Rubin, 1978). There is no interference between units and no different versions of any single treatment arm.

A2. *Monotonicity.* $D_i(1) \geq D_i(0)$ for all i, ruling out the principal stratum of defiers.

A3. *Exclusion restriction* (ER). If $D_i(1) = D_i(0)$, then $Y_i(1) = Y_i(0)$ for all i, implying always-takers and never-takers experience no direct effect of treatment on response.

A4. *No unmeasured confounding.* $(Y_i(0), Y_i(1), S_i) \perp\!\!\!\perp T_i | X_i$ for all i. This is often referred to as strong ignorability of assignment (Rubin, 1978).

A1 and A2 are assumed for the remainder of the article. However, this presentation departs from the typical causal inference set-up and does not assume A3 and A4; rather, the work assesses sensitivity of estimates to A3 and A4. As a step towards sensitivity analyses, unmeasured confounding is first characterized and the consequences for inference when A3 and A4 are incorrect but applied regardless are shown.

3.2.1 Characterizing Unmeasured Confounding

Let $Y_i = (Y_i(0), Y_i(1))$. The no unmeasured confounding assumption can be expressed as

$$\Pr(Y_i, S_i | T_i = 1, X_i) \quad = \quad \Pr(Y_i, S_i | T_i = 0, X_i), \tag{3.1}$$

for all i. Under (3.1), the PS setting may be represented graphically by Figure 3.1a. Unmeasured confounding arises, and (3.1) fails, when some possibly multidimensional variable U that effects $(Y_i(0), Y_i(1))$ and S—after adjustment for X—also influences Z. This situation is represented in Figure 3.1d.

When such a U exists, (3.1) factors as

$$\begin{aligned} & \Pr(Y_i, S_i | T_i = t, X_i, U_i) \\ = \; & \Pr(Y_i | T_i = t, S_i, X_i, U_i) \Pr(S_i | T_i = t, X_i, U_i) \\ = \; & \Pr(Y_i | T_i = t, S_i, X_i, U_i^{Y|S}) \Pr(S_i | T_i = t, X_i, U_i^S). \end{aligned}$$

Here, the unmeasured confounders are partitioned into $U^{Y|S}$ and U^S, which are the possibly overlapping subsets of U that affect each component of the likelihood. This factorization suggests that unmeasured confounding can arise via two pathways.

1. *S-confounding*: the distribution of S varies with T because of U^S (see Figure 3.1e), implying that

$$\Pr(S_i|T_i = 1, X_i) \quad \neq \quad \Pr(S_i|T_i = 0, X_i);$$

2. *Y-confounding*: within principal stata S, the distribution of Y varies with T because of $U^{Y|S}$ (see Figure 3.1f), implying that

$$\Pr(Y_i|T_i = 1, S_i = s, X_i) \quad \neq \quad \Pr(Y_i|T_i = 0, S_i = s, X_i).$$

When S-confounding or Y-confounding exists, (3.1) no longer holds, and inferences predicated on this assumption can be biased.

A third form of confounding, O(utcome)-confounding, which is displayed in Figure 3.1c, may be relevant in the PS framework. O-confounding represents the relationship, incidental to T, that drives the association between S and Y. O-confounding does not violate any of the derived results in this presentation, and so is not examined further.

3.2.2 Implications of Unmeasured Confounding and a False Exclusion Restriction

The importance of A3 and A4 can be illustrated in the simple setting of no covariates. Here, a common estimand of interest is the complier average causal effect (CACE),

$$\theta_c = E(Y_i(1)|S_i = c) - E(Y_i(0)|S_i = c).$$

Under A1–A4, the CACE is identifiable, and a consistent moment estimator is

$$\begin{aligned} \bar{\theta}_c^{obs} &= E(Y_i(1)|S_i = c, T_i = 1) - E(Y_i(0)|S_i = c, T_i = 0) \\ &= \frac{(p_{11}\pi_{11} - p_{10}\pi_{10}) + (p_{01}\pi_{01} - p_{00}\pi_{00})}{1 - \pi_{01} - \pi_{10}} \end{aligned} \tag{3.2}$$

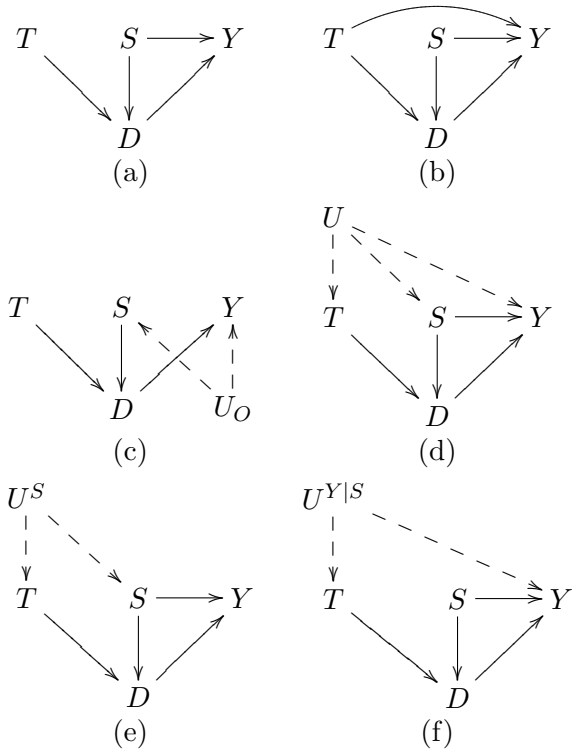

Figure 3.1: Possible sources of confounding: (a) None, (b) None, but direct effect of T on Y, (c) O (outcome), (d) Unmeasured, (e) S-confounding, and (f) Y-confounding.

where $p_{dt} = \Pr(Y_i^{obs} = 1|D_i = d, T_i = t)$ and $\pi_{dt} = \Pr(D_i = d|T_i = t)$, for $d = 0, 1, t = 0, 1$, and $Y_i^{obs} = T_i Y_i(1) + (1 - T_i)Y_i(0)$ (Imbens and Angrist, 1996; Angrist *et al.*, 1996). All quantities in (3.2) are estimable from the observed proportions. The value of p_{11} results from a mixture of compliers and always-takers, and the value of p_{00} results from a mixture of compliers and never-takers. A1-A4 identifies the complier contribution in each mixture.

The ER implies that, for always-takers and never-takers, there is no direct effect of treatment on response, so that $\Pr(Y_i(1) = 1|S_i = s, T_i = t) = \Pr(Y_i(0) = 1|S_i = s, T_i = t)$, $s \in \{a, n\}$. If, instead, there is a possibly unknown direct effect of treatment on response

(as illustrated in Figure 3.1b) for the always-takers or never-takers, then for some $s \in \{a, n\}$

$$\begin{aligned} \tau_s &= \Pr(Y_i(1) = 1 | S_i = s, T_i = t) - \\ &\quad \Pr(Y_i(0) = 1 | S_i = s, T_i = t) \neq 0, \quad \text{for } t = 0, 1. \end{aligned}$$

This assumes the direct effect τ_s is constant across treatment arms t for $s \in \{a, n\}$. An analogous τ_c is not defined as it would equal θ_c if the direct effect is also assumed constant across t; hence, the CACE includes both direct effect of treatment on response and indirect effects carried through the intermediate variable. For examples of direct effects in the always-taker or never-taker strata, see Hirano *et al.* (2000) and Imbens and Rubin (2010).

No unmeasured confounding implies that the distribution of principal strata are the same across treatments, so that $\Pr(S_i = a) = \Pr(S_i = a | T_i = 0) = \pi_{10}$ and $\Pr(S_i = n) = \Pr(S_i = n | T_i = 1) = \pi_{01}$. This is not true in the presence of S-confounding, where for some $s \in \{a, n\}$

$$\xi_s = \Pr(S_i = s | T_i = 1) - \Pr(S_i = s | T_i = 0) \neq 0.$$

With the ER, no unmeasured confounding further implies that the distribution of outcomes for the always-takers and never-takers are the same across treatments, so that

$$\begin{aligned} p_{10} &\overset{A2}{=} \Pr(Y_i^{obs} = 1 | S_i = a, T_i = 0) \\ &\overset{A4}{=} \Pr(Y_i(0) = 1 | S_i = a) \overset{A3}{=} \Pr(Y_i(1) = 1 | S_i = a), \\ p_{01} &\overset{A2}{=} \Pr(Y_i^{obs} = 1 | S_i = n, T_i = 1) \\ &\overset{A4}{=} \Pr(Y_i(1) = 1 | S_i = n) \overset{A3}{=} \Pr(Y_i(0) = 1 | S_i = n). \end{aligned}$$

It also implies that $\Pr(Y_i | S_i = c, T_i = 1) = \Pr(Y_i | S_i = c, T_i = 0)$. These are no longer true

in the presence of Y-confounding, since even with the ER, for some $s \in \{a, n, c\}$

$$\begin{aligned} \eta_s &= \Pr(Y_i(t) = 1 | S_i = s, T_i = 1) - \\ &\quad \Pr(Y_i(t) = 1 | S_i = s, T_i = 0) \neq 0, \quad \text{for } t = 0, 1. \end{aligned}$$

This assumes η_s is constant across treatment arms t for each $s \in \{a, n, c\}$.

When A3 and A4 fail, i.e., $\tau_{\{a,n\}} \neq 0, \eta_{\{a,n,c\}} \neq 0$, and $\xi_{\{a,n\}} \neq 0$, (3.2) is a biased estimator of θ_c since, letting $p_{st} = \Pr(Y_i(t) = 1 | S_i = s, T_i = t)$ and $\pi_{st} = \Pr(S_i = s | T_i = t)$,

1. $p_{10} = p_{a1} - \tau_a - \eta_a$ rather than p_{a1} when $\tau_a, \eta_a \neq 0$,
2. $p_{01} = p_{n0} + \tau_n + \eta_n$ rather than p_{n0} when $\tau_n, \eta_n \neq 0$,
3. $\pi_{10} = \pi_{a1} - \xi_a$ rather than π_{a1} when $\xi_a \neq 0$,
4. $\pi_{01} = \pi_{n0} + \xi_n$ rather than π_{n0} when $\xi_n \neq 0$, and
5. $E(Y_i(1) | S_i = c, T_i = 1) - E(Y_i(0) | S_i = c, T_i = 0) = \theta_c + \eta_c$, when $\eta_c \neq 0$.

These facts define a new estimator of θ_c when A3 and A4 do not hold, namely

$$\begin{aligned} \bar{\theta}_c^{adj} &= -\eta_c + \frac{(p_{11}\pi_{11} - (p_{10} + \tau_a + \eta_a)(\pi_{10} + \xi_a))}{1 - \pi_{01} - (\pi_{10} + \xi_a)} \\ &\quad - \frac{p_{00}\pi_{00} - (p_{01} - \tau_n - \eta_n)(\pi_{01} - \xi_n)}{1 - \pi_{10} - (\pi_{01} - \xi_n)}. \end{aligned} \tag{3.3}$$

A key insight of this formulation is that for $s \in \{a, n\}$, observable direct effects τ_s and Y-confounding effects η_s are not distinguishable since both always appear together as a sum. This is apparent in Figures 3.1b and 3.1f, which are indistinguishable as data generating mechanisms. Thus, operationally, for $s \in \{a, n\}$, τ_s and η_s can be treated as a single parameter $\delta_s = \tau_s + \eta_s$ representing the direct effect plus Y-confounding.

Sjölander *et al.* (2008) consider a similar setting, but instead identify τ_s under A1, A2, and A4 by specifying $\Pr(Y_i(0) = 1 | S_i = c) - \Pr(Y_i(0) = 1 | S_i = n)$ and $\Pr(Y_i(1) = 1 | S_i = a) - \Pr(Y_i(1) = 1 | S_i = c)$. These sensitivity parameters were not designed to assess sensitivity to A3 (and A4), but they have a one-to-one mapping with $\tau_s = \delta_s$ that

$p_{00} = .088$	$p_{10} = .112$	$p_{01} = .083$	$p_{11} = .069$
$\pi_{00} = .88$	$\pi_{10} = .12$	$\pi_{01} = .69$	$\pi_{11} = .31$

Table 3.1: Observed proportions in the McDonald (1992) study.

can be used to examine CACE sensitivity to both the ER and Y-confounding (but not S-confounding).

3.2.3 Illustration of Confounding

We now illustrate that applying (3.2) in the presence of S-confounding and Y-confounding can result in invalid conclusions. In doing so, we also show that adjusting for unmeasured confounding using (3.3) corrects these problems. For the illustration, we generate data that mimic the observed data from the study of McDonald *et al.* (1992), excluding any covariates, and artificially induce various types of confounding. For now, the response and treatment are left context-free to emphasize the generality of these issues.

Table 3.1 presents observed p_{dt} and π_{dt} for the generated data. Table 3.2 presents the underlying population proportions corresponding to the case of no unmeasured confounding. Here, proportions of always-takers and never-takers, and proportions of $Y^{obs} = 1$ within these two strata, do not change with T. The true $\theta_c = .001 - .117 = -0.116$, and $\bar{\theta}_c^{obs}$ correctly estimates θ_c. However, because only D and not S is observed, many population proportions exist that are consistent with the observed data in Table 3.1. Table 3.3 shows one S-confounding example, where the proportions of principal strata differ across T ($\xi_a = 0.13$ and $\xi_n = -0.09$). Similarly, Table 3.4 shows one Y-confounding example, where the outcome proportions differ across T within the always-taker and never-taker strata ($\delta_a = -0.019$ and $\delta_n = -0.020$).

Regardless of whether the population proportions are in accordance with Table 3.2, 3.3, or 3.4, $\bar{\theta}_c^{obs} = -0.116$, while the true θ_c for Tables 3.2-3.4 are approximately -0.116, -0.053, and 0.023, respectively. Clearly, $\bar{\theta}_c^{obs}$ is biased for θ_c in the cases with S- and Y-confounding. The bias is striking when interpreted as proportion change from baseline: The true θ_c values

	$T=0$	$T=1$	$T=0$	$T=1$
$S=n$	$\pi_{n0}=.69$	$\pi_{n1}=.69$	$p_{n0}=.083$	$p_{n1}=.083$
$S=c$	$\pi_{c0}=.12$	$\pi_{c1}=.12$	$p_{c0}\approx.117$	$p_{c1}\approx.001$
$S=a$	$\pi_{a0}=.19$	$\pi_{a1}=.19$	$p_{a0}=.112$	$p_{a1}=.112$

Table 3.2: Example of population probabilities with no S-confounding and no Y-confounding that are consistent with observed data in McDonald study.

	$T=0$	$T=1$	$T=0$	$T=1$
$S=n$	$\pi_{n0}\approx.56$	$\pi_{n1}=.69$	$p_{n0}=.083$	$p_{n1}=.083$
$S=c$	$\pi_{c0}\approx.25$	$\pi_{c1}\approx.21$	$p_{c0}\approx.099$	$p_{c1}\approx.046$
$S=a$	$\pi_{a0}=.19$	$\pi_{a1}\approx.10$	$p_{a0}=.112$	$p_{a1}=.112$

Table 3.3: Example of population probabilities with S-confounding but no Y-confounding that are consistent with observed data in McDonald study.

represent a 99% reduction, a 54% reduction and a 230% increase in rates.

Using (3.3) with correctly specified values of $\delta_{\{a,n\}}$ and $\xi_{\{a,n\}}$ (and $\eta_c = 0$), i.e., those implied by Tables 3.3-3.4, appropriately adjusts $\bar{\theta}_c^{adj}$ so that it is consistent for θ_c. These results also hold for S-confounding and Y-confounding simultaneously. In practice, of course, δ_s and ξ_s are not known. Instead, analysts can carry out sensitivity analyses by specifying a range of plausible values of these parameters.

3.3 Parametric Sensitivity Analysis

In settings where there are observed covariates imbalances, it will be advantageous to adjust for the effects rather than add them to the sensitivity analysis. This article focusses on regression. Such modeling could be applied on the set of treated records and their matched controls after propensity score matching, as in Hill *et al.* (2004). This strategy helps insure that regression models are not extrapolated over large covariate spaces.

