Bayesian Estimation and Sensitivity Analysis for Causal Inference

by

Abbas Zaidi

Department of Statistical Science
Duke University

Date: __________________________

Approved:

____
Sayan Mukherjee, Advisor

____
Fan Li

____
Alexander Volfovsky

____
Alexandre Belloni

____
Jerome Reiter

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

2019

ABSTRACT

Bayesian Estimation and Sensitivity Analysis for Causal Inference

by

Abbas Zaidi

Department of Statistical Science
Duke University

Date: ______________________________
Approved:

____
Sayan Mukherjee, Advisor

____
Fan Li

____
Alexander Volfovsky

____
Alexandre Belloni

____
Jerome Reiter

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

2019

Abstract

This disseration aims to explore Bayesian estimation and sensitivity analysis methods for causal inference. In chapter 1, we present an overview of fundamental ideas from causal inference along with an outline of the methodological developments that we hope to tackle along with motivating applications. In chapter 2, we develop a Gaussian-process mixture model for heterogeneous treatment effect estimation that leverages the use of transformed outcomes. The approach we will present attempts to improve point estimation and uncertainty quantification relative to past work that has used transformed variable related methods as well as traditional outcome modeling. Earlier work on modeling treatment effect heterogeneity using transformed outcomes has relied on tree based methods such as single regression trees and random forests. Under the umbrella of non-parametric models, outcome modeling has been performed using Bayesian additive regression trees and various flavors of weighted single trees. These approaches work well when large samples are available, but suffer in smaller samples where results are more sensitive to model misspecification – our method attempts to garner improvements in inference quality via a correctly specified model rooted in Bayesian non-parametrics. Furthermore, while we begin with a model that assumes that the treatment assignment mechanism is known, an extension where it is learnt from the data is presented for applications to observational studies. Our approach is applied to simulated and real data to demonstrate our theorized improvements in inference with respect to two causal estimands: the conditional average treatment effect and the average treatment effect. By leveraging our correctly specified model, we are able to more accurately estimate the treatment effects while reducing their variance. In chapter 3, we parametrically and hierarchically estimate the average causal effects of different lengths of stay in the Udayan Ghar Program under the assumption that selection into different lengths is based on a set of observed covariates. This program was piloted in New Delhi, India as a means of providing a residential surrogate to vulnerable and at risk children with the hope of improving their psychological development. We find that the estimated effects

on the psychological ideas of self concept and ego resilience (measured by the standardized Piers-Harris score) increase with the length of the time spent in the program. We are also able to conclude that there are measurable differences that exist between male and female children that spend time in the program. In chapter 4, we supplement the estimation of hierarchical dose-response function estimation by introducing a novel sensitivity-analysis and summarization strategy for assessing the robustness of our results to violations of the assumption of unconfoundedness. Finally, in chapter 5, we summarize what this dissertation has achieved, and briefly outline important areas where our work warrants further development.

Acknowledgements

First and foremost, I would like to thank my advisors Sayan Mukherkjee and Fan Li. Both of them have had an immeasurable impact on my view of statistics, mathematics and how to truly have fun pursuing original research. Their guidance has been instrumental in teaching me the subtleties of machine learning (Sayan) and the art of causal inference (Fan). Fan especially, taught me to rigorously put my thoughts to paper in a manner not reminiscent of a high schooler. I will forever be indebted to both of them.

Credit is also due to the members of committee, Jerome Reiter, Alexander Volfovsky and Alexandre Belloni. Alexander Volfovsky and Alexandre Belloni both played a key role in helping me think of the ideas in this dissertation as a means of answering broader questions outside the field of statistics. Jerome Reiter is a model for what a good educator is, and I can sincerely say that my love of teaching statistics is largely inspired by him. I would also like to extend my deepest gratitude to my collaborators Sumedha Ariely-Gupta and Felipe Barrientos without whom the last two chapters of this dissertation would never have come to fruition.

I also gratefully acknowledge the support of Andrea Mercatanti (Department of Statistics, Bank of Italy) and Kiran Modi (Udayan Ghar Program), for providing the data that motivated the questions that this dissertation has aimed to address.

Finally, there are two extraordinary women without whose support I never would have gotten the level of self-aggrandizing delusion needed to quit my job as a banker and return to school: my sister, Aliza Haider, and my friend, Jingmin Wu. Of course, no mention of my time at the Duke Department of Statistical Science, would be complete without a shoutout to the clowns and homies in the *social office*, Old Chem 222: Michael “Magic-Mike” Akande, Michael “Hatin’-Hipster” Jauch, Elizabeth “Yinzer” Lorenzi and Lindsay “Chuck” Berry.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Causal Inference is one of the core disciplines of modern statistics and has a myriad of applications in the social and natural sciences. The goal of standard statistical analysis techniques is to infer associations using samples drawn from the population. This objective is well managed by any number of classical methods assuming of course that the the data generating conditions remain static. Causal inference aims to go further – its objective is to learn the dynamics of evolving statistical conditions and that can help establish causal relationships in observed data.

The main statistical contributions to causal inference is the estimation of treatment effects. This dissertation is focused on treatment effect estimation under various scenarios. The difficulty in this type of estimation task and establishing causality, varies based on the design of the study, and the collection of the data, but its application is massively important to addressing a plethora of causal questions. These are in vital to understanding and quantifying concerns in economic and public policy effectiveness – the application that this dissertation focuses on.

There are two major frameworks for the quantification of treatment effects. The first of these is referred to as the *potential outcomes framework* or the *Rubin Causal Model* as detailed in [77]. A second approach that relies on directed a-cyclical graphs has been presented in [62] – the relationships between these approaches is summarized in [61]. This dissertation will focus on the former and we will present methodological developments in how this approach can be extended to address complexities in our application areas.

The remainder of this chapter has been organized as follows: section 1.1 summarizes the potential outcomes framework, followed by a discussion of notation in section 1.2, and key assumptions in section 1.3. The assumptions that we discuss will motivate the causal estimands in section 1.4, and the propensity score methods needed for their estimation from observational data are given in section 1.6. A brief overview of our data and motivating questions is given in section 1.7 and the methodological developments this gives rise to are given in section 1.8.

1.1 The Potential Outcomes Framework

The potential outcomes framework establishes a framework for quantifying causal effects as comparisons between how recipients of treatments respond under various conditions. This approach is rooted in three key components. First, are the *potential outcomes*, that is each unit of observation (i.e. the recipient of the treatment) has a possible outcome value under each level of the treatment condition. This definition leads to the natural definition of the treatment effect as the magnitude of the difference between the potential outcomes under different treatments for the same set of units. However, since a unit of observation cannot both receive the treatment and also *not* receive it, unit level treatment effects are non-identifiable in the absence of some key assumptions, that are discussed at length in the next section. This is often referred to as the *fundamental problem of causal inference*.

The lack of identifiability of the individual causal effects helps us define the second component of the model: the *assignment mechanism*. This is the process that determines which observations receive the treatment – knowing it defines which potential outcomes are observed, and which ones are missing. As a general comment, the taxonomy of causal studies is defined by the various ways in which the potential outcomes and treatment assignment can be separated. The final component of the potential outcomes framework, is the *model* that ties the observed covariates to the potential outcomes. Jointly these three components define the framework introduced in the pioneering work presented in [76].

This estimation of causal effects discussed so far rely on comparisons between binary treatment conditions i.e. whether a unit of observation receives the treatment or the control. Conceptually, this can can be extended to situations where treatments are non-binary and in the most extreme case *continuous* as introduced in [42]. Under this more complex case, the effect of the treatment is no longer a difference between the potential outcomes of the treated and control conditions; rather, for each individual, the potential outcome is a *dose-response function* evaluated at various levels of the treatment along some continuum of possible levels. Analogous to the binary treatment case, the individual dose-response function is also non-identifiable since not all potential outcomes at each treatment exposure are observed for each observation. Additional assumptions are needed for inference.

In this disseration, we develop methodology that leverages both of these ideas in a Bayesian framework for estimating causal effects flexibly and robustly. These methods rely on assumptions that are necessary for the identification of the causal effects – the next two sections will define notation and leverage them to describe the aforementioned assumptions in context.

1.2 Notation

For units of observation $i = 1, \ldots, n$, we will denote the treatment $W_i = \omega$. In the binary case $\omega \in \{0, 1\}$ where $\omega = 0$ and $\omega = 1$ correspond to the treated and control conditions of the treatment variable respectively. In the case of continuous treatments $\omega \in \mathcal{W}$ where $\mathcal{W}$ is an interval indicating a continuously measured exposure to the treatment. In both cases, we use X_i as a collection of baseline characteristics or pre-treatment covariates for each unit observed – based on the assumptions we will present in the next section these covariates are crucial for garnering causal inference. Each unit has a potential outcome denoted as $Y_i(\omega)$ which can correspond to either a continuous treatment or a binary one.

1.3 Assumptions

Under the potential outcomes framework, in order to identify causal effects certain assumptions about the treatment assignment mechanism need to be satisfied. We begin our discussion of these assumptions in the context of binary treatments first since this is motivated by the first methodological development that we present in this dissertation.

We provide brief summaries of the assumptions below, using the same notation defined above (see [49] for detailed descriptions of these assumptions):

A1: *Individualistic* — the assignment mechanism is separable with respect to the unit assignment probabilities and independent of the pre-treatment covariates,

$$\mathbb{P}\left(\{W_i\}_{i=1}^n \mid \{Y_i(1), Y_i(0), X_i\}_{i=1}^n\right) = \prod_{i=1}^n \mathbb{P}(W_i \mid Y_i(1), Y_i(0), X_i). \tag{1.1}$$

This assumption is satisfied under independence of the treatment assignments which is a stronger condition and therefore implies that the treatment assignment is individualistic.