In PS, two models need to be specified, one for the distribution of principal strata S_i and one for marginal distributions of the potential outcomes $Y_i(t)$ given S_i. When both T_i and D_i are binary, a natural model for S_i is the multinomial logit model. Using compliers

	$T=0$	$T=1$	$T=0$	$T=1$
$S=n$	$\pi_{n0}=.69$	$\pi_{n1}=.69$	$p_{n0}\approx.102$	$p_{n1}=.083$
$S=c$	$\pi_{c0}=.12$	$\pi_{c1}=.12$	$p_{c0}\approx.010$	$p_{c1}\approx.033$
$S=a$	$\pi_{a0}=.19$	$\pi_{a1}=.19$	$p_{a0}=.112$	$p_{a1}\approx.092$

Table 3.4: Example of population probabilities with Y-confounding but no S-confounding that are consistent with observed data in McDonald study.

as the reference group,

$$\log\frac{\Pr(S_i=s|T_i,X_i)}{\Pr(S_i=c|T_i,X_i)}=X_i\beta_s^S+T_i\xi_s,\quad s\in\{a,n\},\tag{3.4}$$

where X includes an intercept term and $\Pr(S_i=c|T_i,X_i)=1-\sum_{s\in\{a,n\}}\Pr(S_i=s|T_i,X_i)$. As in Section 2, the parameters ξ_s in (3.4) represent S-confounding. However, they are re-defined to be on multiplicative scales, so that for $s\in\{a,n\}$

$$\exp(\xi_s)=\frac{\Pr(S_i=s|T_i=1,X_i=x)/\Pr(S_i=c|T_i=1,X_i=x)}{\Pr(S_i=s|T_i=0,X_i=x)/\Pr(S_i=c|T_i=0,X_i=x)}.\tag{3.5}$$

For simplicity, ξ_s are assumed to be constant across x. As an example of interpretation, Table 3.3 was created using $\exp(\xi_a)=1/1.5$ and $\exp(\xi_n)=1.5$, meaning that the ratio of never-takers to compliers increases by a factor of 1.5 when going from $T=0$ to $T=1$.

In practice, analysts should select the ranges of ξ_s to be examined based on subject-matter knowledge. For example, they can speculate the extent to which the ratio of never-takers to compliers could differ across T as a result of confounding. This consideration can be separated into smaller parts by focusing on $\Pr(S_i=s|T_i=1,X_i=x)/\Pr(S_i=c|T_i=0,X_i=x)$, for each $s\in\{a,n\}$. Plausible limits for these proportions can be used to reconstruct ranges for ξ_s. Realistic bounds may also be available for

$$\frac{\exp(\xi_a)}{\exp(\xi_n)}=\frac{\Pr(S_i=a|T_i=1,X_i=x)/\Pr(S_i=a|T_i=0,X_i=x)}{\Pr(S_i=n|T_i=1,X_i=x)/\Pr(S_i=n|T_i=0,X_i=x)}.$$

For example, an analyst may decide that the ratio of proportions for always-takers in

$T = 1$ versus $T = 0$ could not be less than half of that ratio for never-takers. Such a statement in conjunction with information about ξ_n would provide plausible values for ξ_a. A complementary approach is to specify bounds for ξ_s using the observed covariate magnitudes. For example, researchers may hypothesize that the magnitude of ξ_n could be up to twice the magnitude of the largest (standardized) estimated β_s^Y.

Binary potential outcomes $Y_i(t)$ can be modeled using a logistic regression, where

$$\begin{aligned} & \operatorname{logit} \Pr(Y_i(t) = 1 | T_i = t, S_i, X_i) \\ = \;& X_i \beta_x^Y + 1_{S_i = c} T_i (\theta_c + \eta_c) + \sum_{s' \in \{a,n\}} 1_{S_i = s'} (\beta_{s'}^Y + T_i \delta_{s'}). \end{aligned} \tag{3.6}$$

Here, $1_{S_i = s'}$ is an indicator function that equals 1 if $S_i = s'$ and equals zero otherwise. The parameters $(\theta_c, \eta_c, \delta_a, \delta_n)$ in (3.7) have similar interpretations as their counterparts in Section 2 but are redefined using log odds ratios. For instance, for $s \in \{a, n\}$

$$\exp(\delta_s) = \frac{\Pr(Y_i(1) = 1 | S_i = s, T_i = 1, X_i = x) / \Pr(Y_i(1) = 0 | S_i = s, T_i = 1, X_i = x)}{\Pr(Y_i(0) = 1 | S_i = s, T_i = 0, X_i = x) / \Pr(Y_i(0) = 0 | S_i = s, T_i = 0, X_i = x)}. \tag{3.7}$$

For simplicity, that the parameters $(\theta_c, \eta_c, \delta_a, \delta_n)$ are assumed to be constant across x. As an example of interpretation, Table 3.4 was created using $\exp(\delta_a) = \exp(\delta_n) = 1/1.25$, meaning the odds that $Y_i(t) = 1$ are 1.25 times greater when $t = 0$ than when $t = 1$ in both the always-takers and never-takers strata. The CACE θ_c is not distinguishable from the Y-confounding effect in the compliers strata; however, it is identifiable after the analyst specifies η_c (and the other sensitivity parameters). For example, in Table 3.3, $\exp(\theta_c + \eta_c) = \exp(-0.053)$, and so θ_c is not identified until η_c is specified.

As with ξ_s, analysts should specify plausible values of $\delta_{\{a,n\}}$ and η_c based on subject-matter knowledge. For η_c, this involves the extent to which the odds for $Y_i(t) = 1$ within the complier strata could change across t as a result of Y-confounding only. For $\delta_{\{a,n\}}$, interpretation involves the extent to which the odds for $Y_i(t) = 1$ within the always-taker and never-taker strata could change across t as a result of direct effects or Y-confounding. Analysts can decompose δ_s into its components from Y-confounding and direct effects. If

the analyst does not suspect direct effects, δ_s could reflect only the effects of confounding.

3.3.1 Estimation of CACE with Sensitivity Parameters

Given T_i and X_i, the analyst can model Y_i^{obs} and D_i with

$$\begin{aligned} & \Pr(Y_i^{obs}, D_i | T_i, X_i) \\ = & \sum_{s \in \mathcal{S}(T_i, D_i)} \Pr(Y_i^{obs} | S_i = s, T_i, X_i) \Pr(S_i = s | T_i, X_i), \end{aligned} \tag{3.8}$$

where $\mathcal{S}(T_i, D_i)$ denotes the set of all possible principal strata that are consistent with T_i and D_i. The two distributions on the right side of (3.8) are specified by (3.7) and (3.4).

When the sensitivity parameters are left to be freely estimated from the data, (3.8) may be maximized in regions of the the space that are not consistent with the true θ_c. Estimating the sensitivity parameters is analogous to choosing which of the Tables 3.2-3.4 is the truth on the basis of largest likelihood. That is, a priori some models have higher likelihoods than others, and when the latent S_i are left to be freely imputed on the basis of the likelihood, the imputation will favor higher likelihood models.

For example, notice that

$$\begin{aligned} L(\cdot|y) &= \prod p_{a1}^{y_i}(1-p_{a1})^{1-y_i} \prod p_{n1}^{y_i}(1-p_{n1})^{1-y_i} \prod p_{c1}^{y_i}(1-p_{c1})^{1-y_i} \\ &\times \prod p_{a0}^{y_i}(1-p_{a0})^{1-y_i} \prod p_{n0}^{y_i}(1-p_{n0})^{1-y_i} \prod p_{c0}^{y_i}(1-p_{c0})^{1-y_i} \end{aligned}$$

has a higher likelihood for certain data generating mechanisms than others. For instance,

$$\begin{array}{c} (0.3)^{30}(0.7)^{70}(0.3)^{30}(0.7)^{70}(0.3)^{30}(0.7)^{70} \\ \times (0.3)^{30}(0.7)^{70}(0.3)^{30}(0.7)^{70}(0.3)^{30}(0.7)^{70} \end{array} < \begin{array}{c} (0.6)^{60}(0.4)^{40}(0.3)^{30}(0.7)^{70}(0)^0(1)^{100} \\ \times (0.3)^{30}(0.7)^{70}(0.6)^{60}(0.4)^{40}(0)^0(1)^{100} \end{array}$$

so that, if we have no S-confounding and observe the data

	$D=0$	$D=1$
$Z=0$	60/140	30/70
$Z=0$	30/70	60/140

$Y^{obs}=1/Y^{obs}=0$ Counts

the data generating mechanism on the right will have a larger likelihood than the one of the left.

Thus, if the sensitivity parameters are left free to be estimated, the model will prefer the data generating mechanism on the right even if the true data generating mechanism is the one on the left. A further example of this general fact will be given in Section 3.3.3.

Because of this lack of identification, the following sensitivity procedure should be used. First, specify ξ_s and δ_s for $S \in \{a, n\}$ to enable estimation of $\theta_c + \eta_c$. Second, specify η_c to identify θ_c. This process can be repeated for the range of plausible values of ξ_s, δ_s, and η_c.

The estimation of $\theta_c + \eta_c$ can proceed using an Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977) or Bayesian data augmentation (Imbens and Rubin, 1997). EM tends to find posterior modes comparatively quickly, whereas full Bayesian inference conveniently provides measures of inferential uncertainty. The EM algorithm alternately replaces the unobserved S_i with their expected values given current draws of the parameters, and maximizes the parameters given the expected values of all S_i. For the Bayesian analysis, after first specifying a prior distribution—we use the added data conjugate prior of Hirano *et al.* (2000) (p. 78)—analysts can sample from the posterior distributions of $\theta = (\beta_a^S, \beta_n^S, \beta_x^Y, \beta_a^Y, \beta_n^Y, \theta_c + \eta_c)$ using Metropolis proposals within a Gibbs sampler. To accomplish this, the posterior distributions of each S_i must be sampled, each instance of which results in new covariate matrices and response vectors in (3.7) and (3.4), respectively. Since a given imputation of S_i may result in a likelihood that is maximized on the boundary of the parameter space (e.g., $\theta_c = -\infty$), the prior plays a key role in stabilizing the sampling. Mixing can be improved by parameterizing (3.7) without an intercept, as in Hirano *et al.* (2000), and by using a Metropolis subchain rather than a single proposal for θ (to better follow the fast mixing S_i). The Gibbs sampler is initialized using the MLEs obtained

from EM.

Prior to model estimation, matching may be used to remove potentially bias-inducing observations from the control group that are unlike those from the treated group. Matching balances covariates across T, not necessarily within S. Correction of further imbalance within principal strata covariate relies on the covariance adjustment implicit in (3.4) and (3.7).

3.3.2 Sensitivity Demonstration Using Introduced Confounding

Using data from the second year (1979-1980) of the study run by McDonald *et al.* (1992), we show that unadjusted model-based PS estimation of the CACE is biased in the presence of S-confounding and Y-confounding, but that analysts can recover the truth using the sensitivity methodology. To do so, we manipulate the data to induce unmeasured confounding in ways we can keep track of, and then examine point estimates computed via EM using the known correct sensitivity parameter specifications. Full Bayesian analysis is pursued in Section 4, assuming A1 and A2. SUTVA is tenuous in disease contexts (Hudgens and Halloran, 2008), this complication is not dealt with further. Monotonicity is plausible in this setting.

The McDonald *et al.* (1992) study is a randomized encouragement design, where $T_i = 1$ if person i is encouraged to take an influenza vaccine and $T_i = 0$ otherwise. The intermediate variable is actual receipt of the vaccine, with $D_i = 1$ if person i indeed takes the vaccine and $D_i = 0$ otherwise. The response, flu outcome, is $Y_i^{obs} = 1$ if person i gets the flu and $Y_i^{obs} = 0$ otherwise.The randomization is done at the level of physician rather than patient, so that the data are actually clustered. This feature feature of the data is ignored for illustration.

The available covariates comprise age in years, sex, race (white/non-white), chronic obstructive pulmonary disease (COPD), heart disease (HD), diabetes, renal disease, and liver disease for 2901 participants. The covariates are closely balanced across encouragement. Regression analyses indicate that higher age and COPD are predictive of taking the vaccine, whereas HD and COPD are predictive of getting the flu. The predictive role of these three

variables closely mirrors that of the first three principal components, which capture age, an approximate COPD/sex/race relationship, and an approximate heart disease/diabetes relationship. The variation captured in the remaining components does not provide any further predictive benefit. Age is transformed to a four level factor (< 40; $(40, 60]$; $(60, 80]$; ≥ 80) based on the observed relationship between age and taking the vaccine. Restriction to complete cases using age, COPD, and HD yields 2893 participants.

Adopting a naive interpretation, these relationships suggest that (1) older populations generally have more always-takers and compliers than comparable younger populations, i.e., the distribution of S_i varies with age; and, (2) HD pervasive populations generally have greater flu prevalence relative to comparable heart healthy populations, i.e., the distribution of Y_i varies with heart disease prevalence. Thus, in a similar but hypothetical observational study, if elderly people are more likely to receive encouragement but age was not controlled for, there would be a higher proportion of compliers and always-takers in the treatment arm, resulting in S-confounding. Likewise, if HD patients are more likely to receive encouragement but HD was not controlled for, there would be a greater proportion of individuals at risk for flu in the same arm, resulting in Y-confounding. The operational distinction between S-confounding and Y-confounding is not clear cut since age and HD have some association. Indeed, the example of COPD directly suggests that S-confounding and Y-confounding may be intimately connected. Nonetheless, separating S-confounding and Y-confounding facilitates sensitivity checks, as will be discussed.

For the demonstration, discretized age and COPD are used in the sub-model for S_i, and COPD and HD in the sub-model for $Y_i(T_i)$. Analysis uses only complete cases, and two more observations (with $T = 0$, $D = 1$, $Y = 1$, age$= 1$, COPD$= 1$, and HD $= 0/1$) are discarded so that no S-confounding and no Y-confounding are plausible descriptions of the remaining data. Without this adjustment, a specification of no S-confounding and no Y-confounding results in an MLE that lies on the boundary of the parameter space, i.e., $\Pr(Y_i(1) = 1 | S_i = c, T_i = 1) = 0$. This data is referred to as the test data.

The truth is taken to be the estimated coefficients in (3.8) for the test data without any

sensitivity adjustments, i.e., all sensitivity parameters equal zero. The MLE for the CACE in the test data is $\hat{\theta}_c = -1.87$.

S-confounding is introduced by removing half of the observed never-takers in the $T_i = 1$ arm and half of the observed always-takers in the $T_i = 0$ arm. Removal was done randomly but ensuring that $\Pr(Y_i(0) = 1|S_i = a, T_i = 0)$ and $\Pr(Y_i(1) = 1|S_i = n, T_i = 1)$ were not changed from the observed probabilities in the test data. This guards against inadvertently inducing Y-confounding, and ensures that the covariate and flu outcome relationships are not changed. Observed covariate balance remains good for age, COPD, and HD after this manipulation. Since half of the never-takers in the $T_i = 1$ arm and half of the always-takers in the $T_i = 0$ arm have been removed, and $\Pr(Y_i^{obs} = 1|S_i, T_i, X_i)$ does not change, the S-confounding sensitivity parameters for this manipulation are $\exp(\xi_a) \approx 2$ and $\exp(\xi_n) \approx 1/2$.

Y-confounding is introduced by keeping only HD $= 1$ individuals in the $T_i = 1$ arm, and not using HD as a covariate so that HD is an unmeasured confounder. After the manipulation, 56.2% of individuals have HD $= 1$ in the $T_i = 0$ arm, and 100% of individuals have HD $= 1$ in the $T_i = 1$ arm. The distributions of age and COPD remain balanced in the treatment arms after the manipulation. This was applied on top of the S-confounding manipulation, but since HD does not associate with D, there is reason to suspect that it will not drastically alter the previously induced S-confounding of $\exp(\xi_n) \approx 1/2$. However, if HD is more prevalent among compliers than always-takers, or vice-versa, then $\exp(\xi_n) \approx 1/2$ will no longer hold. The log odds ratios of the outcomes without covariate adjustment before and after the manipulations were -0.163 and 0.086, respectively, which corresponds to an approximately correct sensitivity parameters for the Y-confounded data of $\delta_a = \eta_a = \delta_n = \eta_n = \eta_c \approx 0.25$.

The model implied by (3.8) is fit to the confounded data with a variety of possible values for the sensitivity parameters. Figure 3.2 displays the results in two panels. The top panel shows contour plots for $\hat{\theta}_c$ across a variety of combinations of $\exp(\xi_a) \in [1/3, \cdots, 3]$ and $\exp(\xi_n) \in [1/3, \cdots, 3]$ with $\eta_c = \delta_a = \delta_n = 0$. The bottom panel shows the contours for the same range for ξ_a and ξ_n with $\eta_c = \delta_a = \delta_n = 0.25$.

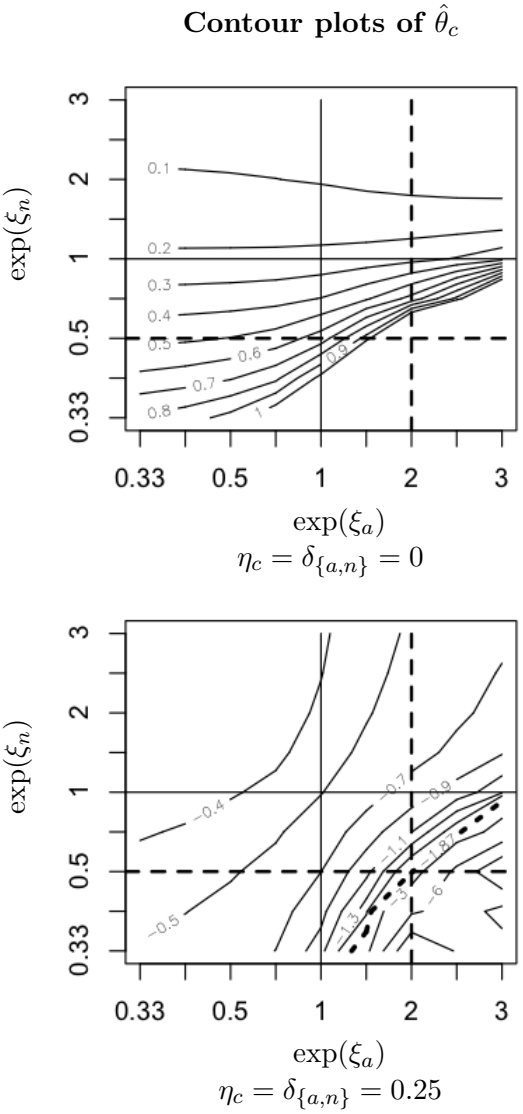

Figure 3.2: Illustrations of sensitivity contour plots for manipulated McDonald data. The top plot shows MLE contours for $\hat{\theta}_c$ across the possible combinations of ξ_s with $\eta_c = \delta_{\{a,n\}} = 0$. The bottom plot shows the same when $\eta_c = \delta_{\{a,n\}} = 0.25$, which are the true values. The dashed cross-hairs are at the approximately correct S-confounding sensitivity parameter values, $\exp(\xi_a) = 2$ and $\exp(\xi_n) = 1/2$. The dashed curve in the bottom plot indicates where $\hat{\theta}_c$ equals θ_c; this curve does not appear in the top plot because it is off the graph. The plots show that standard PS estimates of θ_c are biased in the presence of unmeasured confounding, and it is possible to recover the true θ_c when correct sensitivity parameters are used.

As seen in the left panel of Figure 3.2, fitting PS in the confounded data ignoring unmeasured confounding results in a biased estimate of the CACE. Correctly specifying $\exp(\xi_a) = 2$ and $\exp(\xi_n) = 1/2$, but wrongly setting $\eta_c = \delta_a = \delta_n = 0$ also results in a biased estimate. As evident in the right panel of Figure 3.2, using the approximately correct sensitivity specifications for S-confounding and Y-confounding nearly recovers the CACE estimate. Plots analogous to Figure 3.2 for the β^Y and β^S parameters show similar results. When the sensitivity parameters are estimated by the data rather than be pre-specified, the EM algorithm finds $\hat{\eta}_a = \hat{\eta}_n = .027$, $\hat{\xi}_a = 1.42$ and $\hat{\xi}_n = .80$. These result in $\hat{\theta}_c = .128$ (with $\eta_c = 0$). Estimating when allowing $\hat{\eta}_a \neq \hat{\eta}_n$ resulted in $\hat{\eta}_a = -.32, \hat{\eta}_n = .011, \hat{\xi}_a = 1.77, \hat{\xi}_n = .88$, and $\hat{\theta}_c = .42$. Hence, in all cases, using MLE with free sensitivity parameters results in unreliable estimates.

Figure 3.2 provides a visualization of the topographical nature of potential confounding. The primarily benefit of these plots, however, is to provide a diagnostic alternative to careful bound specification for $\xi_{\{a,n\}}$, η_c, and $\delta_{\{a,n\}}$. Analysts instead can consider large spaces of potential values for the parameters, construct plots like Figure 3.2, and identify the levels of confounding that would alter study conclusions. These values can be interpreted using (3.5) and (3.7), so that analysts can decide if the identified levels are plausible enough to cast doubt on conclusions. This approach is related to the sensitivity checks done by Rosenbaum (2002) in observational study contexts that do not involve PS.

3.4 Applying Sensitivity Analysis in Practice

Using the model specifications of Section 3.2, we now apply the sensitivity methodology to the original complete cases data ($N = 2893$) from McDonald *et al.* (1992). Even though treatment assignment was randomized to physicians, we still check for unmeasured confounding to illustrate the methodology. Additionally, it can be beneficial to perform such checks even in experimental settings, particularly when the number of units in one of the treatment arms is modest. We also consider possible Direct effects, which, as discussed by Hirano *et al.* (2000), might exist because encouragement could affect patients' other health

habits. For simplicity, patient clustering is ignored in the analysis.

When all sensitivity parameters are set to zero (standard PS analysis), the median and 95% credible interval for θ_c are -2.26 and $(-7.29, 2.39)$, so at best there is marginal evidence for a beneficial effect of flu vaccination. Hence, sensitivity checks focus on the types of confounding that would obscure a significant treatment effect. If the interval for θ_c was entirely negative, we would focus on specifications that make θ_c insignificant (or significant in the opposite direction). To implement the analysis, a grid of plausible S-confounding and Y-confounding sensitivity parameters is examined, and the joint model in (3.8) is fit separately at each point in the grid. The grid is determined by using plausible upper and lower limits for the parameters. One could easily examine other ranges if desired.

For S-confounding, ξ_a, ξ_n is set to range between -0.2 and 0.2. From (3.5), this region implies that the relative proportion of always-takers (or never-takers) to compliers could change with T by as much as a factor of $\exp(.2) \approx 1.2$ within each level of X. Further, since ξ_a and ξ_n vary independently over the range, the analysis includes the possibility that the relative proportion of encouraged to unencouraged always-takers could differ from the same relative proportion for never-takers by a factor of roughly $(1.2)(1/1.2) = 1.45$. In the S sub-model of the standard PS model, the largest coefficient is 0.54 and the smallest is -0.37, so that the grid bounds for unmeasured S-confounding are less than the strongest observed effects.

For Y-confounding, the analysis assumes $\eta_a \neq \eta_n \neq \eta_c$, and examines the range -0.2 to 0.2 for each value. This results in sensitivity to unmeasured confounders that may change the outcome odds across T in each strata independently by up to a factor of 1.2 within each level of X.

For the direct effect of treatment on response, the analysis sets $\tau_n = 0$ since, as suggested by (Hirano *et al.*, 2000, p. 82), "if these patients and their physicians did not regard the risk of flu as high enough to warrant inoculation, they might not be subject to other medical considerations either, and so it might be reasonable to assume all these patients were completely unaffected by their physicians' receipt of letter." For τ_a, the analysis

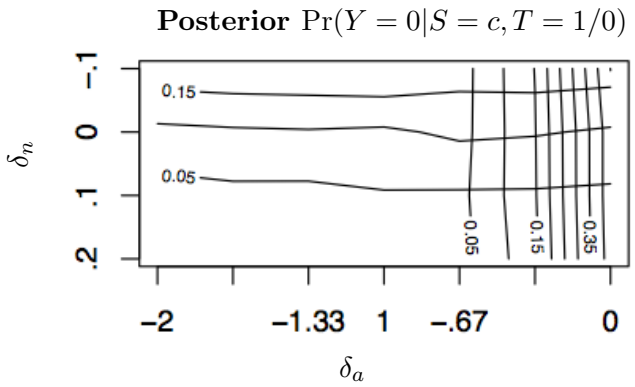

Figure 3.3: Posterior probability of assignment of zero positive flu outcomes to the complier groups in the treatment and control arms from the original McDonald data as a function of δ_a and δ_n, assuming no S-confounding. Vertically oriented lines show probabilities for compliers in the treatment group, and horizontally oriented lines show probabilities for compliers in the control group. These graphs can be used to determine ranges of implausible values of sensitivity parameters, e.g., where the probabilities often equal zero (here, when $\delta_a > 0$ and $\delta_n < -.1$).

assumes that its magnitude does not exceed the strongest effects in the Y sub-model of the standard PS analysis. In this model, the largest coefficient for the Y sub-model is 0.79, and the smallest is -1.83. Thus, combined with the range for η_a, analysis considers $\delta_a \in [-2, 1]$. Equivalently, using (3.7), the analysis investigates scenarios where the odds of $Y(T)$ in the always-taker strata could change by as much as a factor of $\exp(2)$ from $T = 1$ to $T = 0$ within each level of X.

After fitting the models over the grid, we further constrain the sensitivity region to $\delta_a \in [-2, 0]$ and $\delta_n \in [-.1, .2]$. Using $\delta_a > 0$ or $\delta_n < -.1$ results frequent assignment of zero positive flu outcomes to the complier groups in the treatment group and control group, respectively, as displayed in Figure 3.3. If not for the stabilizing prior, this would result in in 95% credible interval lower bounds of $-\infty$ for θ_c and α_0, respectively.

This suggests that the specifications of $\delta_a > 0$ and $\delta_n < -.1$ are not supported by the observed data. In other words, receiving encouragement is unlikely to increase the chance of contracting the flu for people who would take the vaccine regardless, and people who would never take the flu vaccine do not greatly benefit from being encouraged to take the

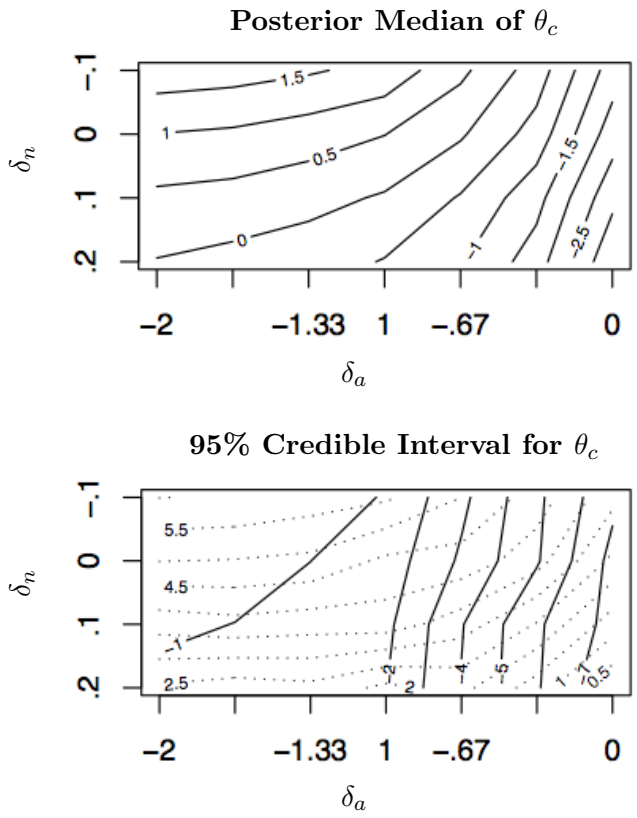

Figure 3.4: Posterior medians and 95% intervals for θ_c for original McDonald data as a function of δ_a and δ_n, assuming no S-confounding. In the bottom panel, the dotted lines represent upper limits and the solid lines represent lower limits of the intervals. The 95% intervals always contain zero, indicating that the conclusions from the standard PS estimation are not overly sensitive to unmeasured confounding in this range of sensitivity paramters.

vaccine.

Using the adjusted bounds, S-confounding and η_c did not have a significant impact on the results. For $\xi_a = \xi_n = 0 = \eta_c$, Figure 3.4 displays the posterior medians and 95% credible intervals for θ_c for the values of δ_a and δ_n in the feasible range. At all points in the sensitivity space, the 95% credible intervals contain zero, suggesting that the results are not overly sensitive to unmeasured confounding or direct effects in the specified plausible range.

Sensitivity examination of only direct effects is embedded in the above analysis by setting

$\xi_a = \xi_n = \eta_c = 0$. As noted already, values of δ_a, and hence direct effects, in the range considered here would not be strong enough to alter the conclusions from the standard PS estimation. Hirano *et al.* (2000) investigated the possibility of direct effects by estimating τ_a and τ_n as parameters of a weakly identified model rather than by evaluating over a grid of pre-set, plausible values.

3.5 Discussion

To implement the sensitivity checks, we recommend first creating plots such as Figure 3.2 with wide bounds for ξ_s and δ_s, and then using these diagnostically to find S- and Y-confounding levels that alter study conclusions. After the fact, the plausibility of the effective S- and Y-confounding regions can be evaluated based on subject specific knowledge. Here, (3.5) and (3.7) can aid in interpretations. Note that the sensitivity approach can be used to examine the amount of confounding that changes the significance of results for either a significant or insignificant CACE from standard PS.

While facilitating rich investigations of the robustness of results to unmeasured confounding, conducting a full sensitivity analysis involves specification of several parameters. Some analysts may choose to reduce the number of free parameters for rougher but faster checks. For example, analysts could set $\delta_a = \delta_n = \eta_c$ to collapse Y-confounding to one parameter. This implies that the unobserved confounders are distributed uniformly across the strata, i.e., that Y-confounding does not vary by S, which may not be strictly true but may be a useful simplification. Future research will evaluate the effectiveness of such simplifications.

While this Chapter focused on the non-compliance setting, the methodology can be analogously applied to other PS contexts. For example, for the outcome low birthweight, intermediate variable hypertension, and treatment smoking—where all variables are binary—the principal strata can be individuals protected against hypertension regardless of smoking (never-takers), individuals who experience hypertension regardless of smoking (always-takers), and individuals susceptible to hypertension on the basis of smoking (compliers).

This methodology also readily extends to continuous outcomes. In the above setting, for instance, birthweight in grams may be considered rather than low birthweight. Further methodological development is required to adapt the methodology to continuous intermediate variables, such as gestational age in days rather than hypertension in the above setting.

Chapter 4

Flexible Bayesian Semi-Parametric PS Modeling for Continuous Intermediate Variables

4.1 Introduction

In causal inference studies, treatment comparisons often need to be adjusted for intermediate variables, i.e., post-treatment variables affected by treatment and also affecting the response. In some randomized trials, for example, intermediate variables are present in the form of non-compliance or partial compliance to assigned treatment, surrogate endpoints, unintended missing outcome data, or truncation by death of primary outcomes. More generally, in both experimental and observational studies, researchers are interested in knowing not only if the treatment is effective, but also to what extent the treatment effect on the outcome is mediated by intermediate variables. That is, concepts of direct and indirect effects are of interest.

It is well documented that directly applying standard methods of pre-treatment variable adjustment, such as regression methods, to intermediate variables can result in estimates that generally lack causal interpretation (e.g., Rosenbaum, 1984; Robins and Greenland,

1992). On the other hand, in the presence of the above mentioned complications, common problems in the current mediation or causal analysis literature are that the causal estimands are not clearly defined, or are defined within the context of the estimation procedure used, and the assumptions needed for a causal interpretation of the estimands are not always made explicit.

This chapter addresses these problems using the potential outcomes approach to causal inference, also known as the Rubin Causal Model (RCM, Rubin, 1974, 1978). In this perspective, a causal inference problem is viewed as a problem of missing data, where the assignment mechanism is explicitly modeled as a process for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects. A commonly invoked identifying assumption is strong ignorability (Rosenbaum and Rubin, 1983b), which usually holds by design in randomized experiments. However, even under such an assumption, inference on causal effects may be invalidated due to the presence of the post-treatment complications mentioned above.

Within the RCM, a relatively recent approach to deal with such complications is principal stratification (PS) (Frangakis and Rubin, 2002). A PS is a cross-classification of subjects into latent classes defined by the joint potential values of the intermediate variable under each of the treatments being compared. Thus, principal strata comprise units having the same values of the intermediate potential outcomes and so are not affected by treatment assignment. This means that comparison of potential outcomes under different treatment levels within a principal stratum, or union of principal strata, are well-defined causal effects that do not suffer from the complications of standard post-treatment-adjusted estimands. These comparisons are called a principal causal effects (PCE) and are the mechanism that PS uses to address post-treatment complications that cannot be ignored for inferring causal effects.

PS is a general framework that can be used to tackle different problems. While PS analyses are mathematically equivalent, they often differ on fundamental issues of PCEs interpretation, specific (union of) principal strata of interest, and potential identifying structural

and modelling assumptions.