A2: *Probabilistic* — each unit has positive probability of receiving all levels of the treatment, so for $i = 1, ..., n$ and all w,

$$0 < \mathbb{P}(W_i = w \mid Y_i(0), Y_i(1), X_i) < 1. \tag{1.2}$$

In many disciplines, this assumption is referred to as the *overlap* assumption.

A3: *Unconfounded* — the treatment assignment does not depend on the potential outcomes conditional on the pre-treatment covariates,

$$\mathbb{P}\left(\{W_i\}_{i=1}^n \mid \{Y_i(1), Y_i(0), X_i\}_{i=1}^n\right) = \mathbb{P}\left(\{W_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n\right), \tag{1.3}$$

or equivalently $\{Y_i(1), Y_i(0)\}_{i=1}^n \perp \{W_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n$. This assumption is also referred to as the *ignorable* assignment assumption.

Assumptions A2 and A3 are together known as the *strong unconfoundedness* assumption and grants the indentifying equivalence between the potential outcome and the causal conditioning, $Y(W = w) \stackrel{P}{=} Y \mid W = w$.

The unconfoundedness assumption necessary for continuous treatments is *weak unconfoundedness* that can be characterized as assumption A3.1, A3.1: Weakly ignorable — $Y_i(w) \perp W_i | X_i \ \ \forall \omega \in \mathcal{W}$

The difference between strong and weak unconfoundedness is one of joint independence between the treatment assignment and the potential outcomes. However, in practice this difference is negligible.

All of the above assumptions summarized here are always satisfied in randomized experiments since the assignment mechanism is known and can therefore be explicitly controlled for. In observational studies (the focus of this dissertation) the assumptions may hold to varying degrees and need to be carefully considered. For instance, A2, which is also sometimes referred to as the overlap condition can be directly assessed. However, by comparison A3 and A3.1 are untestable and therefore indirect techniques are warranted to determine how their violations impact the analysis. A tertiary objective of this dissertation is to develop a new technique for doing so.

Beyond these, we make one additional assumption that allows us to simplify the statistical the models we specify in this dissertation.

A4: Stable Unit Treatment Value Assumption (SUTVA) — This condition assumes no interference between observations, and that there are no multiple versions of the treatment.

In its absence, we would need to define a different potential outcome for the unit of observation not just for each treatment received by that unit but for each combination of treatments received by every other observation in the experiment. Pioneering work in conducting causal inference in the presence of interference has been considered in [45] and [84], and presents an interesting area to which the ideas in this dissertation can be potentially extended.

1.4 Causal Estimands

Unit level causal effects, both in the case of binary and continuous treatments are non-identifiable due to the fundamental problem of causal inference. However, under the assumptions detailed in the last section population level treatment effects can be defined and estimated.

For binary treatments, the most common class of estimands are average treatment effects, including the population average treatment effect (ATE) that is given as,

$$\tau^P = \mathbb{E}[Y_i(1) - Y_i(0)]$$

However, the ATE as defined earlier, can often mask information about subpopulations defined by levels of the various covariates – an ATE that is zero, is not generally an ATE that is zero for everyone in the subpopulation. This necessitates an estimand that can quantify this type of heterogeneity. The average measures also include average effects for various subpopulations defined by values of covariates such as the conditional average treatment effect (CATE) which is defined as,

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$$

Analogously, the estimand in the context of continuous treatment exposures, is no longer a single parameter – we are interested in the average dose-response function (DRF) that is defined as follows as discussed at length in [42].

$$\mu(\omega) = \mathbb{E}[Y_i(\omega)]; \quad \text{for all } \omega > 0$$

It is imperative to recognize here that the estimands defined under the potential outcomes framework that we have discussed are not tied to any type of specific model. This is a departure from the practice of defining causal estimands as the parameters of models i.e. the coefficient of the treatment indicator in a linear regression model. This *model-free* definition of the causal estimands means that there is a clear division of the definition and estimation of treatment effects. It also indicates that working under this framework, we can approach the problem of estimating causal effects with various flexible modeling techniques.

1.5 Understanding Unconfoundedness

From the summary definition in section 1.3 that a core assumption central to causal inference is that of ignorability or unconfoundedness. Based on its mathematical definition, it indicates that such an assignment mechanism drops out of the likelihood and can thus be ignored in statistical inference under mild conditions. In the literature, it is also commonplace to refer to the ignorable assignment assumption as the *no unmeasured confounders* assumption or *selection on observables* assumption. It is vital to note that while ignorability is always satisfied in randomized experiments, and is commonly assumed to hold in the context of observational studies, it is not directly testable. It is an assumption on unmeasured data: under the potential outcomes framework, the distributions of the unobserved counterfactual is not informed by the data that we do observe. When satisfied, the potential outcomes corresponding to both treatment conditions are balanced – in practice, we use the balance of covariates as a proxy for balance of the potential outcomes.

In conjunction with the assumption of the assignment mechanism being individualistic, unconfoundedness grants that the assignment mechanism can be summarized using propensity scoring techniques we summarize in section 1.6. A number of approaches can be used to indirectly test its plausibility and to determine the robustness of causal analyses to its violations. There are a number of techniques for indirect testing, the most commonly used of which is the method of lagged outcomes as introduced in [18] while robustness checks using sensitivity analyses as introduced in [69] are easiest to generalize. One methodological contribution that this dissertation makes is a novel technique for performing this aforementioned type of sensitivity analysis for continuous treatment regimes.

1.6 Propensity Score Methods

Techniques that use the propensity score under the assumption of unconfoundedness can be used to help garner causal results from observational data. In such sitations, the assumptions around the assignment mechanism are difficult to evaluate, and furthermore, the assignment mechanism is usually unknown and therefore cannot be controlled for perfectly. This difficulty is further exacerbated by the existence of both measured and unmeasured confounders that may be balanced between various treatment groups or levels of the treatments.

Propensity scoring techniques can help address questions around covariate imbalance for measured confounders, both under binary and continuous treatments. Formally, for binary treatments, the *propensity score* is defined as the conditional probability of receiving the treatment given the pre-treatment covariates. Mathematically,

$$e(x) = \mathbb{P}(W_i = 1 | X_i = x)$$

while for continuous treatments, we define the *generalized propensity score* as a conditional density of receiving a particular level of the treatment,

$$r(\omega, X) = f_{\omega|X}(\omega | X = x)$$

In both presentations of the propensity score, it retains its *balancing property*. For the binary treatment case,

$$W_i \perp X_i | e(X_i)$$

which means that the propensity score balances the distribution of all the pre-treatment covariates between the two treatment groups. A similar balancing property can also be defined for the continuous covariates that we can state as follows,

$$\mathbf{1}\{W = \omega\} \perp X | r(\omega, X)$$

Under unconfoundedness, adjusting for the propensity score between the various treatment conditions can remove all biases that exist as a result of covariate imbalance. In addition, to its balancing properties, unconfoundedness given the pre-treatment covariates extends to unconfoundedness given the propensity score (under both types of treatment regimes). Past work has indicated that a consistent estimate of the propensity score leads to more efficient estimation of the ATE and the CATE than the true propensity score does – therefore, even in randomized trials, estimating the propensity score has known and documented benefits [49].

Techniques that use the propensity score are usually two-staged: first, estimating the propensity score, and second, estimating the causal effect using the propensity score. In the second stage, classical, non-parametric methods, use the propensity score as a scalar summary of the covariates, and uses this with matching, weighting or some combination of the two as a means of comparing averages across different treatment levels. Intuitively, these methods rely on the idea that units with similar propensity scores, are approximately balanced in the possibly multi-dimensional covariates, and thus a simple comparison of the observed outcomes between treatment and control groups is an unbiased estimator of the causal effects. However, in application, ignoring how the covariates relate to the measured outcomes can potentially induce large biases.

In an effort to minimize this type of *residual* bias i.e. bias that exists as a result of imbalances, can be corrected using a combination of outcome modeling combined with propensity score methods to give rise to *doubly-robust* techniques. Past work, as reviewed in [49] shows that doubly-robust estimators yields consistent estimators if either the model at the propensity score stage, or at the outcome stage is correctly specified, but not necessarily both. The first two parts of this dissertation involves developing methodology that use the propensity scoring with flexible modeling to estimate causal effects under both binary and continuous treatment regimes – the work we present is in spirit, very similar to the ideas of double-robustness.

1.7 Summary of Data

The methodological developements presented in this dissertation are motivated by questions from two datasets. The first, comes form work developed by [59] to explore the causal effects of debit card usage on annual household spending. This data has been studied in detail in past work, and our focus has been to extend the earlier analysis which focused on homogeneous treatment effect estimation. Our goal is to help uncover heterogeneities in the effects of debit card usage as functions of various demographic and economic variables.

Our second dataset, comes from a pilot study aimed at understanding the effects of a new foster care system on various mental health outcomes for orphaned and abandoned children in India. The objective of this dataset is to understand how growing up in this system impacts children in terms of their psychological health, as measured by ego-resilience metrics such as the Piers-Harris score [15]. In addition to understanding what the overall response to the system, we are also interested in investigating how these outcomes vary between different sub-populations such as those defined by gender and developmental stage both of which are fundamental policy development factors.

All three methods that we will briefly introduce in section 1.8 are motivated by limitations of current approaches that became apparent while analyzing our two datasets. The second chapter of this dissertation focuses exclusively on the first dataset and uses the technique we propose to model treatment effect heterogeneity for debit card usage. The third and fourth chapters focus on the second dataset. In chapter three we estimate a hierarchical dose-response function to investigate how spending time in this new residential care system impacts mental health outcomes. Our analysis is augmented by sensitivity analyses in chapter four to determine the reliability of the results presented and determine whether our causal findings are defensible. In the final section, we will present closing remarks to motivate the future of this work.

1.8 Methodological Contributions

This dissertation hopes to make three contributions to statistical methodology for causal inference. First, we focus on the identification and modeling of heterogeneous treatment effects using the CATE under binary treatment regimes. Our approach is related to the ideas of inverse probability weighting (see [20]) in the form of transformed outcomes as popularized in [5]. We leverage the implications of the transformation to introduce a new correctly specified model that retains the flexibility of the original technique, but aims to improve the overall quality of inference in terms of both point estimation and uncertainty quantification. The inference for this model is rooted in Bayesian non-parametric ideas, and we introduce a joint approach to modeling the propensity score and the causal effects in an effort to better learn both.

The second part of this dissertation moves into the realm of non-binary treatments and focuses on dose-response function estimation under continuous exposure regimes as explored in [24] extended to multi-stage clustered data. Past work with this type of data has been restricted to estimating causal effects under binary treatments as in [53, 94], while continuous exposure regimes have been restricted to single stage analyses as explored in [42, 24]. We leverage ideas from both to develop an a new parametric approach using hierarchical Bayesian models for estimating the dose-response function. Our approach uses the generalized propensity score both accounting for idiosyncrasies at the exposure and outcome stages via hierarchical models to garner more reliable inference.

Our final contribution develops a new technique for conducting sensitivity analyses to assess the impacts of unmeasured confounding in a hierarchical setting for continuous treatment regimes. Earlier work on this problem has largely focused on assessing this impact under binary treatments (see [69] for the seminal work on this topic) and is generally restricted to single stage data. We develop these ideas further and propose a flexible, likelihood based method, that is rooted in concepts of Bayesian sensitivity analysis using simulated unmeasured confounder values, as introduced in [58, 34]. We want to understand how the dose response function varies as the effects of unmeasured confounding are varied in some systematic way. Furthermore, our approach can be generalized to assess various types of unmeasured confounding i.e. where the unmeasured confounder is binary or continuous. These three contributions provide a rich set of future areas in causal inference that we are interested in exploring.

Chapter 2

Heterogeneous Treatment Effect Estimation with Gaussian Process Mixtures

2.1 Introduction

The estimation of treatment effects is one of the core problems in causal inference. A treatment effect is a measure used to compare interventions in randomized experiments, policy analysis, and medical trials. The treatment effect measures the difference in outcomes between units assigned to the treatment versus those assigned to the control. There have been a variety of related approaches for estimating treatment effects including those based on graphical models [62] and the potential outcomes framework [77]. In this chapter, we develop methodology that builds on the potential outcomes framework as defined in [76] to estimate treatment effects.

In the potential outcomes framework we compare the observed outcome to the outcome under the counterfactual, that is, what the outcome would be under a different set of treatment conditions. If the counterfactual outcome were known then the treatment effect on an individual unit is the difference between the outcome under the observed and counterfactual interventions. The *fundamental problem of causal inference* is that in general for any unit one can only observe the outcome under a single treatment condition. As a consequence unit level causal effects are not identifiable. However, population level causal effects can be identified under some standard assumptions (see Section 2.2). An estimator of population level effects is the average treatment effect (ATE) which is a measure of the difference in the mean outcomes between units assigned to the treatment and units assigned to the control. If treatment effects are homogenous across individuals then estimators such as the ATE that consider causal effects at an aggregate level are reasonable, however such estimators will overlook subgroup or covariate-level specific heterogeneity in treatment effects. There is evidence that heterogeneity in treatment effects is more the rule than the exception [39, 31, 93].

A quantity in addressing heterogeneous treatment effects is the conditional average treatment effect (CATE) which is the average treatment effect conditional on the covariate level of a unit of observation. One can consider the CATE as a difference of two regression functions – the average response given treatment at a set of covariate levels minus the the average response assuming the control condition and the same set of covariate levels. One can estimate the ATE by marginalizing the CATE over the joint distribution of the covariates. There are a number of approaches for estimating the two aforementioned causal estimands. The main approach for modeling heterogeneous treatment effects based on the CATE is conditional mean regression. Under this approach, we model the CATE as a difference between the conditional mean outcome given the treatment for particular covariate levels minus the mean outcome given the control at the same covariate levels [20]. The implementation of these models can be approached both parametrically and non-parametrically.

The most popular parametric methods for estimating the difference between the conditional mean outcomes include linear and polynomial regression [62], along with penalized regression approaches such as least absolute subset selection operator and ridge regression [86]. The non-parametric techniques apply non-parametric regression models to estimate the difference between the conditional means. Examples of non-parametric regression methods that have been applied to this problem include boosting [64], Bayesian additive regression trees (BART) [41, 36, 17] as well as classical regression trees [5, 11] and random forests [89, 25, 12]. These methods have some limitations to their use and we provide a brief discussion of these.

The use of random forests for CATE estimation as defined in [89] provides some interesting theoretical results that allow for probabilistically valid statistical inference. These methods are theorized to outperform classical methods particularly in the presence of irrelevant covariates. This technique however, has been demonstrated to be outperformed in application [36]. In addition, without a procedure for imposing a degree of regularization, random forests are difficult to actually deploy for heterogenous treatment effect estimation [91]. BART and its variants [36, 41] present a persuasive argument for their use in application, but there is limited work on their formal inferential properties [89] for learning heterogeneous treatment effects. Specifically for BART, formal statistical analysis is hurdled by the lack of theory arguing posterior concentration around the true conditional mean function – the key quantity of interest in heterogenous treatment effect estimation via conditional mean regression.

An alternative to modeling the difference in conditional mean outcomes is the use of Transformed Response s or Outcome Variables (TRV) [21, 10]. This approach introduces a transformation for the outcome and the treatment indicator variable for which the conditional expectation given a covariate level is equivalent to the CATE. This allows it to be used with *off-the-shelf* machine learning techniques and has been applied to optimal treatment policy estimation. More recent work on the TRV has attempted to model it as a function of the observed covariates via regression trees [5] and boosting [64]. This has resulted in questions of estimation quality of the approach given the high variance of the procedure. We assert that this is a consequence of the properties of the TRV that have not been explicitly accounted for in the model since past work has relied on using it as a benchmark for other methods [4].

In this chapter we introduce a novel non-parametric Bayesian model based on Gaussian process regression [81, 67] for inference of the TRV that allows us to infer a posterior distribution on the CATE. The model we propose is a finite mixture of Gaussian-processes [66] that leverages the distribution implied by the transformation. This specification is aimed at improving the overall quality of inference on the treatment effects with a correctly specified model.

This approach has benefits over both conditional mean regression and other TRV based techniques. In practice, we never estimate either the treatment nor the control function perfectly and different covariate distributions for the treatment and control groups can lead to biases in the treatment effect estimation [64]. The TRV allows for the joint modeling of information from both the treated and control groups which can help circumvent the aforementioned estimation challenge which for instance has been discussed as a specific limitation of conditional mean regression with random forests [89]. This joint modeling is also an improvement over Bayesian techniques that place individual vague priors on the treatment and control outcome models since the prior on the treatment effect as the difference of the two is *doubly* vague. This can make inference a challenge since it is difficult to control the degree of heterogeneity that the model adapts to. Furthermore the TRV generates unbiased estimates for the CATE [64].

In addition to its benefits over conditional mean regression methods, the model we introduce offers four advantages over other TRV modeling approaches. First, we significantly improve the accuracy of point estimation by explicitly modeling the distribution of the transformed outcome. Second, by modeling the distribution of the transformed outcome specifically we are able to greatly reduce the variance of causal estimands i.e. the average treatment effect and the conditional average treatment effect. Reducing the variance of the estimators is crucial since this has been the main criticism of the TRV approach [5, 64]. This provides tighter uncertainty intervals relative to the approaches discussed in [5] and [89]. Third, our approach is well suited for instances when the treated and control groups share information since our proposed mechanism jointly models the behavior of both via the transformation.

Our main contribution is that we improve the overall quality of inference by improving the point estimation with a correctly specified model. In addition, the proposed framework is flexible in that we do not assume a functional form for how heterogeneity of treatment effects are driven by the levels of the observed covariates. Finally, our proposed framework is easily adapted to studies where the mechanism by which individuals receive the treatment is unknown. For this problem, past work has relied on a two-stage procedure for learning this treatment assignment mechanism first and then utilizing this in the model. We instead propose an approach whereby the treatment assignment mechanism and the treatment effects are jointly learnt in a unified framework. By working under this paradigm we have a twofold gain. First, the uncertainty quantification from our proposed model accounts for the uncertainty for all stages of inference including the learning of the assignment mechanism, and the treatment effects. Second as a by-product of the *feedback* in between the two estimation stages, the assignment mechanism makes more complete use of the data, which can improve estimation of causal effects.

The remainder of this chapter is organized as follows: in Section 2.2, we introduce the TRV, the relevant notation and the assumptions inherent to the TRV approach. We state our new model in Section 2.3. Our approach is benchmarked against to TRV regression trees and random forests, along with non-TRV weighted tree methods as discussed in [4], as well as Bayesian tree models in [36, 41] on both simulated and real data in Section 2.4. We close with a summary of our findings and possible areas of future work.

2.2 Transformed Response Variables Framework

In this section, we will define the notation used throughout this chapter, the foundational assumptions for causal estimation, and formulate the transformed response variables (TRV) approach.