Consider the problem of dealing with non-compliance in a simple all or none compliance setting. Principal strata are defined on the joint compliance behavior under treatment and under control. Here, interest often lies in the effect on compliers, i.e., those for whom treatment assignment and treatment receipt coincide, since this is the only stratum where we may learn something about the effect of treatment receipt. Identifying assumptions usually include some plausible form of exclusion restrictions (i.e., ruling out the presence of *direct* effect of assignment) and monotonicity (i.e., no defier principal strata).

Consider instead the problem of disentangling direct and indirect effects. For example, Sjölander *et al.* (2008) assess the effects of physical activity on circulation diseases, such that the effects are not channeled through body mass index (BMI). In this case, PCEs naturally provide information on the extent of the causal effect of the treatment (physical activity) on the primary outcome (disease) that occurs together or separately with a causal effect of the treatment on the intermediate outcome (BMI). Specifically, a principal strata direct effect (PSDE) of the physical activity, after controlling for BMI, exists if there is a causal effect of physical activity on disease outcome for subjects belonging to principal strata where the BMI is not affected by the physical activity. On the other hand, a PSDE does not exist if there is no observed effects of physical activity on the disease outcome for these subjects, i.e., a causal effect of physical activity on disease exists only in the presence of a causal effect of physical activity on BMI. In this context, focus is then on strata where the intermediate variables take on the same values irrespective of the level of the treatment. Identifying assumptions cannot usually include exclusion restrictions, because direct effects are indeed the causal estimands of interest, and cannot thus be ruled out *a priori*.

PS analysis is challenging due to the latent nature of principal strata, i.e., only one potential intermediate outcome is observed for each subject. Thus, identification and estimation strategies generally involve techniques for incomplete data and usually require strong structural or modeling assumptions. Much of the literature discusses settings with binary intermediate variables. If the treatment is also binary, there are at most four prin-

cipal strata. Depending on the setting, simple method of moments estimators for the PCE of those strata are sometimes available under a set of reasonable assumptions (e.g., Angrist *et al.*, 1996). On the other hand, continuous intermediate outcomes lead to an infinite number of possible principal strata in theory, which introduces substantial complications in both inference and interpretation. Few papers have dealt with many (e.g., Frangakis *et al.*, 2004) or continuous principal strata (e.g., Jin and Rubin, 2008), and one common method is to dichotomize continuous intermediate variables (e.g., Sjölander *et al.*, 2008). However, dichotomization is often subject to information loss that might miss important underlying structure, and the results can be sensitive to cutoff points. An improved approach is to specify fully parametric continuous models for principal strata (e.g., Jin and Rubin, 2008). However, restricting models to a single parametric family is inadequate for complex data distributions involving, for example, outliers, skewness and multi-modality common in real applications. This is a serious concern because the current literature shows that the modeled association between intermediate potential outcomes plays a crucial role in inference since it implicitly defines the latent structure that drives PS analysis.

To address this shortcoming, this thesis proposes a Bayesian nonparametric model for continuous intermediate variables based on DP mixture models (DPM) (Escobar and West, 1995b; Müller and Quintana, 2004). The resulting principal strata model has support over the space of all mixed continuous and discrete distributions, and thus greatly mitigates concerns of model inflexibility and inappropriateness, as well as concerns on the sensitivity of inference relative to specification of the joint distribution of intermediate outcomes. In addition, because DP methodologies exhibit clustering properties, they are a particularly appropriate choice for the PS setting for two reasons. First, clustering encourages information sharing. In the PS setting with so much unobserved data requiring imputation, sharing information across subjects with similar characteristics can be desirable. Second, PS can be viewed as a latent class model, and clustering provides opportunities to model, explore, and potentially interpret the latent structure of data.

The remainder of the article is organized as follows. In Section 2, we introduce the

general Bayesian semi-parametric model and develop a procedure for posterior inference for the model parameters. In Section 3, we illustrate the method by estimating the treatment effect of a drug in the randomized clinical trial with partial compliance presented by Efron and Feldman (1991), and compare the results with those from previous studies. In section 4, the method is applied to an observational study, the Swedish National March Cohort (NMC), to investigate the effect of physical activity on coronary heart disease as it relates to BMI. Section 5 concludes with a discussion.

4.2 Models and Computation

4.2.1 Overview of Principal Stratification

PS was first introduced by Frangakis and Rubin (2002) to address post-treatment complications within the RCM. In the RCM, for a study with a binary treatment T and post-treatment variable Y, each unit i ($i = 1, \ldots, N$) is assigned to either treatment ($T_i^{obs} = 1$) or control ($T_i^{obs} = 0$), but has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, representing the hypothetical outcomes for each i under simultaneous assignment of treatment and control. The potential outcome to be observed, $Y_i^{obs} = Y_i(T_i^{obs})$, depends on the realized value of T_i. The remaining potential outcome, $Y_i^{mis} = Y_i(1 - T_i^{obs})$, is unobserved. The causal effect of T on Y for unit i is defined as a comparison between $Y_i(1)$ and $Y_i(0)$, e.g., $Y_i(1) - Y_i(0)$. The fact that only two potential outcomes for each unit exist reflects the acceptance stable unit treatment value assumption, (Rubin, 1996), i.e., that there is no interference between units and that both levels of the treatment define a single outcome for each unit. This assumption is maintained throughout this work.

Intermediate variables D cannot be avoided in many settings. For example, in a noncompliance setting, D is an indicator for treatment receipt; in a partial compliance setting, D is a continuous variable indicating the extent of compliance; in a truncation by death setting, D is an indicator of outcome missingness or outcome censoring; and, in mediation analysis, D is a variable lying on the causal pathway. Intermediate variables are post-treatment

variables, and so also have potential outcomes.

Throughout this chapter, the strong ignorability assumption (Rosenbaum and Rubin, 1983b) on the assignment mechanism is maintained. This asserts that treatment assignment is unconfounded given a vector $\boldsymbol{X}$ of observed pre-treatment covariates and that in infinite samples treated and controls can be compared for all values of $\boldsymbol{X}$. Formally, this is

$$Y(0), Y(1), D(0), D(1) \perp\!\!\!\perp T|\boldsymbol{X}, \quad \text{and} \quad 0 < P(T = 1|\boldsymbol{X}) < 1.$$

This assumption is often referred to as no unmeasured confounding of assignment. If true, it guarantees that the comparison of treated and control units with the same value of $\boldsymbol{X}$ leads to valid inference on causal effects. However, it is in general improper to condition on D_i^{obs}, i.e., compare $\{Y_i^{obs}\colon D_i^{obs} = D_0, T_i^{obs} = 1, \boldsymbol{X}_i\}$ and $\{Y_j^{obs}\colon D_j^{obs} = D_0, T_i^{obs} = 0, \boldsymbol{X}_j\}$, because $\{i\colon D_i^{obs} = D_0, T_i^{obs} = 1\}$ and $\{j\colon D_j^{obs} = D_0, T_i^{obs} = 0\}$ are not exchangable if T affects D. Instead, PS compares $\{Y_i(1) : S_i = (D_0, D_1)\}$ and $\{Y_i(0) : S_i = (D_0, D_1)\}$, where $S_i = (D_i(0), D_i(1))$ is called a (basic) principal stratum. More generally, a PS with respect to D is a partition of the units into sets that are unions of the basic principal strata.

By strong ignorability, $Y(0), Y(1) \perp\!\!\!\perp T|(D(0), D(1), \boldsymbol{X})$, so comparisons within principal stratum, called principal causal effects (PCE), are well-defined causal effects. Specifically, the PCE is

$$PCE(D_0, D_1) \quad = \quad E(Y_i(1) - Y_i(0)|S_i = (D_0, D_1)) \tag{4.1}$$

Intuitively, S_i may be regarded as a pre-treatment covariate since it is invariant under different treatment assignment. PCEs are the primary focus of this chapter.

Depending on the settings and the causal questions, some PCEs can be more interesting and relevant than others. For example, if D is a censoring indicator, e.g., an indicator of death, and Y is an outcome defined only on non censored observations, e.g., quality of life, then the principal stratum of the always survivors (Rubin, 2006), $\{i : D_i(0) = D_i(1) = 0\}$, is the only group where the effect on Y is well defined. As a consequence, different settings

require different structural and modeling assumptions for identification of PCEs. Those usually include assumptions that allow one to reduce the number of strata, e.g., monotonicity with respect to some post-treatment variables, exclusion restrictions on the intermediate variables, dichotomization or other coarsening of continuous/categorical intermediate variables; restrictions on potential outcome distributions, e.g., exclusion restrictions on the outcome or stochastic dominance; restrictions based on covariates, e.g., effects assumed to be constant across values of some covariates; restrictions on the specification of the outcome distribution, e.g., assuming specific parametric distribution for the outcome and incorporating some other restrictions; and strong ignorability to insure that the distribution of S_i has the same distribution in both treatment arms within cells defined by pre-treatment variables.

When some of these identification assumptions are relaxed, there is usually a lack of nonparametric point identification, so that no unique moment estimate of PCEs exist. Sometimes, even adding parametric assumptions is only enough to weakly identify PCEs (Imbens and Rubin, 1997; Hirano *et al.*, 2000), i.e., the likelihood function can be flat around its maximum. Alternatively, posterior PCE distributions are usually well defined and allow improved investigation in these weakly identified models. This is a possible reason to adopt a fully Bayesian approach.

In the PS framework, the quantities for each subject i that are relevant to estimating PCEs are $(Y_i(0), Y_i(1), D_i(0), D_i(1), T_i, \boldsymbol{X}_i)$. Under SUTVA and strong ignorability, they may be regarded as a joint realization from a general model,

$$\begin{aligned} & \Pr((Y(0), Y(1)), (D(0), D(1)), T, \boldsymbol{X}) \\ = & \prod_i \Pr(Y_i(0), Y_i(1) | S_i, \boldsymbol{X}_i) \Pr(S_i | \boldsymbol{X}_i) \Pr(T_i | \boldsymbol{X}_i) \Pr(\boldsymbol{X}_i). \end{aligned}$$

Since, $\Pr(T|\boldsymbol{X}_i)$ and $\Pr(\boldsymbol{X}_i)$ do not contain information on potential outcomes, this factorization suggests that model-based PS inference usually involves two sets of models: One for the marginal distribution of potential outcomes $Y(0), Y(1)$ conditional on the principal

strata S and covariates (hereafter referred to as the Y-model), and one for the distribution of principal strata conditional on the covariates (hereafter referred as the S-model). The Y-model provides a structure to recover the relationship between Y_i^{obs} and D_i^{mis} given the observed D_i^{obs} and $\boldsymbol{X}_i$, and the S-model directly specifies the relationship between D_i^{mis} and the observed D_i^{obs} and $\boldsymbol{X}_i$. Together, these inform each D_i^{mis}, and therefore each S_i, which in turn informs the PCEs.

The S-model forms the basis of PCE analysis, but only its marginal distributions $D(0)$ and $D(1)$ are observed. With continuous D, correct specification of the joint structure becomes important, but can only be checked after the fact on the basis of observed imputations that depend on the assumed specifications to start with. Flexible S-model specifications are desirable so that model forms do not adversely influence PCE estimation by inappropriately restricting the estimation of S_i.

4.2.2 Bayesian semi-parametric model

We propose a Bayesian semi-parametric model that consists of a parametric Y-model and a Bayesian nonparametric S-model based on Dirichlet process (DP). While it can be also useful to assume flexible Y-models, this chapter uses fully parametric outcome models in order to focus on the S-model, where the benefit of Bayesian nonparametric models is more pronounced.

Parametric Y-model. Since $Y_i(0)$ and $Y_i(1)$ are not both observed for the same subject, it is typical to assume they are marginally independent and model them separately as $\Pr(Y_i(t)|S_i, \boldsymbol{X}_i; \lambda_t^Y)$ for $t = 0, 1$, where λ_t^Y are the corresponding parameters. Parametric model assumptions depend on specific application and are generally guided by the observed marginal distribution of Y given D and X in each assignment arm. Section 3 and 4 give examples of this. Y-model specification is crucial for identifying the unobserved correlation between $D_i(0)$ and $D_i(1)$ in the S-model, because this information is implicitly embedded in the Y-model where $D_i(0)$ and $D_i(1)$ appear together. Thus, careful specification of Y-model underlies any reasonable analysis.

Sensitivity of results to the independence assumption may be checked by joining the two potential outcomes $Y_i(0)$ and $Y_i(1)$ with a correlation parameter ρ (Jin and Rubin, 2008). In fact, it is known that ρ generally only affects precision, and that inference, whether Bayesian or frequentist, does not depend on ρ in large samples.

Bayesian nonparametric model for principal strata. Even though never jointly observed, the potential intermediate outcomes need to be modeled jointly to be compatible with the underlying correlation imposed implicitly by the Y-model. Flexible S-models with strong structuring features are desirable here to capture subtle information from the possibly complex distribution of S_i. A Dirichlet process mixture (DPM) provides such an S-model,

$$\Pr(S_i|\boldsymbol{X}_i;\boldsymbol{\beta}^D,\boldsymbol{\theta}) = \int K(D_i(0),D_i(1)|\boldsymbol{X}_i;\boldsymbol{\beta}^D,\theta)dG(\theta), \quad \text{with} \quad G\sim \mathrm{DP}(\alpha G_0), \qquad (4.2)$$

where the kernel $K(D_i(0),D_i(1)|\boldsymbol{X},\theta)$ is a bivariate distribution (described later) and the probability measure G is generated from a Dirichlet process (DP), $\mathrm{DP}(\alpha G_0)$, with strength parameter α and base measure G_0 (Ferguson, 1974).

A random probability measure G is sampled from $\mathrm{DP}(\alpha G_0)$ if for any Borel set partition of the space A, $\{A_1,\cdots A_k\}$, where G_0 (and G) are defined, the distribution of the realized probabilities follows a Dirichlet distribution,

$$\{\Pr_G(A_1),\cdots,\Pr_G(A_k)\}\sim \mathrm{Dir}(\alpha\Pr_{G_0}(A_1),\ldots,\alpha\Pr_{G_0}(A_k)).$$

Large values for the scalar strength parameter α imply less variation of a realized distribution G from the base measure G_0, where $E(G)=G_0$ in the sense that $E(G(A_1))=G_0(A_1)$ for any Borel set A_1 in A.

The stick-breaking (SB) representation of the DP (Sethurman, 1994) shows that $G\sim \mathrm{DP}(\alpha G_0)$ may be constructed as

$$\mathrm{G}(\cdot)=\sum_{h=1}^{\infty} w_h\delta_{\theta_h}(\cdot), \quad \theta_h\overset{iid}{\sim} G_0, \quad w_h=w'_h\prod_{k<h}(1-w'_k),\ w'_h\overset{iid}{\sim}\mathrm{Be}(1,\alpha), \qquad (4.3)$$

where θ_h are called atoms and w_h are probabilities that sum up to 1. The stick breaking nature of the DP encourages decreasing weights, $w_i > w_j$ for $i > j$, *a priori* since $E[w_h] = 1/(1+\alpha) * (\alpha/(1+\alpha))^{h-1}$. Small α corresponds to sparser models, i.e., models with fewer nontrivial weights, that provide a coarser approximation to G_0.

The SB representation shows that samples from a DP are discrete distributions, so the DP cannot be directly used as a prior distribution for continuous data models. However, the discrete atoms and associated weights may be used to define an infinite mixture of continuous distributions, as in (4.2). In the setting of continuous joint potential intermediate variables, a convenient choice for the kernel of the mixture, $K(D_i(0), D_i(1)|\boldsymbol{X}; \boldsymbol{\beta}^D, \theta)$ in (4.2) is a (truncated) bivariate Gaussian distribution with linear effect from covariates $\boldsymbol{X}$,

$$K(D_i(0), D_i(1)|\boldsymbol{X}_i; \boldsymbol{\beta}^D, \theta) \propto \mathrm{N}((\eta_0 + \boldsymbol{X}_i\boldsymbol{\beta}_0^D, \eta_1 + \boldsymbol{X}_i\boldsymbol{\beta}_1^D)', \Sigma)1_A,$$

where $\theta = (\eta_0, \eta_1, \Sigma)$, and A is the support of $(D_i(0), D_i(1))$ in the (possibly truncated) distribution (e.g., R^2, or $[0,1]\times[0,1]$). Using the SB representation (4.3), (4.2) is equivalent to

$$\Pr(S_i|\boldsymbol{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\theta}) = \sum_{h=1}^{\infty} w_h c_h \mathrm{N}((\eta_{0h} + \boldsymbol{X}_i\boldsymbol{\beta}_0^D, \eta_{1h} + \boldsymbol{X}_i\boldsymbol{\beta}_1^D)', \Sigma_h)1_A, \tag{4.4}$$

where the atoms $(\theta_h = (\eta_{0h}, \eta_{1h}, \Sigma_h))$ and associated weights (mixture probabilities w_h) are nonparametrically specified via $\mathrm{DP}(\alpha G_0)$, and c_h is the normalizing constant resulting from the truncation to support space A. The coefficients $\boldsymbol{\beta}^D$ are assumed to be common across mixture components, but this may be relaxed. This specification results in a flexible nonparametric mixture structure for the distribution of the principal strata that has support on a very large space of continuous bivariate distributions defined on A.