The observed data $\mathcal{D}$ consists of a sample of size n where for each observation we are given a response variable $Y_i \in \mathbb{R}$ and a covariate vector$X_i \in \mathbb{R}^p$. In addition to the observed data, we denote as $W_i \in \{0, 1\}$ the treatment assignment. The corresponding treatment assignment probability is denoted as e_i. Finally, the potential outcome is denoted as $Y_i(W_i = w)$.

Under the potential outcomes framework, in order to estimate treatment effects from observational data certain assumptions about the treatment assignment mechanism need to be satisfied. Briefly, these assumptions are that the treatment assignment is *individualistic* (A1), *probabilistic* (A2) and *ignorable*(A3). Details of these assumptions are left to the reader in [49]. A1 and A2 are implied under the assumption that the units of observation are a simple random sample from the target population that are independent and identically distributed.

Assumptions A2 and A3 are together known as the *strong ignorability* assumption and grants the indentifying equivalence between the potential outcome and the causal conditioning, $Y(W = w) \stackrel{P}{=} Y \mid W = w$. All three of the assumptions summarized here are always satisfied in randomized trials; In observational studies the assumptions may hold to varying degrees.

For instance, A2, which is also sometimes referred to as the overlap or common support condition condition can be directly assessed. However, by comparison A3 is untestable and therefore indirect techniques are needed to determine the degree to which it is satisfied most commonly via sensitivity analyses [73]. We explore sensitivity analyses in detail in Chapter 4. These assumptions are necessary for the formal results in the transformed response variable framework to hold.

Beyond these, we make one additional assumption that allows us to simplify the statistical the model we specify in this chapter: Stable Unit Treatment Value Assumption (SUTVA) — This condition assumes no interference between observations, and that there are no multiple versions of the treatment (A4). In its absence, we would need to define a different potential outcome for the unit of observation not just for each treatment received by that unit but for each combination of treatments received by every other observation in the experiment. Relaxing these assumptions will be discussed in Section 2.5 as an avenue that our future work will aim to explore.

The causal estimands considered are the conditional average treatment effect (CATE), that we denote as $\tau(x)$and the average treatment effect (ATE) that we denote as τ^p. $\tau(x)$ is the primary estimate of interest in modeling heterogenous treatment effects and is defined as,

$$\tau(x) = \mathbb{E}_Y[Y_i(1) \mid X_i = x] - \mathbb{E}_Y[Y_i(0) \mid X_i = x], \tag{2.1}$$

the ATE can be derived by integrating over the the joint distribution of the covariates

$$\tau^p = \mathbb{E}_X\Big[\mathbb{E}_Y[Y(1) \mid X_i = x] - \mathbb{E}_Y[Y(0) \mid X_i = x]\Big] = \mathbb{E}_X[\tau(X_i)]. \tag{2.2}$$

The idea behind the transformed response variable apporach is to define a variable Y_i^* for which the conditional expectation with respect to the response recovers the CATE under A3 (see Appendix .1 for a proof of this result). These estimands implicitly assume that SUTVA holds. A transformation that satisfies the above condition is,

$$Y_i^* = f(W_i, Y_i, e_i) = \frac{W_i - e_i}{e_i(1 - e_i)} Y_i \tag{2.3}$$

The transformation requires knowledge of the probability of receiving the treatment. We assume that the treatment assignment probability depends on the observed covariate levels, or $e_i = \mathbb{P}(W_i = 1|X_i = x)$ and is therefore a *propensity score*. A trivial example is when the propensity score is a fixed covariate independent value, $e_i = e$ as in randomized trials. This is not an example commonly seen in real observational causal inference problems and is as such not considered as a part of the model presented here, albeit [5, 3, 4] consider it as a means of model validation.

2.2.1 Strengths and Weaknesses of Past Work in TRV Modeling

TRV modeling offers three main advantages when used for estimating treatment effects as demonstrated in prior studies. Foremost amongst these is that the TRV can easily be modeled with any supervised learning method. For instance, regression trees and random forests have been used [5, 3, 89] as has boosting [64]. This is not an exhaustive list, and there are a myriad of other methods that can be used in conjunction with the TRV to estimate heterogeneous treatment effects. Furthermore, relative to conditional mean regression, this method does not ignore the propensity score which explicitly enters the estimation via the transformation. Finally, based on the modeling approach used, we can tackle this problem flexibly and therefore avoid issues arising from model misspecification since it is likely that there are complex relationships between the covariates and heterogeneity of the treatment effects. Despite their usefulness, the TRVs have some key weaknesses.

First, as mentioned in [5] and [64] using TRVs as CATE estimators results in high variance estimates of the causal estimands. By construction the treatment assignment probability and the assignment itself only enter the model implicitly via the transformation. Specifically, the treatment assignment probability only appears in the denominator, and if close to zero or one, the variance can spike. In conjunction with the high variance predictions of some flexible models such as regression trees, this means that the method suffers in terms of efficiency and the quality of inference is therefore degraded. Second, uncertainty quantification in this framework relies entirely on the use of the bootstrap – this is applicable both to single regression trees as well as the other ensemble learning methods which have been used for TRV modeling. This presents two concerns. First, prior work [90] has suggested that in certain applications the Monte Carlo error can dominate the uncertainty quantification produced – therefore in conjunction with the high variance inherent to the aforementioned approaches, we might be unable to garner useful insights. If treatment effects are small (near zero), the conflation of the Monte Carlo noise with the underlying sampling noise may lead us to overstate the variance and therefore lower the power of our analysis. In addition note that when the sample size is small, [64] demonstrate that the variance of the TRV is small as well – it increases with increasing sample size. Hence, in situations where bootstrapping is likely to do well i.e. in large samples, the high variance of the TRV is even more so an issue.

Based on these limitations, we propose the Gaussian process mixture model in Section 2.3. Our proposed model attempts to overcome the aforementioned limitations by leveraging the mixture distribution implied by the transformation. In addition, we still aim to model the TRV flexibly and capture the complexity of treatment effect heterogeneity. We garner gains in the quality of inference by constructing a likelihood that reflects the error structure imposed by the TRV under some basic assumptions that earlier work with this technique has ignored. The details of these findings will be discussed in greater depth in Section 2.4 where these approaches are applied to real and simulated data.

2.3 The Gaussian Process Mixture Model

In this section we specify a non-parametric Bayesian model based on a mixture of Gaussian processes to model heterogenous treatment effects. Our model is based on the transformed response variable framework. It is motivated by three objectives: (1) to explicitly model the distribution implied by the transformed outcome with the goal of reducing the variance of the TRV generated estimates that have hitherto been produced using non-probabilistic models, (2) model the two treatment groups jointly so we can borrow strength and therefore improve inference even relative to non-TRV based methods for estimating treatment effects, and (3) making more complete use of the data by jointly modeling the transformed response as well as the treatment assignment probabilities in a one step model. The *feedback* between the two stages in joint modeling can improve the point estimation of treatment effects and the propensity scores [95]. Throughout this section we assume A1-A4 are satisfied.

2.3.1 Model Specification

A natural starting point is to consider two non-parametric regression functions for the response under treatment and control, respectively

$$\begin{aligned} Y_i(1) &= f_1(X_i = x) + \varepsilon_i(1), \quad \epsilon_i(1) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \\ Y_i(0) &= f_0(X_i = x) + \varepsilon_i(0), \quad \epsilon_i(0) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2). \end{aligned}$$

Substituting these non-parametric regression functions under the treatment and control cases in the definition of the TRV, we can define two new functions $g(\cdot)$ and $h(\cdot)$. These can be interpreted in terms of the non-parametric regression functions as

$$g(X_i = x) = f_1(X_i = x) - f_0(X_i = x), \quad h(X_i = x) = \frac{f_1(X_i = x)}{e_i} + \frac{f_0(X_i = x)}{1 - e_i}$$

A detailed derivation of this model is given in Appendix .2. This definition allows us to specify the following mixture model for the transformed outcome that takes treatment-control heterogeneity into account by simultaneously modeling the contributions from both groups

$$Y_i^* = g(X_i = x) + \varepsilon_i^*, \quad \varepsilon_i^* \sim e_i \mathrm{N}\left((1 - e_i)h(X_i = x), \frac{1}{e_i^2}\sigma^2 \right) + (1 - e_i)\mathrm{N}\left(- e_i h(X_i = x), \frac{1}{(1 - e_i)^2}\sigma^2 \right) \tag{2.4}$$

The argument for specifying the TRV mixture model rather than individual models for the treatment and control is that in many cases the conditionals $Y_i \mid X_i, W_i = 1$ and $Y_i \mid X_i, W_i = 0$ may not be perfectly estimable and ignoring shared information between the treated and untreated groups is a potential source of bias in the treatment effect estimation [64]. Furthermore, under this joint approach, we can show a direct equivalency to the CATE

$$\mathbb{E}[Y_i^* | X_i = x] = \mathbb{E}[g(x) | X_i = x] = \mathbb{E}[f_1(x) - f_0(x) | X_i = x]$$

This model can be considered under two specifications – when the treatment assignment probabilities are known and when they need to be inferred from the data. The details of each specification are given in Sections 2.3.1 and 2.3.1 for the two cases respectively.

Model specification with known assignment probabilities

We will place Gaussian process priors on both g and h and will specify an inverse gamma prior on σ^2 to leverage conjugacy. For the case where the treatment assignment probabilities are unknown, we will specify two additional levels in the prior structure.

For the case where the treatment probabilities are known we specify the following model

$$
\begin{aligned}
Y_i^* &= g(X_i = x) + \varepsilon_i^* \\
\varepsilon_i^* \sim e_i \mathrm{N}\Big((1-e_i)h(X_i = x), \frac{1}{e_i^2}\sigma^2\Big) &+ (1-e_i)\mathrm{N}\Big(-e_i h(X_i = x), \frac{1}{(1-e_i)^2}\sigma^2\Big) \\
g &\sim \mathrm{GP}(0, \kappa_g), \\
h &\sim \mathrm{GP}(0, \kappa_h), \\
\sigma^2 &\sim \mathrm{IG}(a, b).
\end{aligned} \tag{2.5}
$$

Here $\mathrm{IG}(a,b)$ is the inverse gamma distribution with hyper-parameters a and b and $\mathrm{GP}(\mathbf{0}, \boldsymbol{\kappa})$ denotes the Gaussian process priors on the function g and h. Both priors are zero mean and have covariance kernels specified (1) a non-stationary linear kernel $\kappa_g(u,v) = s_0^2 + \sum_{i=1}^{p} s_i^2 (u_i - c_i)(v_i - c_i)$, with hyper-parameters $s_0^2, \ldots s_p^2$ on g and (2) a square exponential, $\kappa_h(u,v) = s_h^2 \exp\{-\tau^2 \|u - v\|^2\}$ with hyper-parameters τ, s^2 on h. These kernels rely on the notion of similarity between data points – if the inputs are closer together than the target values of the response, in this case the TRV are also likely to be close together. Under the Gaussian process prior, the kernel functions described above formally define what is near or similar.

The hyper-parameters $s_0^2, \ldots s_p^2$ can be interpreted in the context of linear regression with $\{\text{Normal} \sim (0, s_j^2)\}_{j=0}^{p}$ priors on the $p+1$ regression coefficients including the intercept. The offset $\{c_i\}_{i=1}^{p}$ determines the x coordinate of the point that all the lines in the posterior is meant to go through. This provides some insight into how these can be set for applied modeling problems. In cases where there is a large number of covariates, many of which are thought to share information, the prior variance for those dimensions can be made small, with a higher degree of mass concentrated near zero to induce more shrinkage. In contrast, where there is a small number of important covariates the prior variance can be set to make the prior more diffuse. The offset can be set to the average of each covariates observed value. This is a general overview of the strategy that we have employed.

Model specification with unknown assignment probabilities

Computing the TRV requires knowledge of the treatment assignment probabilities $\{e_i\}_{i=1}^n$. In the case where these are unknown we consider them as latent variables and add extra levels to the hierarchical model specified in (2.5) to model the treatment assignment probabilities. We model the assignment probabilities individually so for notational ease, later in this chapter we use $\boldsymbol{e} = \{e_i\}_{i=1}^n$. Our specification, *apriori*, assumes that the assignment mechanism and the outcome model are independent.

Modeling the Propensity Score In order to learn the treatment assignment probabilities, we specify a probit regression model that is layered onto the model defined in (2.5).

$$\begin{aligned} W_i &\sim \text{Ber}(e_i) \\ e_i &= \Phi(X_i\boldsymbol{\beta}) \\ \boldsymbol{\beta} &\sim \text{N}_{p+1}(0, \Psi_{p+1\times p+1}). \end{aligned} \tag{2.6}$$

Where Φ denotes the standard Normal cumulative distribution function. In this chapter we will only consider the above Gaussian prior on $\boldsymbol{\beta}$ with prior covariance Ψ. However, additional complexity can be added by allowing the coefficient vector $\boldsymbol{\beta}$ to vary via a hierarchical prior structure as may be motivated by more complex hierarchical or clustered data.

2.3.2 Posterior Sampling with Known Assignment Probabilities

Inference for the model specified in Section 2.3.1 involves sampling from a posterior distribution via straightforward Gibbs-sampling.

We define $\mathbf{g} = (g(X_1), \ldots, g(X_n))$ and $\mathbf{h} = (h(X_1), \ldots, h(X_n))$ as the values of the two regression functions on the training data. We denote the TRV as $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)$. In this case the target joint posterior distribution is

$$\pi(\mathbf{g}, \mathbf{h}, \sigma^2 \mid \mathcal{D}). \tag{2.7}$$

Due to prior conjugacy the conditional distributions: $\pi(\mathbf{g} \mid \mathbf{h}, \sigma^2, \mathcal{D})$, $\pi(\mathbf{h} \mid \mathbf{g}, \sigma^2, \mathcal{D})$ and $\pi(\sigma^2 \mid \mathbf{h}, \mathbf{g}, \mathcal{D})$ all have simple forms that we can easily sample from. We first state some matrices and vectors that will enter our calculations: $\mathbf{D}$ is an $n \times n$ diagonal matrix with entries $\mathbf{D}_{ii} = \left(\frac{W_i}{e_i^2}\sigma^2 + \frac{1-W_i}{(1-e_i)^2}\sigma^2\right)$, $\mathbf{\Lambda}$ is also an $n \times n$ diagonal matrix with entries $\mathbf{\Lambda}_{ii} = \left(W_i(1-e_i) + (1-W_i)(-e_i)\right)$, $\mathbf{K}$ is also an $n \times n$ diagonal matrix with entries $\mathbf{K}_{ii} = \sigma^2 \mathbf{D}_{ii}$, and $\mathbf{m} = \Lambda \mathbf{H}$. We also denote the covariance matrix $\boldsymbol{\kappa}_g$ with the ij-th entry as taking the value $\kappa_g(x_i, x_j)$ and similarly $\boldsymbol{\kappa}_h$ is a matrix with the ij-th entry taking the value $\kappa_h(x_i, x_j)$. We now state the conditional distributions that will enter our Gibbs sampler

$$\begin{aligned}
&\pi(\mathbf{g} \mid \mathbf{h}, \sigma^2, \mathcal{D}) \sim \mathrm{N}\big((\boldsymbol{\kappa}_g^{-1} + \mathbf{D}^{-1})^{-1}(\mathbf{D}^{-1}\mathbf{Y}^* - \mathbf{m}\}, \{\boldsymbol{\kappa}_g^{-1} + \mathbf{D}^{-1})^{-1}\big), \\
&\pi(\mathbf{h} \mid \mathbf{g}, \sigma^2, \mathcal{D}) \sim \mathrm{N}\Bigg((\boldsymbol{\kappa}_h^{-1} + \mathbf{\Lambda}^T\mathbf{D}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{D}^{-1}(\mathbf{Y}^* - \mathbf{g}), (\boldsymbol{\kappa}_h^{-1} + \mathbf{\Lambda}^T\mathbf{D}^{-1}\mathbf{\Lambda})^{-1}\Bigg), \\
&\pi(\sigma^2 \mid \mathbf{h}, \mathbf{g}, \mathcal{D}) \sim \mathrm{IG}\Bigg(a + \frac{n}{2}, b + \frac{(\mathbf{Y}^* - \mathbf{g} - \mathbf{m})^T\mathbf{K}^{-1}(\mathbf{Y}^* - \mathbf{g} - \mathbf{m})}{2}\Bigg).
\end{aligned} \tag{2.8}$$

Given the above conditional distributions we can run the following Gibbs sampling procedure to generate a sequence $(\mathbf{g}^{(j)}, \mathbf{h}^{(j)}, \sigma^{((j)})_{j=1}^K$

a) Initialize $\mathbf{h}^{(0)}$, $\sigma^{(0)}$, and $\mathbf{g}^{(0)}$;

b) For $j = 1, ..., K$

  1) $\mathbf{g}^{(j)} \sim \pi(\mathbf{g} \mid \mathbf{h}^{(j-1)}, \sigma^{(j-1)}, \mathcal{D})$;

  2) $\mathbf{h}^{(j)} \sim \pi(\mathbf{h} \mid \mathbf{g}^{(j)}, \sigma^{(j-1)}, \mathcal{D})$;

  2) $\sigma^{(j)} \sim \pi(\sigma \mid \mathbf{h}^{(j)}, \mathbf{g}^{(j)}, \mathcal{D})$.

Given the sequence $(\mathbf{g}^{(j)}, \mathbf{h}^{(j)}, \sigma^{((j)})_{j=1}^K$ we discard an initial K_0 of the samples to address burn-in of the chain and we thin the remaining samples by a small factor γ to obtain independent samples from the joint posterior given in equation (2.7). We will specify the burn-in and thinning settings whenever we discuss applications of the method.

2.3.3 Posterior Sampling with Unknown Assignment Probabilities

There are two additional problems with respect to inference when the assignment probabilities are unknown: one needs to estimate the assignment probabilities $\boldsymbol{e}$ and use these to compute the TRV $\mathbf{Y}^*$. The following target posterior distribution corresponds to the model when the treatment probabilities vary across individuals as specified by the probit model in (2.6).

$$\pi(\mathbf{g}, \mathbf{h}, \mathbf{Y}^*, \sigma^2, \boldsymbol{e}, \boldsymbol{\beta} \mid \mathcal{D}). \tag{2.9}$$

In this setting the joint posterior is more complicated than equation (2.7) and is harder to sample from.