In addition to flexibility, a natural byproduct of the mixture structure of the DPM is clustering, which is appealing in the PS context. Clustering allows information to be shared locally between the S_i in the same cluster, whether they are unobserved or not. Increased local information sharing is encouraged because the DPM naturally promotes sparse clustering. *A posteriori*, a given w_h will be nearly 0 unless the inclusion of the additional

component is strongly suggested by the data. Parsimonious clustering also provides opportunities for meaningful interpretation of the principal strata. Principal strata are essentially latent classes of subjects, and the DPM non-parametrically allocates similar subjects into the same clusters. As will be elaborated in the applications, this automatic latent structure recovery may be treated as the continuous analogue to discrete PS analysis.

4.2.3 Posterior inference

Bayesian PS inference utilizes the complete-data likelihood

$$\begin{aligned} & \Pr(Y_i(0), Y_i(1), S_i | \boldsymbol{X}_i; \lambda_0^Y, \lambda_1^Y, \boldsymbol{\beta}^D, \boldsymbol{\theta}) \\ = & \prod_{i=1}^{n} \Pr{}_0\left(Y_i(0) | S_i, \boldsymbol{X}_i; \lambda_0^Y\right)^{(1-T_i^{obs})} \Pr{}_1\left(Y_i(1) | S_i, \boldsymbol{X}_i; \lambda_1^Y\right)^{T_i^{obs}} \Pr\left(S_i | \boldsymbol{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\theta}\right), \end{aligned}$$

where $\Pr\left(S_i | \boldsymbol{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\theta}\right)$ is the DPM (4.4). The Bayesian model is completed by specifying the DPM and assuming prior distributions for the parameters. The choice of G_0 suggests the support of θ to be explored, and generally depends on the range of the data being modeled. Convenient choices for G_0 are inverse Wishart $\mathrm{IW}(q, \Sigma_0)$ for Σ_h, and normal $\mathrm{N}(m, v^2)$ or uniform $\mathrm{Unif}(A)$ for (η_{0h}, η_{1h}). Inference on the level of sparsity of the DP demanded by the data is available by specifying a prior distribution for α. A standard choice for α is a Gamma distribution $\mathrm{Ga}(a, b)$ with hyperparameters a, b. A standard prior distribution for the coefficients $\boldsymbol{\beta}^D$ is a diffuse normal distribution $\mathrm{N}(0, s^2 I)$. Prior distributions for λ_t^Y depend on the nature of the parameter, but for many common Y-models conjugate choices are available.

All PCEs are functions of the parameters of the complete-data likelihood, so full Bayesian inference for PCEs is based on the posterior distributions the model parameters conditional on the observed data. However, because $S_i = (D_i(0), D_i(0))$ is not fully observed, the posterior distribution, $\Pr(\boldsymbol{\beta} | \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs})$, is proportional to

$$\Pr(\lambda_0^Y, \lambda_1^Y, \boldsymbol{\beta}, \boldsymbol{\theta}^D) \int \Pr(Y_i(0), Y_i(1), S_i | \boldsymbol{X}_i, \boldsymbol{\lambda}^Y, \boldsymbol{\beta}^D, \boldsymbol{\theta}) d\boldsymbol{D}^{mis},$$

which is not directly tractable for most specifications. However, because both $\Pr(\boldsymbol{\beta}|\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{D}^{mis})$ and $\Pr(\boldsymbol{D}^{mis}|\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{\beta})$ are generally tractable, the joint posterior distribution $\Pr(\boldsymbol{\beta}, \boldsymbol{D}^{mis}|\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs})$ can be simulated using a data augmentation approach for $\boldsymbol{D}^{mis}$ (Tanner and Wong, 1987). Inference for the joint posterior distribution then provides inference for the marginal posterior distribution $\Pr(\boldsymbol{\beta}|\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs})$. Imbens and Rubin (1997) and Jin and Rubin (2008) provide further discussion on general Bayesian inference in PS.

The implementation of the DPM (4.4) used here first selects the number of mixture components $H < \infty$ used to approximate the SB representation of the DPM (Ishwaran and James, 2001). Then, latent class indicators with a multinomial distribution,

$$Z_i \sim \text{MN}(\boldsymbol{w}),\ Z_i \in \{1, \cdots, H\},$$

are introduced to associate each observation i with a cluster h of the DPM. The marginal distribution implied by integrating out $\boldsymbol{Z}$ is the original approximation (based on H) to (4.4), so this augmentation expands the parameter space but does not change the original model specification. It does, however, greatly simplify (4.4), so that for each individual i, conditional on $Z_i = h$,

$$\Pr(S_i|X_i, Z_i = h; \boldsymbol{\beta}^D, \boldsymbol{\theta}) = c_h \text{N}((\eta_{0h} + \boldsymbol{X}\boldsymbol{\beta}_0^D, \eta_{1h} + \boldsymbol{X}\boldsymbol{\beta}_1^D)', \Sigma_h) 1_A.$$

This approximation is justified through the sparsity property of the DPM, which effectively provides an automatic selection mechanism for the number of active components $H^* < \infty$ in the SB representation, i.e., the number of nontrivial w_h. Thus, when the sample size is fixed, and only a small number of w_h are nonzero, the nonparametric behavior of the DPM may be approximated with a finite mixture model that truncates the SB representation at some large $H^* < H$ (Ishwaran and James, 2001). This approach gives satisfying results in our applications.

Using the DPM approximation, the complete-data model may be estimated using a data augmentation Gibbs sampler approach with the following steps.

1. Given $\lambda_0^Y, \lambda_1^Y, \boldsymbol{\beta}^D, \boldsymbol{\theta}$, and $\boldsymbol{Z}$, draw each D_i^{mis} from

$$\Pr(D_i^{mis}|-) \quad \propto \quad \Pr{}_{T_i}\left(Y_i^{obs}|S_i, \boldsymbol{X}_i; \lambda_{T_i}^Y\right) \Pr(S_i|\boldsymbol{X}_i, \boldsymbol{\beta}^D, \theta_{Z_i}).$$

2. Given $\boldsymbol{\beta}^D, \boldsymbol{w}$, and $\boldsymbol{S}$, draw each Z_i from a multinomial distribution with

$$\Pr(Z_i = h|-) \propto w_h \Pr\left(S_i|\boldsymbol{X}_i, Z_i = h; \boldsymbol{\beta}^D, \theta_h\right).$$

3. Given $\boldsymbol{Z}$, set $w'_H = 1$, and for each $h \in \{1, \cdots, H-1\}$ draw w'_h from

$$\Pr(w'_h|-) = \text{Be}\left(1 + \sum_{i:Z_i=h} 1, \alpha + \sum_{i:Z_i>h} 1\right),$$

   and update $w_h = w'_h \prod_{k<h} w'_k$.

4. Given $\boldsymbol{Z}$, draw α from

$$\Pr(\alpha|-) \propto \Pr(\alpha) \prod_{h=1}^{H} \text{Be}\left(1 + \sum_{i:Z_i=h} 1, \alpha + \sum_{i:Z_i>h} 1\right).$$

5. Given $\boldsymbol{\beta}^D, \boldsymbol{Z}$, and $\boldsymbol{S}$, draw each θ_h from

$$\Pr(\theta_h|-) \propto G_0(\theta_h) \prod_{i:Z_i=h} \Pr\left(S_i|\boldsymbol{X}_i, Z_i; \boldsymbol{\beta}, \theta_h\right).$$

6. Given $\boldsymbol{\theta}, \boldsymbol{Z}$, and $\boldsymbol{S}$, draw $\boldsymbol{\beta}^D$ from

$$\Pr(\boldsymbol{\beta}^D|-) \propto \Pr(\boldsymbol{\beta}^D) \prod_{i=1}^{n} \Pr\left(S_i|\boldsymbol{X}_i, Z_i; \boldsymbol{\beta}^D, \theta_{Z_i}\right).$$

7. Given $\boldsymbol{S}$, draw each λ_t^Y from:

$$\Pr(\boldsymbol{\beta}_t^Y|-) \propto \Pr(\lambda_t^Y) \prod_{i:T_i=t} \Pr_t\left(Y_i^{obs}|S_i, \boldsymbol{X}_i; \lambda_t^Y\right).$$

Cycling through these steps provides correlated draws whose stationary distribution upon convergence is the joint posterior distribution of the parameters. Posterior distributions of a function of the parameters can be obtained by sequentially transforming the posterior parameter samples according to the desired functional. The stationary distribution of the resulting sample is the posterior distribution of the functional. Since $E[Y(1)|X_i, S_i] = f(X_i, S_i; \lambda_0^Y)$ and $E[Y(0)|X_i, S_i] = f(X_i, S_i; \lambda_0^Y)$, samples from the posterior distribution of the PCE for a given X_i and S_i are obtained by taking the difference of these values at each posterior sample of λ_0^Y and λ_1^Y. Point and interval estimation is done using quantiles of posterior samples.

4.3 Application to randomized trial with partial compliance

4.3.1 Data and Models

To compare with the existing approaches, in this section we apply the proposed Bayesian semi-parametric model to a frequently studied data set from the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) in Efron and Feldman (1991), hereafter EF. The LRC-CPPT was a placebo-controlled double-blind randomized clinical trial that assigned 164 men to receive cholestyramine and 171 men to receive a placebo, and sought to examine the effect of cholestyramine in lowering cholesterol. Compliance with treatment assignment was not enforced, but was monitored, and percent of compliance over the approximately 7 years of the study was reported. The cholesterol level of each subject was recorded before and after the study, and the outcome was the decrease in cholesterol level over the course of the study. No covariate information is available.

An interesting question arising from this experiment concerns the causal effect of different doses of cholestyramine. However, because cholestyramine has side-effects, the compliance distributions are different in the drug and placebo arms. Compliance to drug and compliance to placebo are different subject characteristics. Thus, comparisons between experimental arms within a given level of observed compliance to drug and placebo do

not define proper causal effects and so do not capture drug efficacy or any dose-response relationship.

Jin and Rubin (2008), hereafter JR, re-analyzed the data, and present two analyses. The first is a PS analysis that estimates PCE(D_0, D_1) for every combination of placebo and treatment compliance using a fully parametric Bayesian approach. The second is a dose-response analysis which imposes additional assumptions to estimate dose-response curves. Here, we limit comparison to their first PS analysis, leaving the use of our approach to the estimation of dose-response relationships to future research.

Let $D_i(0)$ denote the percentage of placebo taken by subject i when assigned to control and $D_i(1)$ denote the percentage of active treatment taken by subject i when assigned to treatment. It is worth noting the slight abuse of notation here: Variable D does not have the same meaning under treatment and under control. For this reason, JR use a different notation for the two compliance variables. Also because of this, standard exclusion restriction assumptions that would rule out any effect of assignment on individuals with $D_i(0) = D_(1)$ are not plausible. The only exclusion restriction assumption that may be plausible is $Y_i(0) = Y_i(1)$ where $\{i : D_i(0) = D_i(1) = 0\}$, i.e., for individuals who take no placebo and no active treatment, assignment to treatment should not have any effect on the outcome. JR make this assumption, as well as a side-effect monotonicity assumption, $D_i(1) < D_i(0)$, which appears to necessary to define their parametric S-model.

Bartolucci and Grilli (2010), hereafter BG, conducted a PS analysis on the same data, proposing a more flexible S-model based on a Plackett copula that does not require a side-effect assumption and models the marginal distributions of the intermediate variables nonparametrically. They provide a likelihood based method to estimate the model, and compare several alternative outcome models using likelihood ratio tests. They suggest the following Y-model, which is adopted here:

$$Y_i(t)|D_i(0), D_i(1) \sim \mathrm{N}[\mu_t(D_i(0), D_i(1)), \exp(\sigma_t^2(D_i(0), D_i(1)))], \quad t = 0, 1, \tag{4.5}$$

with the mean and variance depending on $D_i(0)$ and $D_i(1)$ as follows,

$$\begin{aligned}
\mu_0(D_i(0), D_i(1)) &= \beta_0^Y + \beta_1^Y D_i(0); \\
\mu_1(D_i(0), D_i(1)) &= \beta_0^Y + \beta_1^Y D_i(0) + \beta_2^Y D_i(1) + \beta_3^Y D_i(0) D_i(1); \\
\sigma_0^2(D_i(0), D_i(1)) &= \gamma_0; \\
\sigma_1^2(D_i(0), D_i(1)) &= \gamma_0 + \gamma_1 D_i(1).
\end{aligned}$$

For a subject that would not take any placebo or drug under either assignment, it is reasonable to assume, as JR do, his/her cholesterol level under either assignment, i.e. the intercepts β_0^Y in the mean functions μ_1 and μ_0, are the same. Model specification has also been suggested by the scatter plots of observed Y and D in the two random halves of the experiment. They show a linear relationship between Y and D in control group and a quadratic relationship in treatment group, which lead to specifications of the Y-model as detailed above. Coefficient β_1^Y represents how the baseline mean outcome varies with the personal characteristic of compliance to placebo, and is also assumed to be constant under both assignment. The key model assumption lies in the mean function of the potential outcome $Y_i(1)$, μ_1, where both $D_i(1)$ and $D_i(0)$ enter as regressors. Because the association between $D_i(1)$ and $D_i(0)$ cannot be directly estimated from the data, as they are never observed jointly, indirect evidence on their association is provided by the regression of $Y(1)$.

The prior distributions for the parameters are assumed to be,

$$\Pr(\gamma_0) = \mathrm{N}(3, 1), \quad \Pr(\gamma_1) = \mathrm{N}(0, 1.5^2), \quad \Pr(\boldsymbol{\beta}^Y) = \mathrm{N}(\mathbf{0}, 10^2 I_4),$$

where $\boldsymbol{\beta}^Y = (\beta_0^Y, \beta_1^Y, \beta_2^Y, \beta_3^Y)$ and I_4 is a 4-dimensional identity matrix. The specifications for γ_0 and γ_1 are vague, but are based on observed estimates of the variance in the two treatment arms.

The DPM model (4.4) without covariates is assumed for S-model, with

$$A \equiv [0,1] \times [0,1] \quad G_0 = \mathrm{N}((.5,.5)', .25^2 I_2)\mathrm{IW}(2, I_2) \quad \Pr(\alpha) = \mathrm{Ga}(1,1)$$

4.3.2 Results

The parametric Y-model (4.5) and Bayesian nonparametric S-model (4.4) were fitted using the LRC-CPPT. Five parallel MCMC chains of 205,000 iterations with the first 5,000 as burn-in period were run, each having different starting values. None of the chains showed signs of adverse mixing and all chains lead to highly similar posterior summary statistics. Sensitivity of results to alternative hyperparameter specifications for α within the Gamma prior distribution for α is minimal and the approximation truncation level $H = 10$ appears to be adequate for DP approximation in this analysis.

Table 4.1 provides the posterior medians and 95% credible intervals for the coefficients in the Y-model (4.5), and the corresponding MLE and standard errors under the copula-likelihood approach of BG. The point estimates of $\beta_0^Y, \beta_1^Y, \gamma_0, \gamma_1$ are similar between the two methods, as their estimation is based primarily on observed data. The DPM provides slightly tighter intervals than the copula does. However, there is a large discrepancy in the point estimates of β_2^Y and β_3^Y. The length of the corresponding interval estimates based on the DPM are roughly half the length of those produced by the copula. The sum of β_2^Y and β_3^Y are comparable between methods, however. The term, $\beta_2^Y D_i(1) + \beta_3^Y D_i(0) D_i(1)$, in the μ_1 function is approximately equal to $\beta_2^Y + \beta_3^Y$ when $D_i(1) = D_i(0) = 1$. Since the majority of the subjects have high compliance (close to 1) under both assignments (as shown in Figure 4.1), this suggests that the marginal distribution of $Y_i(1)$ are similarly estimated by the DPM and copula methods, but the DPM more clearly delineates the effects of the linear term $D_i(1)$ and the interaction term $D_i(0)D_i(1)$ compared to the copula.