For the full posterior stated in equation (2.9) a standard Gibbs sampling procedure cannot be specified for sampling the treatment assignment probabilities. For the full posterior given in equation (2.9), we use a näive approach to sampling the assignment probabilities in addition to the other model parameters with an additional Metropolis-within-Gibbs step. This results in the following procedure:

a) Initialize $\mathbf{h}^{(0)}$, $\sigma^{(0)}$, $\mathbf{g}^{(0)}$, and $\boldsymbol{\beta}^{(0)}$. Use $\boldsymbol{\beta}^{(0)}$to compute $\boldsymbol{e}^{(0)}$;

b) Compute $\mathbf{Y}^*$ from the initial $\boldsymbol{e}^{(0)}$ and data;

c) For $j = 1, ..., K$

   1) $\mathbf{g}^{(j)} \sim \pi(\mathbf{g} \mid \mathbf{h}^{(j-1)}, \sigma^{(j-1)}, \boldsymbol{e}^{(j-1)}, \mathbf{Y}^*, \mathcal{D})$;

   2) $\mathbf{h}^{(j)} \sim \pi(\mathbf{h} \mid \mathbf{g}^{(j)}, \sigma^{(j-1)}, \boldsymbol{e}^{(j-1)}, \mathbf{Y}^*, \mathcal{D})$;

   3) $\sigma^{(j)} \sim \pi(\sigma \mid \mathbf{h}^{(j)}, \mathbf{g}^{(j)}, \boldsymbol{e}^{(j-1)}, \mathbf{Y}^*, \mathcal{D})$;

   4) Use Metropolis-Hastings step to sample $\boldsymbol{\beta}^{(j)}$;

   5) Compute $\boldsymbol{e}^{(j)}$ from $\boldsymbol{\beta}^{(j)}$ and data;

6) Compute $\mathbf{Y}^*$ from $\boldsymbol{e}^{(j)}$ and data.

The Metropolis-Hastings step consists of specifying a proposal distribution $q(\boldsymbol{\beta})$, and given a candidate value $\boldsymbol{\beta}^* \sim q(\boldsymbol{\beta})$ is accepted with acceptance probability

$$\alpha = \min\left(1, \frac{\pi(\mathbf{g}, \mathbf{h}, \mathbf{Y}^*, \sigma^2, \boldsymbol{\beta}^*, \boldsymbol{e} \mid \mathcal{D})\, q(\boldsymbol{\beta})}{\pi(\mathbf{g}, \mathbf{h}, \mathbf{Y}^*, \sigma^2, \boldsymbol{\beta}, \boldsymbol{e} \mid \mathcal{D})\, q(\boldsymbol{\beta}^*)}\right). \tag{2.10}$$

where the posterior $\pi(\mathbf{g}, \mathbf{h}, \mathbf{Y}^*, \sigma^2, \boldsymbol{\beta}^*, \boldsymbol{e} \mid \mathcal{D})$ is specified in Sections 2.3.1 and 2.3.1.

Joint Bayesian modeling and the collusion problem The joint Bayesian model specified in this chapter for learning the assignment mechanism $\boldsymbol{e}$ and the transformed outcome $\mathbf{Y}^*$ leads to a feedback problem of the type described in [95]. The treatment assignment probability $\boldsymbol{e}$ appears in the joint posterior distribution both as a part of the transformed outcome model through 2.8 as well as its own model in 2.6. Therefore its posterior samples involve information from both. In the specific context of the assignment model, this means that the posterior samples of parameters in learning $\boldsymbol{e}$ are informed by information from the outcome stage.

Under the classical method of using $\boldsymbol{e}$ as a dimension reduced covariate representation in the outcome stage model (an analog to our transformed outcome), [95] demonstrate that the estimation of causal effects is poor. There is a possibility of considerable bias due to the distortion of the causal effects. Furthermore, the usefulness of the propensity score adjustment as a replacement for the covariates is also compromised.

However, this is not the concern in the modeling scheme proposed in this dissertation. [95] show that the nature of the feedback between the two stages is altered when the outcome stage model is augmented with adjustment for the individual covariates and that this method can recover causal effects akin to when a classical two stage procedure is used. Our approach via the kernels of the Gaussian processes provides individual covariate adjustment therefore alleviating concerns created by the feedback. Therefore we garner the benefits of the joint estimation, but by means of suitably elicited priors, and individually controlled covariates, we bypass the concerns of feedback. In fact, by making more complete use of the data, we are arguably able to improve the overall quality of estimation.

2.4 Results on Simulated and Real Data

In this section we validate our Gaussian process based TRV model on simulated and real data. We use the simulations to show that our approach outperforms other techniques (both TRV as well as conditional mean regression type methods). This holds true both when the treatment assignment probabilities are known or need to be inferred from the data. We also observe on the simulated data that our model does in fact recover the causal effects in the TRV framework in the presence of feedback as theorized earlier. Our assertion is based on comparisons of mean squared error, bias and point-wise coverage of the uncertainty intervals generated by the model, which are computed as follows,

$$\begin{aligned}
\text{Mean Squared Error} &= \frac{1}{n}\sum_{i=1}^{n}(\tau(X_i = x) - \tau(\widehat{X_i = x}))^2 \\
\text{Bias} &= \frac{1}{n}\sum_{i=1}^{n}(\tau(X_i = x) - \tau(\widehat{X_i = x})) \\
\text{Coverage} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\tau(X_i = x) \in [\tau(X_i = x)_{lwr}, \tau(X_i = x)_{upr}])
\end{aligned}$$

The real data analyzed here comes from a study of the causal effects of debit card ownership on household spending in Italy [59] – we will refer to these data as the SHIW data. In the analysis of the SHIW data we jointly infer treatment effects as well as the treatment assignment probability for each individual, as these are not observed.

The most interesting aspect of our analysis of the SHIW data is that we are able to identify heterogeneity in the treatment effects. We find that the impact of debit card usage on aggregate household spending is found to vary based on income and this variability is highest at the lowest levels of income – a notion that is validated under behavioral economic theory which further lends credibility to our proposed model.

2.4.1 Estimands Used and Modelling Approaches Compared

In this section we state the estimands that we will use for comparing our method to other non-parametric methods. We will also state in detail how we compute the relevant estimand for both our method and the other techniques considered. The analysis is focused on the estimation of the CATE.

Gaussian process mixture model: We first specify the procedure we use to estimate the CATE for our model. The model is trained on data $(X_1, ..., X_n)$ and the values of the two functions are

$$\mathbf{g} = (g(X_1), \ldots, g(X_n)),$$
$$\mathbf{h} = (h(X_1), \ldots, h(X_n)).$$

We will use the function values to evaluate the accuracy of our estimators.

Depending on whether the treatment assignment probabilities are observed or not we obtain posterior samples $\left(\mathbf{g}^{(j)}, \mathbf{h}^{(j)}\right)_{j=1}^{K}$ or $\left(\mathbf{g}^{(j)}, \mathbf{h}^{(j)}, \mathbf{e}^{(j)}\right)_{j=1}^{K}$, respectively, using which we can compute posterior samples for the conditional average treatment effect at each location X_i, $i = 1, ..., n$ as

$$\tau_i^{(j)}(X_i) = g^{(j)}(X_i)$$

Given the posterior samples we can compute a posterior mean as a point estimate, $\widehat{\tau(X_i)}$, as well as credible intervals. Where applicable, marginalizing over the values X_i allows us to compute posterior estimates of the average treatment effect.

Summary of alternative methods used: We will compare our proposed Gaussian process mixture model approach to other regression based methods for estimating treatment effects. We have considered random forests and single regression trees for treatment effect estimation via TRV modeling as well as *fit based trees*, *causal trees* [4], and BART[36, 41] as non-TRV alternatives [1].

[1]We use the implementations of these methods in the `R` packages `causalTree` [83], `rpart`[55], `randomForest`[85] and `BART`[68]

None of the aforementioned methods have an obvious framework for learning the treatment assignment probabilities internally. This a crucial step in computing the CATE and ATE both via TRV and non-TRV based estimation techniques. In the case of the regression trees and random forests for TRV modeling, the TRV needs to be computed from the learnt propensity score first before any modeling can commence. The BART model uses the propensity score as an additional covariate, while causal and fit based trees use the propensity score as a weighting mechanism.

Therefore, we will use a two-step procedure where we first use the data to infer the treatment assignment probabilities and then given these estimates, apply the aforementioned regression methods to estimate the treatment effect. The treatment assignment probability vector $\boldsymbol{e}$ is estimated via logistic regression [78], a standard approach for estimating propensity scores in the causal inference literature.

2.4.2 Results on Simulated Data

The objective of the simulation studies presented in this section is to compare the performance of the Gaussian process mixture model to, BART, causal trees, fit based trees, the random forest and single regression tree models. We consider two criteria in our comparison. The first criteria is a comparison of the accuracy of the CATE, in terms of mean squared error and bias. The second criterion involves assessing how well the methods model uncertainty by considering the coverage of the uncertainty intervals produced by all the models.

Simulated Data Model

In order to evaluate the proposed model as well as the other aforementioned approaches, we consider two simulation settings – one high dimensional case (with 40 covariates) and one low dimensional case (with 5 covariates) each with its own covariate level heterogeneity and a sample size of $n = 250$. For the remainder of this analysis, the high dimensional case is referred to as Case A, and the low dimensional case is referred to as Case B. By design neither of these simulation cases has a meaningful average treatment effect. We start with a detailed description of Case A.