To further understand the improved precision in the Y-model, it is useful to look at the results of the DPM S-model. As shown in the scatterplot of a representative posterior draw of the principal strata $(D_i(0), D_i(1))$ in Figure 4.1, there are three predominant clusters:

	DPM		Copula	
coefficient	post. median	95% cred. ints.	MLE	SE
β_0^Y	-0.71	(-5.19, 3.74)	-2.69	2.98
β_1^Y	11.87	(5.95, 17.74)	11.24	3.48
β_2^Y	22.30	(9.36, 35.17)	-21.88	21.33
β_3^Y	23.02	(8.40, 37.65)	73.46	25.24
γ_0	5.28	(5.09, 5.48)	5.26	-
γ_1	1.35	(0.96, 1.74)	1.16	0.16

Table 4.1: Coefficients in the outcome Y-model (4.5) as estimated using the Bayesian DPM S-model, where posterior medians and 95% credible intervals are shown; and the frequentist copula S-model of BG, where MLE and SEs are shown.

the largest cluster in the upper right corner (45%), a second largest cluster in the right middle (30%), and the smallest in the lower left corner (25%). Interestingly, these latent clusters roughly correspond to the principal strata of always-takers, never-takers, and defiers in the standard binary PS classification (Angrist *et al.*, 1996), but with slight different interpretation since $D(0)$ is placebo and $D(1)$ is drug. Here, the defiers are those subjects who experienced negative side-effects, as discussed by both JR and BG. The same cluster structure was consistantly observed in the posterior draws, and the majority of imputed D_i^{mis} maintained a single cluster membership. Thus, for the LRC-CPPT data, there was strong evidence of relevant latent structure recovery, and this information was used to inform D_i^{mis} locally on the basis of cluster membership. The reduced variability in estimating the unobserved D_i^{mis} lead to more precise estimates of the Y-model. The DPM did not assume side-effect monotonicity, but appears to support the assumption.

Figure 4.2 shows the posterior medians of all the PCEs over the entire $(D(0), D(1))$ space. The PCE surface is smoothly increasing as the compliance increases in both assignment arms, suggesting better compliance behavior leading to larger overall reduction in cholesterol level.

Comparison with the results of JR and BG is made on the estimated PCE at four selected principal stratum $S = (D_0, D_1)$, including the stratum of "median complier" under both assignments, $S = (0.68, 0.89)$. The comparison is displayed in Table 4.2, which includes the posterior medians and 95% credible intervals for the PCEs under the DPM approach,

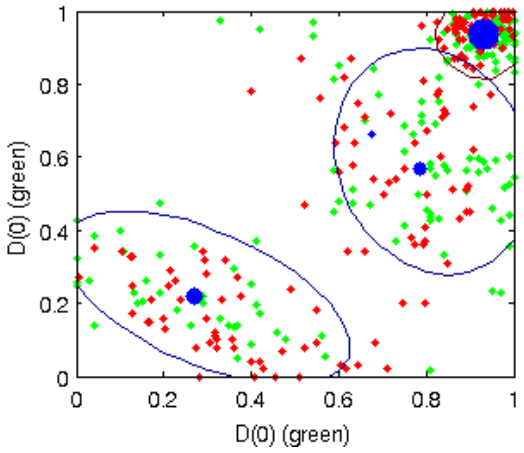

Figure 4.1: A single posterior imputation for $(D(0), D(1))$ based on the DPM S-model. The three predominant clusters appear to be continuous analogues of always-takers, never-takers, and defiers from binary PS.

(D_0, D_1)	DPM		Parametric (JR)		Copula (BG)
$(1, 1)$	45	(38, 52)	50	(39, 59)	51
$(0.68, 0.89)$	29	(25, 34)	24	(17, 30)	30
$(0, 1)$	0	-	-13	(-42, 27)	0
$(0, 0)$	0	-	5	(-6, 16)	0

Table 4.2: Estimated PCE for selected principal stratum (D_0, D_1) using the DPM approach, the fully parametric approach of JR, and the copula approach of BG.

and the corresponding estimates from the fully Bayesian parametric approach of JR and the copula approach of BG. The discrepancy between the entries of the final two rows of Table (4.2) is a result of the mean specifications (4.5). The results are comparable across methods, while the DPM approach provides tighter estimate intervals than the fully parametric approach, which again highlights the improved precision that results from the clustering structure imposed by the DP. Interval estimates were not provided by BG.

Using the randomized LRC-CPPT data, the proposed Bayesian semi-parametric approach obtains comparable point estimates but more precise interval estimates than those from existing methods. In addition, the DPM S-model reveals clusters of the latent principal strata that have natural interpretation as in the binary PS case. The performance of the DPM is now investigated in a more challenging observational study setting with covariates and binary outcomes.

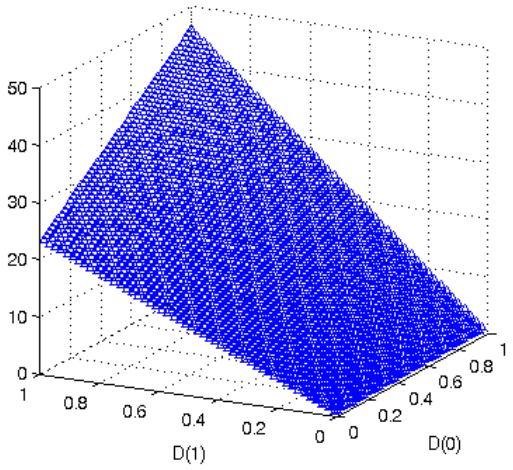

Figure 4.2: Median PCE over the entire $(D(0), D(1))$ space, as estimated by the DPM S-model.

4.4 Application to the Swedish National March Cohort

4.4.1 Data and Models

This section examines the effect of physical activity (PA) on cardiovascular disease (CVD) as it relates to body mass index (BMI) using the Swedish National March Cohort (NMC). The NMC was established in year 1997, when 300,000 Swedes participated in a national fund-raising event organized by the Swedish Cancer Society. Every participant was asked to fill in a questionnaire that included items on known or suspected risk factors for cancer and CVD. Questionnaire data were obtained on over 43,880 individuals. Using the Swedish patient registry, these individuals were followed for the period from year 1997 to 2004, and each CVD event was recorded. Further details on the NMC can be found in Lagerros (2006).

The question of scientific interest here is the extent of causal effect of PA on CVD risk mediated or not mediated through BMI. The principal strata with respect to the intermediate variable BMI is the joint potential values of BMI for an individual under high and low exercise regimes. The PCEs of the principal strata consisting of individuals whose BMI remains the same regardless of exercise can be interpreted as the (principal strata) direct effect (PSDE) of exercise on CVD reduction not mediated through BMI. Similarly, we can

define the (principal strata) indirect effect (PSIE) as the PCEs in the principal strata of individuals whose BMI would change due to exercise.

Sjölander *et al.* (2008), hereafter SJ, analyzed the NMC data using PS, where each subject is classified as either a "low-level exerciser" ($T = 1$) or a "high-level exerciser" ($T = 0$) based on self-reported history of PA; obese ($D = 1$) or not obese ($D = 0$) based on baseline BMI in year 1997 dichotomized at cutoff point 30; and "with disease" ($Y = 1$) or "without disease" ($Y = 0$) based on if the subject had at least one CVD event recorded during follow-up. Age is a strong confounder in this setting, and is the sole covariate reported in SJ. The PSDEs are the main causal estimand in SJ, and SJ found evidence for beneficial PSDE effects. In the current analysis, we follow the definition of T and Y in SJ, but analyze D (BMI) in its original continuous scale and let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ be the centered age and square of age. In addition, we will investigate both PSDEs and PSIEs.

Of the participants, 38,349 reported as "high-exercisers" and 2,956 reported as "low-exercisers". The former included 2,262 cases of CVD, while the latter included 172. The distributions of age are similar in both groups, with the $T = 0$ group being slightly older on average (Figure 4.3 (a)). There is a strong correlation between CVD incidence and both age and BMI (Figure 4.3 (d) and (f)). The distribution of BMI is right skewed, with a heavier tail in the "low-exercise" group, as seen in Figures 4.3 (b). A QQplot of BMI in the two arms (Figure 4.3 (c)) clearly suggests the "high exercisers" have lower BMI than the "low exercisers". To balance with the available 2,956 "low exercise" cases, 3,000 participants were sampled from the "high exercise" group, and these data were used in model fitting. Sampling was done randomly subject to the constraint that there be 177 CVD cases and 2,833 non-CVD cases in order to maintain the observed incidence rate.

Likelihood ratio tests on the observed marginal distribution of $Y(0)$ and $Y(1)$ suggest that both age and BMI are significant predictors of CVD risk in both arms. Among the various specifications that we have experimented, the following form leads to the most stable

model fitting and natural interpretation,

$$\begin{aligned}
\text{logit}\{\Pr(Y_i(1)=1|S_i,\boldsymbol{X}_i)\} &= \beta_0^Y + \beta_1^Y D_i(1) + \beta_2^Y X_{i1} \\
\text{logit}\{\Pr(Y_i(0)=1|S_i,\boldsymbol{X}_i)\} &= \beta_0^Y + \beta_1^Y D_i(1) + \beta_2^Y X_{i1} + \beta_3^Y D_i(0) + \beta_4^Y \frac{D_i(1)}{D_i(0)}.
\end{aligned} \tag{4.6}$$

Sharing β_0, β_1 and β_2 across the $Y(0)$ and $Y(1)$ models in effect assumes a common 'baseline' effect of BMI on CVD risk when people do not exercise. In fact, those are the identification assumptions that underly a succesful model fitting. The $Y(0)$ specification in (4.6) is the key modeling assumption because it allows correlation to be induced between $D_i(1)$ and $D_i(0)$ even though are not jointly observed. The prior distribution for $\boldsymbol{\beta}^Y = (\beta_0^Y, \beta_1^Y, \beta_2^Y, \beta_3^Y, \beta_4^Y)$ is set to be $\Pr(\boldsymbol{\beta}^Y) = \text{N}(\mathbf{0}, 5^2 I_5)$.

The scatter plot of average BMI and age shows a clear positive and curvilinear relationship between age and BMI. Likelihood ratio tests on the observed marginal distribution of Ds suggest that both age and square of age are significant predictors of BMI in both arms. Thus, we assume the following DPM S-model,

$$\Pr((D_i(0), D_i(1))|\boldsymbol{X}_i, \boldsymbol{\beta}^D) = \sum_{h=1}^{\infty} w_h c_h \text{N}\left(\begin{pmatrix} \eta_{0h} + X_{i1}\beta_{01}^D + X_{i2}\beta_{02}^D \\ \eta_{1h} + X_{i1}\beta_{11}^D + X_{i2}\beta_{12}^D \end{pmatrix}, \Sigma_h\right) 1_A, \tag{4.7}$$

with $A = \{(D_i(0), D_i(1)) : 0 < D_i(0), D_i(1) < 100\}$. Specifications for the DP are $G_0 = \text{N}(25I_2, \Sigma_0)\text{IW}(2, 3^2 I_2)$, with $\sigma_0^2 = \sigma_1^2 = 5^2$ and $\rho_{01} = 0.75$, and $\Pr(\alpha) = \text{Ga}(1, 1)$. The prior distribution for $\boldsymbol{\beta}^D$ is specified as $\Pr(\boldsymbol{\beta}^D) = \text{N}(\mathbf{0}, 10^2 I_4)$.

4.4.2 Results

The Y-model (4.6) and DPM S-model (4.7) are fitted to the randomly sub-sampled NMC data. Similarly as before, five parallel MCMC chains with different starting values were run 205,000 iterations, with the first 5,000 as burn-in. Mixing of the chains was determined to be adequate and all chains lead to highly similar posterior summary statistics. Sensitivity of

coefficient	median	95% cred. ints.
β_0^Y	-3.63	(-3.87, 3.40)
β_1^Y	0.05	(0.01, 0.09)
β_2^Y	0.10	(0.09, 0.11)
β_3^Y	-0.06	(-0.13, 0.01)
β_4^Y	-0.28	(-0.50, -0.05)
β_{01}^D	0.20	(0.17, 0.23)
β_{11}^D	0.24	(0.21, 0.27)
$\beta_{02}^D \times 10^2$	-0.10	(-0.18, -0.13)
$\beta_{12}^D \times 10^2$	-0.19	(-0.22, -0.16)

Table 4.3: Posterior medians and 95% credible intervals for the coefficients in the Y-model and S-model.

results to alternative hyperparameter specifications for α within the Gamma prior distribution for α is minimal and the approximation truncation level $H = 10$ appears to be adequate for DP approximation in this analysis. Results varied slightly with the sub-sampling, but led to similar parameter estimation.

Table 4.3 provides posterior medians and 95% credible intervals for the coefficients. Estimates of β_1^Y and β_2^Y suggest that both baseline BMI without exercise and age are positively associate with CVD incidence, while β_4^Y suggests a significant reduction in CVD as a result of reduction in BMI due to PA.

A representative posterior draw of principal strata $(D_i(0), D_i(1))$ along with the DPM configuration from the S-model (4.4) is displayed in Figure 4.4. There are two predominant clusters that are consistently found throughout all analyses: Component 1 in the middle of the 45^o line that consists of around 80% individuals whose BMI is stable regardless of PA; and component 2 above the 45^o line that consists of around 15% individuals whose BMI decreases with PA. The precise configuration of remaining components can vary in different MCMC chains, but still consistently suggest that the 5% remaining individuals are people who have a even larger reduction in BMI as a result of PA.

Figure 4.5 shows posterior medians and 95% credible intervals for PCEs old over the plausible range of $D(1)$ and $D(0)$ for individuals who are 60 years old. The PCE surface increases smoothly with both $D(1)$ and $D(0)$, suggesting that the causal effect of exercise in reducing the probability of developing CVD increases as one's BMI increases. Table 4.4

	Age = 50		Age = 60	
$S = (d_0, d_1)$	median	95% cred. ints.	median	95% cred. ints.
(20, 20)	0.09	(-0.77, 0.94)	0.23	(-1.91, 2.38)
(25, 25)	0.74	(0.22, 1.31)	1.86	(0.54, 3.26)
(30, 30)	1.52	(0.05, 2.57)	3.79	(1.24, 6.26)
(35, 35)	2.45	(0.50, 4.54)	6.03	(1.19, 10.86)
(20, 25)	0.29	(-0.93, 1.26)	0.72	(-2.27, 3.20)
(25, 30)	1.06	(0.31, 1.85)	2.61	(0.74, 4.52)
(30, 35)	2.00	(0.62, 3.69)	4.87	(1.54, 8.73)

Table 4.4: Posterior medians and 95% credible intervals for the percent PCE, $E(Y(1) - Y(0)|S = (d_0, d_1)) \times 100$, for selected principal strata S at age of 50 and 60 years.

provides PCEs for principal strata $S_i = (d_0, d_1)$ that are of scientific interest. BMI values of 18.5, 25, 30, 35 are the standard cut points of underweight, overweight, class I obese and class II obese, respectively. Since there are very few individuals with BMI below 18.5, we present PCEs with BMI being 20 instead. PSDEs are the PCEs on the 45^o line. We can see that the PSDEs increase as both age and BMI increase. For example, for a person whose BMI is 25 no matter whether he or she exercises, the reduction in CVD risk due to exercise is 0.74% and 1.86% when he or she is 50 and 60 years old, respectively, and the reduction increases to 1.52% and 3.79% respectively if his BMI is always 30. This means that even if exercise does not reduce the BMI, it does reduce the risk of CVD, and the benefit is bigger for older and heavier population. Our results also suggests a sizable PSIE of exercise on CVD mediated through BMI. For example, for a person whose BMI reduces from 30 to 25 as a result of exercise, the reduction in CVD risk is 1.06% and 2.61% when he or she is 50 and 60 years old, respectively; and that reduction is 2.00% and 4.87% respectively for a person whose BMI reduces from 35 to 30 when he or she exercise. But neither PSDE nor PSIE is significant for people with normal BMI (below 25).

The PSDE results here match the findings in SJ. However, analysis of PCE based on continuous BMI offers offers a more refined picture of the causal mechanism among PA, BMI and CVD risk than that based on dichotomized BMI. In addition, continuous analysis relies less on some standard but untestable identifying assumptions. Specifically, the monotonicity assumption (i.e., $D_i(0) < D_i(1)$) is not imposed as in SJ. Monotonicity is a key identifying

assumption in the case of binary D. Although practically plausible and supported by the qqplot of observed BMI in two groups (Figure 4.3 (c)), approximately 25% of the posterior draws of $(D_i(0), D_i(1))$ do not adhere to monotonicity when it was not enforced (Figure 4.4). In fact, the PCE results for the NMC data here are insensitive to this assumption. Moreover, PSDEs are directly estimated without relying on pre-fixed sensitivity parameters governing the degree of derivation from exclusion restrictions. Effectively, these assumptions are replaced with the Y-model and S-model specifications. But, the chosen specifications appear to be supported by the data.

4.5 Discussion

Flexible modeling for the joint distribution of continuous intermediate potential outcomes (continuous principal strata) is critical for PS inference. This chapter proposes and develops a Bayesian semi-parametric methodology based on DPMs that achieves the necessary flexibility for PS analysis in the presence of continuous intermediate variables. Additional benefits of the DPM approach are illustrated using the randomized LRC-CPPT data and the observational NMC study: The DPM produces comparable point estimates and more precise interval estimates relative to existing methods, and, provides a more refined view of causal mechanisms (e.g., those between PA, BMI and CVD) compared to non-continuous modeling approaches.

The results so far suggest that the Bayesian semi-parametric model via DPM offers a new approach to continuous intermediate variables in causal inference that has advantages in both inference and interpretability. However, the approach relies heavily on the information about the unobserved principal strata contained in the observed outcome variable. Naturally, model fitting for this context is more challenging for binary outcomes than continuous outcomes, since binary data contain less information. Inclusion of covariates in modeling can help remedy this difficulty by directly providing information about $(D(0), D(1))$ through the S-model, but it also appears that they are critical to synergistically improve the information available from the outcome data in the Y-model, particularly in the binary

outcomes case. There may not be enough information in the outcome model to recover latent PS structure without additional covariate information.

The ignorability assumption has been assumed throughout this chapter. Ignorability is usually plausible in randomized experiment, like the LRC-CPPT trial. It is more questionable in observational studies, like the NMC data. For example, age is the only covariate available in the analysis of the NMC, but there can be many remaining possible confounders, such as sex, genetic profile, etc. Sensitivity to derivation from the ignorability assumption in PS framework has been explored in Chapter 3, but only in the case of binary D. A systematic investigation for such sensitivity in the case of continuous intermediate variable is of interest.

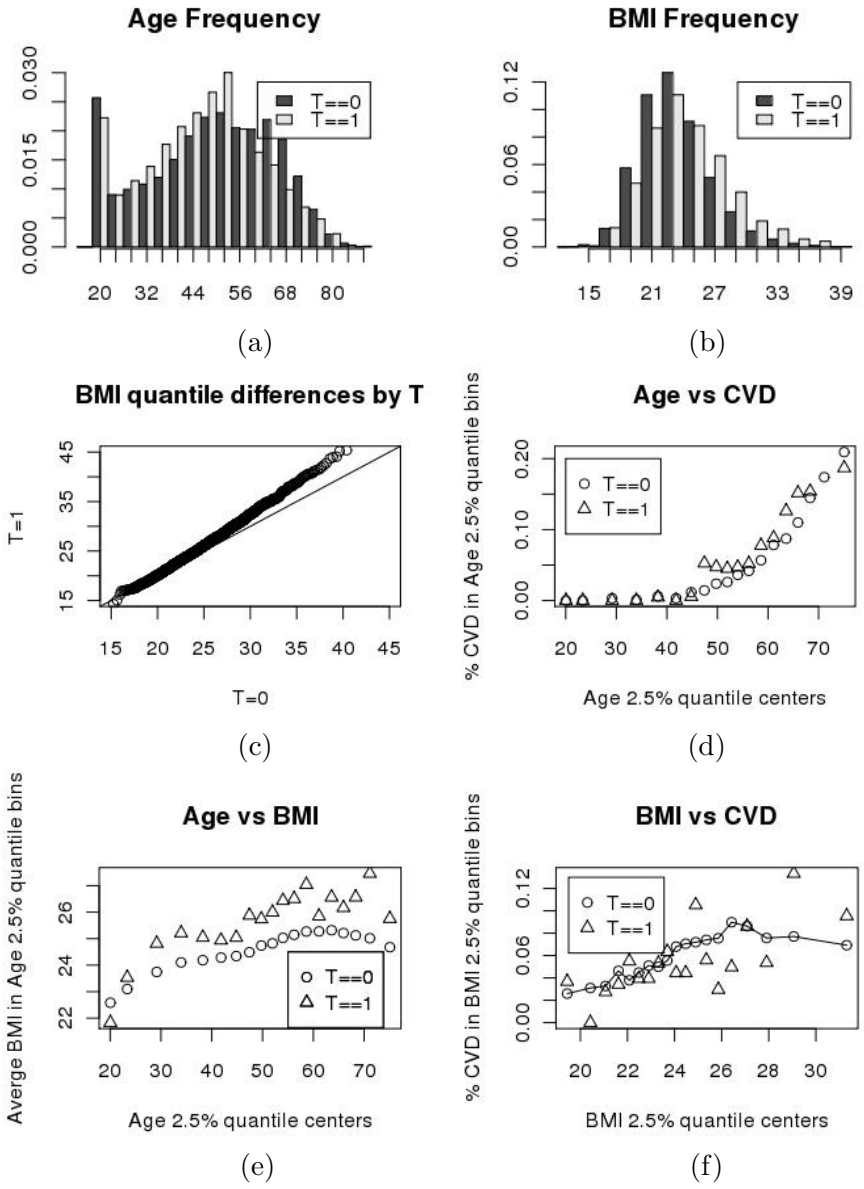

Figure 4.3: (a) is the histogram of age across T. (b) is the histogram of BMI across T. (c) is the qqplot of BMI of $T_i = 0$ versus $T_i = 1$ group: 98% of the points lie above the diagonal. (d) is the scatterplot of age versus CVD incidence, displaying a clear positive correlation. (e) is the scatterplot of age versus BMI. (f) is the scatterplot of BMI versus CVD incidence

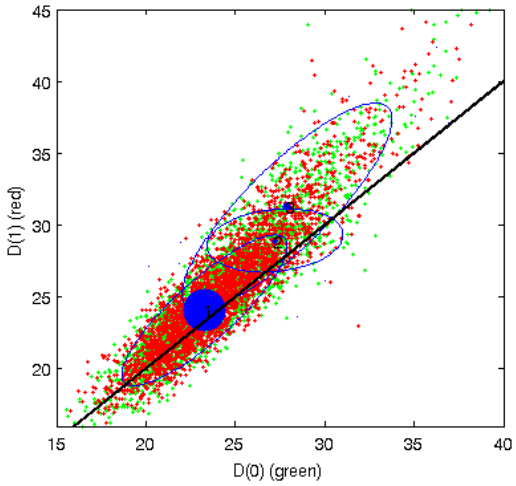

Figure 4.4: A representative posterior draw of principal strata S_i under the DPM S-model. Each component is labeled with a number and a light dot representing its mass contribution, e.g., component 1 contributes 80%. The solid line is the 45^o line.

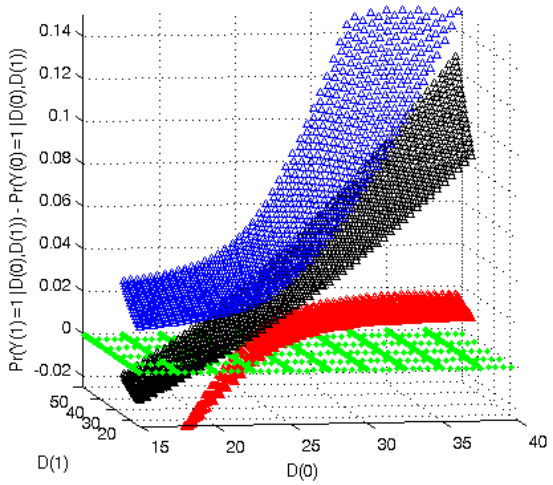

Figure 4.5: Median surface and point-wise 95% credible intervals for the PCE, over the relevant space of $(D(0), D(1))$ for individuals 10 years above the median age (60 years old). The green surface is the reference surface of PCE = 0.

Chapter 5

Conclusion and Future Direction

The motivation behind this thesis is to improve the ability of statistical inference, particularly Bayesian inference, to impact real-world problems through applied analyses. The initial catalyst for this research was the desire of applied and methodological researchers to improve our ability to analyze and understand birth outcomes data using sound, scientifically motivated modeling techniques. The work in this area developed into a study of causal inference methodology. There is much more to do, however.

5.1 Extensions to Joint Modeling Analyses

Chapter 2 provides a joint distribution of birthweight and censored gestational age conditional on covariates, and so readily accommodates inference concerning disparities in birthweight and/or gestational age in a richer way than previously considered. Nonetheless, this methodology leaves several topics unaddressed and so there are several further research opportunities related to this work.

First, thorough attention to mis-measurement in gestational age is of great importance since many forms of gestational age data cannot be accurately measured. This may be accomplished by embedding measurement error models within the current methodology. For example, within each component of the mixture approach, incorrect estimations of g_i^c might be addressed using an integer offset o_i modeled with a discrete distribution centered

on 0, as in

$$\Pr(b_i|g_i^c + u_i + o_i, x_i) \Pr(g_i^c + u_i + o_i|x_i) 1_{[0,1]}(u_i) \Pr(o_i).$$

Further exploration into measurement error for the various forms of mis-measured gestational age is necessary to determine the benefits and tradeoffs of potential modeling approaches, and to provide tools to address mis-measured gestational age. Connecting measurement error models to the mixture modeling approach is also of general methodological interest.

Second, further exploration into the role of covariates in the model's mixing proportions is warranted. This may be accomplished by replacing the current submodel for the mixing proportions with a multinomial logit model, such as

$$\log \frac{\Pr(Z_i = h|X_i)}{\Pr(Z_i = 1|X_i)} = X_i \beta_h^Z,$$

for $h \in \{2, \cdots, H\}$, and $\Pr(Z_i = 1|X_i) = 1 - \sum_{h=2}^{H} \Pr(Z_i = j|X_i)$. This allows covariate information to be incorporated into the predictions of component membership. Since certain components tend to correspond to at risk populations, predicting component membership may be useful in predicting at risk births. However, incorporation of covariates at the outcome level as well as component membership level will require careful consideration regarding interpretations of covariates.

Third, given the longitudinal nature of birth record data, a dynamic perspective could be considered to investigate if and how the joint distribution changes over time. One approach to dynamic modeling might connect the component location parameters over time t through a latent auto-regressive process, such as

$$\begin{aligned} \mu_{h,t} &= \theta_{h,t} + \epsilon_t,\ \epsilon_t \sim N(0, s^2) \\ \theta_{h,t} &= \phi\theta_{h,t-1} + \omega_t,\ \omega_t \sim N(0, v^2). \end{aligned}$$

Such an approach could be used to examine local nonstationarity, or, with the inclusion of

appropriate terms, cyclic temporal patters. Further dynamic structure on the covariance structure of each mixture component might also be considered. A related consideration is the inclusion of a spatial component in the modeling. Birth records are increasingly geocoded, and so in addition to dynamic considerations, examination of possible spatial structure underlying the data would be of interest. One approach might be to allow individuals in component h and at locations $\vec{s}$ to have mean locations drawn from a Gaussian process, e.g., $\vec{\mu}_{h,s} \sim \mathrm{N}(X_i\beta, C(\vec{s}))$, for some covariance function $C(\vec{s})$. Individual extensions of finite mixture models to dynamic and spatial settings are both of general interest, as is the extension of finite mixture models to the complete spatio-temporal setting.

5.2 Extension to Sensitivity Analyses

Inspired by the problems with intermediate variables observed in the setting of Chapter 2, as well as the unaddressed observational nature of Chapter 2, Chapter 3 extends PS analysis for binary treatments and binary intermediate variables into the observational setting by replacing the exclusion restriction and the strong ignorability assumption with a sensitivity analysis. However, several research paths remain open with respect to this methodology, and PS in general.

First, the proposed PS sensitivity analysis requires the standard monotonicity assumption, $D(1) > D(0)$, which rules out the possibility of defiers. This assumption could itself be included as part of the sensitivity analysis, perhaps by specifying $Pr(S_i = (1,0)|X_i)$ or some related quantity. In many settings the expected proportion of defiers would be low and so may have little impact on the results. However, this is not confirmed under the current sensitivity analyses and is currently not a standard consideration in other PS settings.

Second, the current sensitivity approach uses direct effect sensitivity parameters, δ_a and δ_n, to identify an overall treatment effect for compliers, θ_c. Often, however, direct effects themselves may be of primary interest. In this case, perhaps θ_c may instead be used as the sensitivity parameter to examine δ_a and δ_n in a reverse formulation. One interesting application of such a methodology would be to explore the extent of placebo effects in

randomized clinical trials.

Third, the PS sensitivity results provided here are deterministically related to the observed data and the sensitivity specification. Thus, if two replicate data sets differ in their observed values they will produce different sensitivity inference. The uncertainty is reflected in the estimated model parameters, but not in the sensitivity parameters since they are hand selected. Quantifying the potential uncertainty in realized confounding as it relates to sensitivity bound specifications would provide a principled approach to account for sensitivity specification uncertainty as it relates to randomly sampled data.

Fourth, on-the-fly propensity score matching may aid PS analysis. Many forms of PS, such as the sensitivity methodology presented here, require imputation of the latent principal strata. However, even if observed covariates are balanced across treatment arms, a given imputation set may not be balanced within a given principal strata. The properties of a matching procedure that dynamically balances observed covariates during principal strata imputation have not been considered. Perhaps such an approach could even improve imputation determination.

Fifth, missing data is always an issue in real data settings. The current sensitivity analysis does not address missing data, and it is not clear what the role of missing data should be within this methodology. In general, the consequences of missing data in PS analysis has not been examined. It is unclear if the currently available methods for missing data readily extend into the PS setting, or if new methodology is required.

5.3 Extensions to PS for Continuous Intermediate Variables

Continuing the focus on PS analysis from Chapter 3, Chapter 4 improves the current parametric based approach for PS for continuous intermediate variables by replacing it with a nonparametric Bayesian alternative. Because of the potential benefits of nonparametric approaches, such as those shown in Chapter 4, there are likely to be many further uses of similar methodologies for causal inference analysis in the future. With regards to the methodology presented in Chapter 4, there several major related projects to be investigated.

First, Chapter 3 extends PS analysis into the observational setting for binary treatments and intermediate variables only. A sensitivity analysis in the manner of Chapter 3 might be used to extend PS analysis into the observational setting for continuous intermediate variables as well. However, because the continuous setting results in an infinite collection of principal strata rather than three or four (depending on monotonicity), it is not clear that the approach of Chapter 3 can be immediately applied to the setting of Chapter 4. Thus, opportunities for new methodological development are likely to be available through this consideration. In general, methodologies that extend analyses such as PS for continuous intermediate variables to the observational setting will have many application opportunities.

Second, an alternative Bayesian semi-parametric analysis based on a copula may be considered and contrasted with the approach of Chapter 4. Rather than link the observed margins through a bivariate DPM, each observed margin may be modeled with a univariate DPM, or appropriate alternative approach, and then linked with the other through a copula. Copula based approaches appear to be very natural in the causal inference setting where only marginal distributions, and not the joint distributions of interest, are observed. Copulas are a relatively new introduction to causal inference, and so they provide many opportunities for development.

Third, the PS setting of Chapter 4 naturally involves a hierarchy of bivariate modeling. This setting is made difficult because it entails so much missing data during model fitting. However, there are many similar settings that do not entail missing data during fitting. For example, one version of the small area estimation problem involves a cheap psychological instrument Y_{0ij} and an expensive psychological instrument Y_{1ij} for individual i in school j, with the latter providing more accurate reading while also costing too much to implement on a large scale. Interest lies in

$$\begin{aligned}(Y_{0ij}, Y_{1ij}) &\sim \mathrm{N}((X_{0ij}\beta_0 + v_{0i}, X_{1ij}\beta_1 + v_{1i}), \Sigma_Y) \\ (v_{0j}, v_{1j}) &\sim \mathrm{N}((\mu_0, \mu_1), \Sigma_v)\end{aligned}$$

in order to relate the cheap instrument back to the more meaningful expensive instrument

through conditional prediction. This provides a way to estimate highly informative measurements in small areas, e.g., schools. The methodology of Chapter 4 applies to this setting with only minor changes, and provides the opportunity to replace the random effects model with a completely flexible nonparametric model in cases where random effects may not behave well under a Gaussian specification.