In this framework, covariates $X_1, \dots X_{30}$ are independent covariates, $X_{31}, \dots X_{35}$ depend on pairs of covariates, while $X_{36}, \dots, X_{40}$ depend on groups of three as follows,

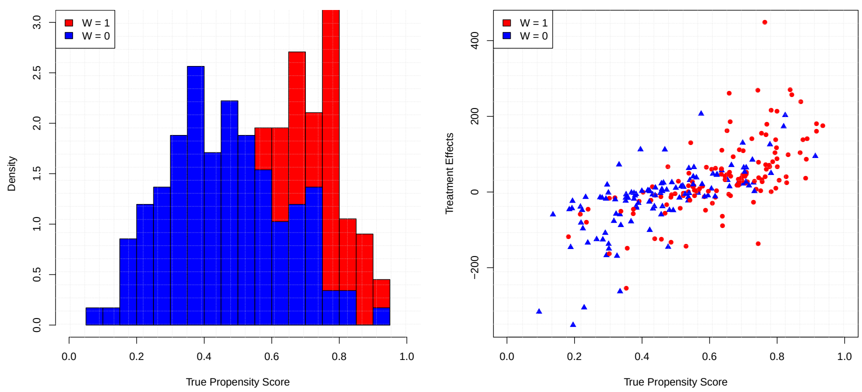
$$
\begin{aligned}
&X_k \sim \text{Normal}(0,1); \quad k = 1, \dots, 15 \\
&X_k \sim \text{Uniform}(0,1); \quad k = 16, \dots, 30 \\
&X_k \sim \text{Bernoulli}(q_k); \quad q_k = \text{logit}^{-1}(X_{k-30} - X_{k-15}); \quad k = 31, \dots, 35 \\
&X_k \sim \text{Poisson}(\lambda_k); \quad \lambda_k = 5 + 0.75 X_{k-35}(X_{k-20} + X_{k-5}); \quad k = 36, \dots, 40
\end{aligned}
$$

Next, we simulate the propensity score and the corresponding treatment assignments. This has been done as a simple linear transformation since the focus of the chapter is not propensity score modeling but rather CATE modeling. The propensity scores and the treatment effects of interest for Case A are given in Figure 2.1.

$$
\begin{aligned}
&p_i = \text{logit}^{-1}(0.3 \sum_{k=1}^{5} X_k - 0.5 \sum_{k=21}^{25} X_k - 0.0001 \sum_{k=26}^{35} X_k + 0.055 \sum_{k=36}^{40} X_k) \\
&W \sim \text{Bernoulli}(p_i)
\end{aligned}
$$

Finally we generate the potential outcomes and the observed outcomes.

$$
\begin{aligned}
f(\mathbf{X}) &= \frac{\sum_{k=16}^{19} X_k \exp(X_{k+14})}{1 + \sum_{k=16}^{19} X_k \exp(X_{k+14})} \\
Y(0) &= 0.15 \sum_{k=1}^{5} X_k + 1.5 \exp(1 + 1.5 f(\mathbf{X})) + \epsilon_i \\
Y(1) &= \sum_{k=1}^{5} \{2.15 X_k + 2.75 X_k^2 + 10 X_k^3\} + 1.25 \sqrt{0.5 + 1.5 \sum_{k=36}^{40} X_k} + \epsilon_i \\
Y &= WY(1) + (1-W)Y(0); \quad \epsilon_i \overset{\text{IID}}{\sim} \text{Normal}(0, 0.0001)
\end{aligned}
$$



(a) (b)

Figure 2.1: Summary plots of Case A (a) Histogram of the true propensity scores for each of the two treatment groups. (b) Treatment effects. The simulation was generated with $n = 250$.

The lower dimensional case, which we have adapted from the simulation study in [36] is presented similarly. We start by simulating the following $p = 5$ covariates.

$$X_k \sim \text{Normal}(0, 1); \quad k = 1, \dots, 3$$
$$X_4 \sim \text{Bernoulli}(p = 0.25)$$
$$X_5 \sim \text{Binomial}(n = 2, p = 0.5)$$

In this scheme, unlike Case A, all the covariates are independent. The propensity score model analogous to the previous case is a linear transformation of the covariates.

$$p_i = \text{logit}^{-1}(0.1X_1 - 0.001X_2 + .275X_3 - 0.03X_4)$$
$$W \sim \text{Bernoulli}(p_i)$$

Finally we generate the potential outcomes and the observed outcomes. The results of this simulation are presented in Figure 2.2.

$$f(\mathbf{X}) = -6 + h(X_5) + |X_3 - 1|$$

$$h(0) = 2, \quad h(1) = -1, \quad h(2) = -4$$

$$Y(0) = f(X) - 15X_3 + \epsilon_i$$

$$Y(1) = f(X) + (1 + 2X_2X_3) + \epsilon_i$$

$$Y = WY(1) + (1-W)Y(0); \quad \epsilon_i \overset{\text{IID}}{\sim} \text{Normal}(0, 0.0001)$$

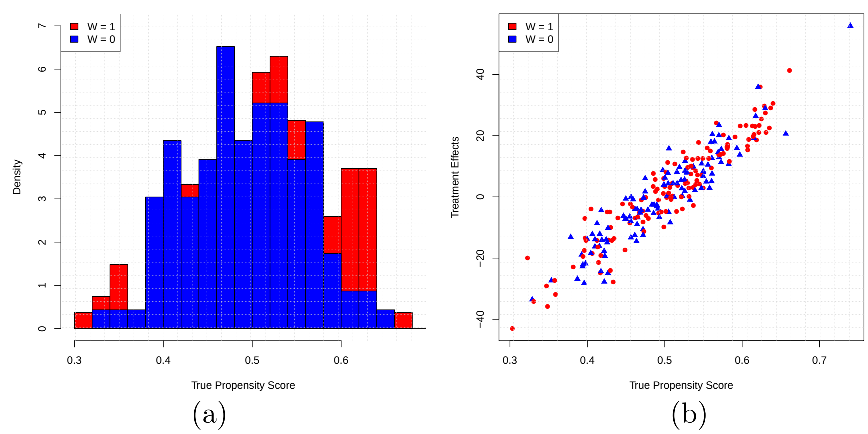


Figure 2.2: Summary plots of Case B (a) Histogram of the true propensity scores for each of the two treatment groups. (b) Treatment effects. The simulation was generated with $n = 250$.

Comparison of Methods

The first stage of our analysis compares the CATE estimation in instances when the treatment assignment probability is assumed to be known. We focus on the mean squared error, bias and coverage of the CATE under Case A and Case B along with visual analyses of model adaptability to gauge fit quality. For the proposed model the samplers were run for $K = 6,000$ steps with $1,000$ initial steps burned off. No thinning of the samplers was needed. Similarly, for the non-Bayesian methods, $K = 5,000$ replications of the bootstrap were generated. The comparison of point estimates of the CATE under Case A is presented in Figure 2.3 and Case B in 2.4 for the sub-case where the treatment assignment mechanism is known; the corresponding diagnostic measures are presented in Tables 2.1 and 2.2.

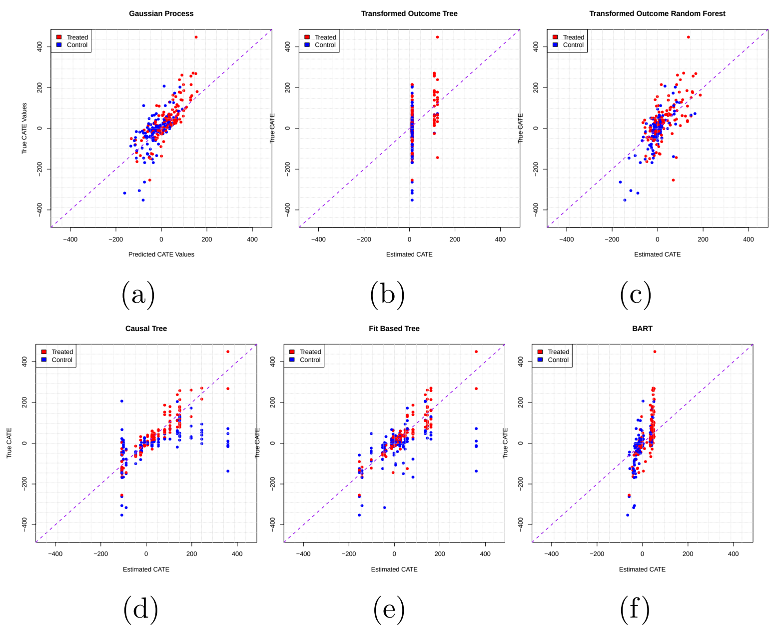


Figure 2.3: Comparison of the CATE estimates when the treatment probabilities are known for Case A (a) the GP mixture model (b) the transformed outcome regression tree (c) the transformed outcome random forest (d) the causal tree (e) fit based tree (f) BART

In Case A, both in terms of point estimation, as well as uncertainty quantification, we can conclude that when the treatment assignment is known, the proposed model is the overall winner. As we can see, it adapts well to the heterogeneity of the treatment effects in the data, and is able to recover the effects to a high degree as observed in Figure 2.3(a). It also has the lowest mean squared error of the models presented and the point-wise coverage of its uncertainty intervals, while low relative to tree based methods, is better than BART (see Table 2.1). Furthermore, the bias of the model is generally lower causal trees, fit based trees and transformed outcome trees.

In Case B, the model performs well in terms of recovering the high degree of heterogeneity but it suffers in terms of mean square error and bias. The model still adapts well to the heterogeneity inherent in the data, and is able to recover the effects as observed in

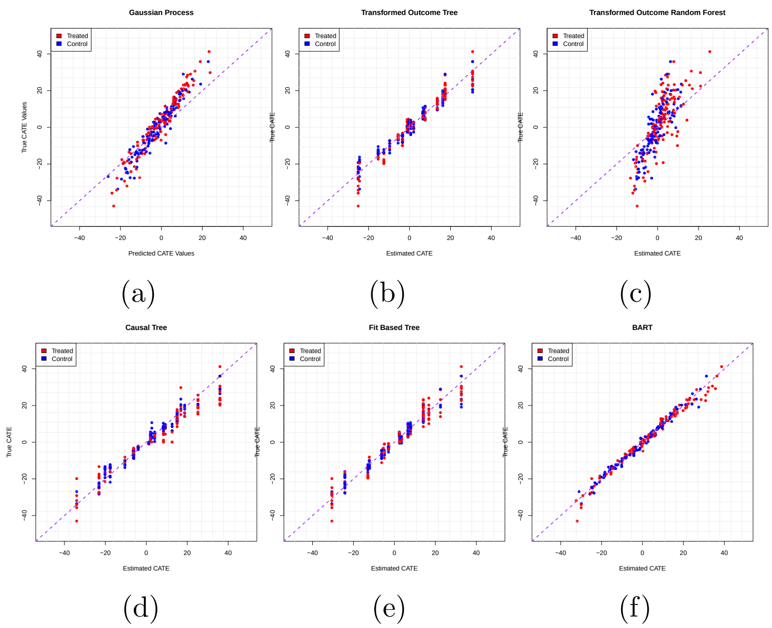


Figure 2.4: Comparison of the CATE estimates when the treatment probabilities are known for Case B (a) the GP mixture model (b) the transformed outcome regression tree (c) the transformed outcome random forest (d) the causal tree (e) fit based tree (f) BART.