Bibliography

Ananth CV, Platt RW. Reexamining the Effects of Gestational Age, Fetal Growth, and Maternal Smoking on Neonatal Mortality. BMC Pregnancy Childbirth 2004; 4.

Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 1996; 91: 444–455.

Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. Journal of the American Statistical Association 1997; 92: 924–933.

Barnard J, Frangakis C, Hill J, Rubin D. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association 2003; 98: 311–314.

Bartolucci F, Grilli L. Modeling partial compliance through copulas in the principal stratification framework. Journal of the American Statistical Association 2010; under revision.

Berkson J. Limitations of the application of fourfold tables to hospital data. Biometrics Bulletin 1946; 2: 47–53.

Bickel P, Hammel E, O'Connell J. Sex Bias in Graduate Admissions: Data From Berkeley. Science 1975; 187: 398–404.

Blyth C. On Simpsons Paradox and the Sure-Thing Principle. Journal of the American Statistical Association 1972; 67: 364–366.

Chung Y, DB D. The local Dirichlet process. Annals of the Institute of Statistical Mathematics 2009; .

Dawid A. Conditional Independence for Statistical Operations. Annals of Statistics 1980; 8: 598–617.

Dawid A. Causal Inference Without Counterfactuals. Journal of the American Statistical Association 2000; 95: 407–424.

Dawid A, Didelez V. Identifying the consequences of dynamic treatment strategies. 2005.

Delbaere I, Vansteelandt S, De Bacquer D, Verstraelen H, Gerris J, De Sutter P, *et al.* Should We Adjust for Gestational Age When Analyzing Birth weights? The Use of Z-Scores Revisited. Human Reproduction 2007; 22: 2080–2083.

Dempster A, Laird N, Rubin D. Maximum Likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 1977; 39: 1–38.

Dey D, Rao C. Handbook of Statistics 25: Bayesian Thinking, Modeling and Computation, chapter 16: Bayesian Modeling and Inference on Mixtures of Distributions. New York: Elsevier, 2005; .

Didelez V, Dawid A, Geneletti S. Direct and indirect effects of sequential treatments. 2006.

Didelez V, Sheehan N. Mendelian Randomisation: Why Epidemiology needs a Formal Language for Causality. 2005.

Diebolt J, Robert CP. Estimation of Finite Mixture Distributions through Bayesian Sampling. Journal of the Royal Statistical Society Series B Methodological 1994; 56: 363–375.

Efron B, Feldman D. Compliance as an Explanatory Variable in Clinical Trials. Journal of the American Statistical Association 1991; 86: 9–17.

Escobar M, West M. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association 1995a; 90: 577–588.

Escobar M, West M. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association 1995b; 90: 577–588.

Fang F, Stratton H, Gage TB. Multiple mortality optima due to heterogeneity in the birth cohort: A continuous model of birth weight by gestational age specific infant mortality. American Journal of Human Biology 2007; 19: 475–486.

Ferguson T. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics 1973; 1: 209–230.

Ferguson T. Prior distributions on spaces of probability measures. Annals of Statistics 1974; 2: 615–29.

Fisher R. The Causes of Human Variability. Eugenics Review 1918; 10: 213–220.

Fisher R. Statistical Methods for Research Workers. London, U.K.: Oliver & Boyd, 1925.

Flores C, Flores-Lagunes A. Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness. IZA DP No 4237 2009; .

Frangakis C, Brookmeyer R, Varadhan R, Mahboobeh S, Vlahov D, Strathdee S. Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a Needle Exchange Program. Journal of the American Statistical Association 2004; 99: 239–249.

Frangakis C, Rubin D. Principal Stratification in Causal Inference. Biometrics 2002; 58: 21–29.

Frimpong EY, Gage TB, Stratton H. Identifiability of bivariate mixtures: An Application to infant mortality models. American Statistical Association Joint Statistical Meetings Proceedings, Statistics in Epidemiology, 2009.

Gage TB. Variability of Gestational Age Distributions by Sex and Ethnicity: An analysis Using Mixture Models. American Journal of Human Biology 2000; 12: 181–191.

Gage TB. Birth-Weight-Specific Infant and Neonatal Mortality: Effects of Heterogeneity in the Birth Cohort. Human Biology 2002; 74: 165–184.

Gage TB. Classification of births by birth weight and gestational age: an application of multivariate mixture models. Annals of Human Biology 2003; 30: 589–604.

Gage TB, Bauer MJ, Heffner N, Stratton H. Pediatric Paradox: Heterogeneity in the Birth Cohort. Human Biology 2004; 76: 327–342.

Gage TB, Fang F, H S. Modeling the pediatric paradox: Birth weight by gestational age. Biodemography and Social Biology 2008a; 54: 95–112.

Gage TB, Fang F, ONeill E, H S. Maternal age and infant mortality: A test of the Wilcox-Russell hypothesis. American Journal of Epidemiology 2008b; 169: 294–303.

Gage TB, Therriault G. Variability of Birth-Weight Distributions by Sex and Ethnicity: An Analysis Using Mixture Models. Human Biology 1998; 70: 517–534.

Gelfand AE, Ghosh SK. Model Choice: A Minimum Posterior Predictive Loss Approach. Biometrika 1998; 85: 1–11.

Gelfand AE, Sahu SK. Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models. Journal of the American Statistical Association 1999; 94: 247–253.

Gelfand AE, Sahu SK, Carlin BP. Efficient Parameterizations for Normal Linear Mixed Models. Biometrika 1995; 82: 479–488.

Gelfand AE, Smith AFM. Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association 1990; 85: 398–409.

Gelman A. Statistical Modeling, Causal Inference, and Social Science. `http://www.stat.columbia.edu/~cook/movabletype/archives/2006/04/amusing_example.html`, July 18 2008.

Gilbert P, Bosch J, Hudgens M. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. Biometrics 2003; 59: 531–541.

Goldberger A. Structural Equation Methods in the Social Sciences. Econometrica 1972; 40: 979–1001.

Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiological research. Epidemiology 1999; 10: 37–48.

Griffin J, Steel M. Order-based dependent dirichlet processes. Journal of the American Statistical Association 2004; 101: 179–194.

Grimes DA. Discussion: Impaired Growth and Risk of Fetal Death: Is the Tenth Percentile the Appropriate Standard? American Journal of Obstetrics and Gynecology 1998; 178: 658–669.

Haavelmo T. The Statistical Implications of a System of Simultaneous Equations. Econometrica 1943; 11: 1–12.

Haavelmo T. The Probability Approach in Econometrics. Econometrica 1944; 12 (Supplement): 1–115.

Heckman J. Econometric Causality. International Statistical Review, International Statistical Institute 2008; 76: 1–27.

Hernán M, Brumback B, Robins J. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. Epidemiology 2000; 11: 561–570.

Hernández-Díaz S, Schisterman EF, Hernán MA. The Birth Weight "Paradox" Uncovered? American Journal of Epidemiology 2006; 164: 1115–1120.

Hill J, McCulloch R. Bayesian Nonparametric Modeling for Causal Inference. Biometrics 2007; .

Hill J, Reiter J, Zanutto E. A comparison of experimental and observational data analyses. In Gelman A, Meng X, editors, Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley, 2004; pp. 49–60.

Hirano K, Imbens G, Rubin D, Zhou XH. Assessing the effect of an inuenza vaccine in an encouragement design. Biostatistics 2000; 1: 69–88.

Holland P. Statistics and Causal Inference. Journal of the American Statistical Association 1986; 81: 945–970.

Hudgens M, Halloran M. Towards causal inference with interference. Journal of the American Statistical Association 2008; 103: 832–842.

Imai K, Yamamoto T. Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis. Americal Journal of Political Science 2010; In press.

Imbens G, Angrist J. Identification and estimation of local average treatment effects. Journal of the American Statistical Association 1996; 91: 444–455.

Imbens G, Rubin D. Causal Inference in Statistics. Cambridge UK: Cambridge University Press, 2010.

Imbens W, Rubin D. Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics 1997; 25: 305–327.

Ishwaran H, James L. Gibbs Sampling Methods for Stick Breaking Priors. Journal of the American Statistical Association 2001; 96: 161–173.

Jasra A, Holmes CC, Stephens DA. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. Statistical Science 2005; 20: 50–67.

Jin H, Rubin D. Principal Stratification for Causal Inference With Extended Partial Compliance. Journal of the American Statistical Association 2008; 103: 101–111.

Joseph KS. Theory of obstetrics: An epidemiological framework for justifying medically indicated early delivery. BMC Pregnancy and Childbirth 2007; 7.

Joseph KS, Demissie K, Platt RW, Ananth CV, McCarthy BJ, Kramer MS. A parsimonious explanation for intersecting perinatal mortality curves: understanding the effects of race and maternal smoking. BMC Pregnancy and Childbirth 2004a; 7.

Joseph KS, Liu S, Demissie K, Wen SW, Platt RW, Ananth CV, *et al.* A parsimonious explanation for intersecting perinatal mortality curves: understanding the effect of plurality and of parity. BMC Pregnancy and Childbirth 2004b; 7.

Karn MN, Penrose LS. Birthweight and Gestation Time in Relation to Maternal Age, Parity and Infant Survival. Annals of Eugenics 1951; 16: 147–160.

Lagerros Y. Physical activity from the epidemiological perspectivemeasurement issues and health effects. Ph.D. thesis, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden., 2006.

Lavine M. More aspects of Polya tree distributions for statistical modeling. The Annals of Statistics 1992a; 22: 1161–1176.

Lavine M. Some aspects of Polya tree distributions for statistical modeling. The Annals of Statistics 1992b; 20: 1222–1235.

McClellan M, McNeil B, Newhouse J. Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables. Econometrica 1994; 11: 859–866.

McDonald C, HIU S, Tierney W. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. MD Computing 1992; 9: 304–312.

McLachlan G, Peel D. Finite Mixture Models. New York: Wiley-Interscience, 2000.

Morgan M. The History of Econometric Ideas. Cambridge, U.K.: Cambridge University Press, 1990.

Müller P, Quintana F. Nonparametric Bayesian Data Analysis. Statistical Science 2004; 19: 95–110.

Müller P, Quintana F, Rosner G. A method for combining inference across related nonparametric Bayesian models. Journal of the Royal Statistical Society, Series B 2004; 66: 735–749.

Neyman J. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. translated in Statistical Science 1990 1923; 5: 465–480.

Oja H, Koiranen M, Rantakallio P. Fitting Mixture Models to Birth Weight Data: A Case Study. Biometrics 1991; 47: 883–897.

Paneth NS. The Problem of Low Birth Weight. The Future of Children 1995; 5: 19–34.

Pearl J. Causal Diagrams for Empirical Research. Biometrika 1995; 82: 669–710.

Pearl J. Causality: Models, Reasoning, and Inference. New York: Cambridge University Press, 2000.

Pearl J. Direct and indrect effects. In Proceedings of the 17th Conference on Uncertainy in artificial intelligence. San Francisco, CA: Morgan Kaufmann, pp. 411–420.

Pearson K, Lee A, Bramley-Moore L. Mathematical Contributions to the Theory of Evolution. - VI. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses. Proceedings of the Royal Society of London 1898; 68: 163–167.

Platt RW, Joseph KS, Ananth CV, Gordines J, Abrahamowicz M, Kramer MS. A Proportional Hazards Model with Time-dependent Covariates and Time-varying Effects for Analysis of Fetal and Infant Death. American Journal of Epidemiology 2003; 160: 199–206.

Robins J. Data, Design, and Background Knowledge in Etiologic Inference. Epidemiology 2001; 12: 313–320.

Robins J. Semantics of causal DAG models and the identification of direct and indirect effects. In Green P, Hjort N, Richardson S, editors, Highly Structured Stochastic Systems. Oxford University Press, 2003; pp. 70–81.

Robins J, Greenland S. Identifiability and Exchangeability for Direct and Indirect Effects. Epidemiology 1992; 3: 143–155.

Robins J, Greenland S, Hu F. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. Journal of the American Statistical Association 1999; 94: 687–700.

Robins J, Hernán M, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology 2000; 11: 550–560.

Rodrguez A, Dunson D, Alan E Gelfand A. The nested Dirichlet process. Journal of the American Statistical Association 2008; 103: 1131–1154.

Rosenbaum P. The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. Journal of the Royal Statistical Society: Series B 1984; 147: 656–666.

Rosenbaum P. Observational Studies. New York: Springer, 2002.

Rosenbaum P, Rubin D. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. Journal of the Royal Statistical Society Series B (Methodological) 1983a; 45: 212–218.

Rosenbaum P, Rubin D. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Journal of the Royal Statistical Society: Series B 1983b; 70: 41–55.

Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974; 66: 688–701.

Rubin D. Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics 1978; 6: 34–58.

Rubin D. Comment on 'Randomization analysis of experimental Data: The Fisher randomization test' by D. Basu. Journal of the American Statistical Association 1996; 75: 591–593.

Rubin D. Causal inference through potential outcomes and principal stratification: application to studies with censoring due to death. Statistical Science 2006; 91: 299–321.

Rubin DB. Direct and Indirect Causal Effects via Potential Outcomes. Scandinavian Journal of Statistics 2004; 31: 161–170.

Sethuraman J. A Const ructive Definition of Dirichlet Priors. Statistical Sinica 1994; 4: 639–650.

Sethurman J. A constructive definition of Dirichlet priors. Statistica Sinica 1994; 4: 639–650.

Sjölander A, Humphreys K, Vansteelandt S, Bellocco R, Palmgren J. Sensitivity Analysis for Principal Stratum Direct Effects, with an Application to a Study of Physical Activity and Coronary Heart Disease. Biometrics 2008; 65: 514–520.

Small D, Rosenbaum P. War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases. Journal of the American Statistical Association 2008; 103: 924–933.

Tanner M, Wong W. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 1987; 82: 528–540.

Tassone EC, Miranda ML, Gelfand AE. Disaggregated Spatial Modeling for Areal Unit Categorical Data. Journal Of The Royal Statistical Society Series C 2010; 59: 175–190.

Teh Y, Jordan M, Beal M, Blei D. Hierarchical Dirichlet Processes. Journal of the American Statistical Association 2003; 101: 1566–1581.

Wasserman L. Comment on "Estimation of the causal effect of time-varying exposure on the marginal mean of a repeated binary outome" by Robins, *et. al.* Journal of the American Statistical Association 1999; 94: 687–690.

Wasserman L. All of Nonparametric Statistics. New York: Springer, 2005.

Wilcox AJ. On the importance – And The Unimportance – of Birthweight. International Journal of Epidemiology 2001; 30: 1233–1241.

Wilcox AJ, Russell I. Why small black infants have a lower mortality rate than small white infants: The case for population-specific standards for birth weight. The Journal of Pediatrics 1990; 116: 7–10.

Wilcox AJ, Skjaerven R. Birth Weight and Perinatal Mortality: The Effect of Gestational Age. American Journal of Public Health 1992; 82: 378–382.

Wright P. Appendix to The Tariff on Animal and Vegetable Oils. New York: MacMillan, 1928.

Wright P. The Method of Path Coefficients. Annals of Mathematical Statistics 1934; 5: 161–215.

Ylppö A. Das Wachstum der Frühgeborenen yon der Geburt bis zum Schulalter. Z Kinderheilkd 1919; 24: 111–178.

Yule U. Notes on the Theory of Association of Attributes in Statistics. Biometrika 1903; 2: 121–134.

Biography

Scott Lee Schwartz was born in Winfield, Illinois, to Kenneth and Cheryl Schwartz on November 7^{th}, 1982. He is the oldest of four, including two brothers, Daniel and Jonathan, and one sister, Kendra. Scott was home schooled until enrolling in the 8^{th} grade at Kirby junior high school, in Kirby, Texas, and in 2000, he completed his high school education in a dual home school/public school program with Judson high school, in Converse, Texas, and San Antonio College, in San Antonio, Texas. Scott attended San Antonio College in 2000 full time for one year, and then transfered to Trinity University, in San Antonio, Texas, where received a BA in Mathematics and BS in Computer Science in 2005. While at Trinity, Scott was a member of the Mens Soccer team, and in 2003, he won the NCAA Mens Division III National Championship with the team. In 2006, Scott began graduate school in the Department of Statistical Science at Duke University, in Durham, North Carolina, and in 2008, he received his MA in Statistical Science from Duke. Scott is currently scheduled to complete is PhD studies at Duke in May, 2010, and will accept a postdoctoral position in Nutrition and Bioinformatics at Texas A&M University in College Station, Texas, pending successful completion of the defense of this Dissertation.