Figure 2.4(a), albeit with a higher degree of overall noise. This noisiness translates to high mean squared error and bias, where the other alternative models perform better, with one minor caveat. Due to the piece-wise nature of the tree based models, they do not adapt to the heterogeneity as well as the proposed model and BART do. Furthermore, the model also has the highest degree of point-wise uncertainty interval coverage (see Table 2.2).

Table 2.1: Case A - Conditional Average Treatment Effect Summary (Known).

	Model Type	Mean Square Error	Bias	95% CI Coverage
1	Gaussian-Process Mixture	4191.665	13.207	0.780
2	Bayesian Additive Regression Tree	5856.135	-5.351	0.596
3	Transformed Outcome Tree	7769.077	14.374	0.876
4	Fit Based Tree	6154.396	15.633	0.928
5	Causal Tree	8390.039	21.923	0.964
6	Transformed Outcome Random Forest	4993.576	0.317	0.932

We also compare the CATE estimation for both cases when the treatment assignment probabilities are unknown and need to be inferred from the data. The comparison of the point estimation is given in Figures 2.5 and 2.6 respectively for the two cases, with the corresponding summary measurements of fit in Tables 2.3 and 2.4.

Table 2.2: Case B - Conditional Average Treatment Effect Summary (Known).

	Model Type	Mean Square Error	Bias	95% CI Coverage
1	Gaussian Process Mixture	50.262	3.174	0.988
2	Bayesian Additive Regression Tree	5.498	0.229	0.808
3	Transformed Outcome Tree	16.421	0.202	0.900
4	Fit Based Tree	15.620	0.282	0.952
5	Causal Tree	21.143	0.974	0.972
6	Transformed Outcome Random Forest	118.745	-0.582	0.816

We see that for Case A, the performance of the model is far superior in terms of adapting to the heterogeneity, as indicated in Figure 2.5(a), in particular compared to the performance of the transformed outcome random forest and BART given in Figures 2.5(c) and 2.5(f). The deterioration in the quality of the estimates from BART is particularly noticeable. Furthermore, while the point-wise coverage of the uncertainty interval is lower relative to the other models, the Gaussian process mixture is the clear winner in terms of the mean square error. The proposed model also outperforms the tree based models (causal and fit based trees as well as transformed outcome trees) in terms of bias (see Table 2.3) and its point-wise interval coverage is stable relative to BART, which speaks to the models overall robustness despite the added layer of complexity from learning the treatment assignments.

We see that for Case B, the results of the analysis are similar to when the treatment assignment was known. The performance of the model is comparable in terms of adapting to the heterogeneity relative to the other models, as indicated in Figure 2.6(a) – albeit again with a similar degree of noisiness as earlier. However, we again out-perform transformed outcome random forests in terms of point estimation with lower mean squared error. The only aspect in which the model out performs all the other methods considered is in terms of point-wise interval coverage.

Our conclusion is that the model performs well when there are a large number of covariates present, and the degree of heterogeneity in the treatment effects is high. The flexibility of the mixture of Gaussian processes ensures adaptability, where tree based models fail particularly when there is shared information in the covariates (as is true in Case A) since the

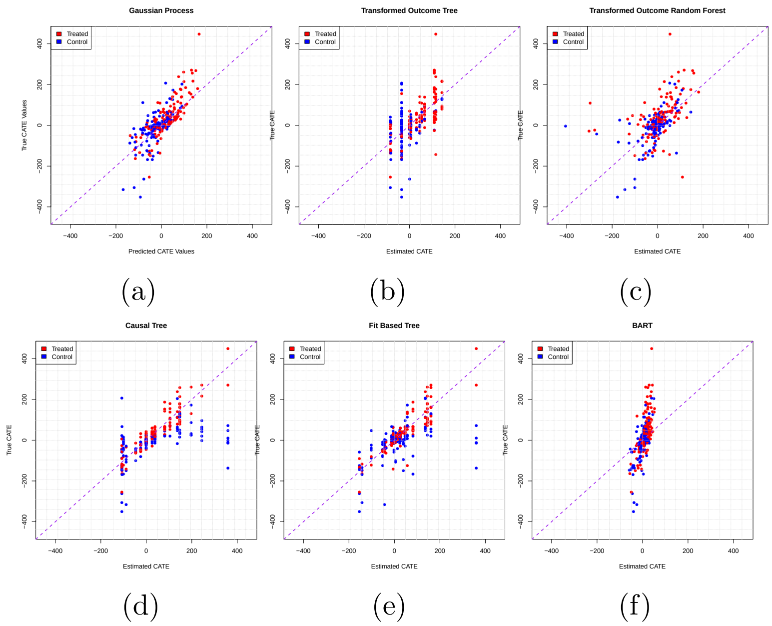


Figure 2.5: Comparison of the CATE estimates when the treatment probabilities are unknown for Case A (a) the GP mixture model (b) the transformed outcome regression tree (c) the transformed outcome random forest (d) the causal tree (e) fit based tree (f) BART.

prior provides built-in regularization. However, when the number of covariates is small, the flexibility of the model hurts its overall performance since we observe that our estimates are noisier than for instance BART. These limitations of the model are discussed as avenues for future work in the last section of this chapter.

Table 2.3: Case A - Conditional Average Treatment Effect Summary (Unknown).

	Model Type	Mean Square Error	Bias	95% CI Coverage
1	Gaussian Process Mixture	3916.562	13.207	0.780
2	Bayesian Additive Regression Tree	6754.058	-5.569	0.624
3	Transformed Outcome Tree	6289.891	7.061	0.880
4	Fit Based Tree	6154.396	15.633	0.932
5	Causal Tree	8390.039	21.923	0.968
6	Transformed Outcome Random Forest	12124.426	-21.958	0.960

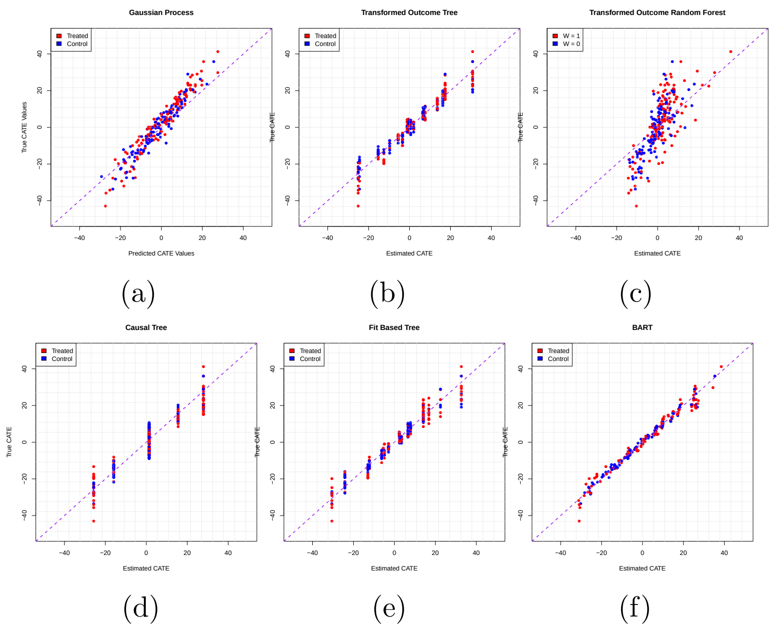


Figure 2.6: Comparison of the CATE estimates when the treatment probabilities are unknown for Case B (a) the GP mixture model (b) the transformed outcome regression tree (c) the transformed outcome random forest (d) the causal tree (e) fit based tree (f) BART.

Table 2.4: Case B - Conditional Average Treatment Effect Summary (Unknown).

	Model Type	Mean Square Error	Bias	95% CI Coverage
1	Gaussian Process	31.517	1.898	1.000
2	Bayesian Additive Regression Tree	6.259	0.118	0.776
3	Transformed Outcome Tree	16.421	0.202	0.892
4	Fit Based Tree	15.620	0.282	0.956
5	Causal Tree	19.652	0.876	0.972
6	Transformed Outcome Random Forest	115.329	-0.349	0.820

2.4.3 Results on the Italy Survey on Household Income and Wealth (SHIW)

Our application of the GP mixture model to a real data aimed at the estimation the causal effects of debit card ownership on household spending. A causal analysis of this question was developed in [59] using data from the Italy Survey on Household Income and Wealth (SHIW) to estimate the population average treatment effect for the treated (PATT). The SHIW is a biennial, national population representative survey run by Bank of Italy. The subset of the SHIW data we considered consists of $n = 564$ observations with 385 untreated and 179 treated observations. The outcome variable is the monthly average spending of the household on all consumer goods. The treatment condition is whether the household possesses one and only one debit card, and the control condition is that the household does not possess *any* debit cards. The covariates we used include: cash inventory held by the household, household income, average interest rate in the province where the household resides, measurement of wealth, and the number of banks in the province in which the household resides. See [59] for more details about the data. Our analysis of these data will consist of comparing estimates of the ATE and CATE (with respect to household income) of our GP mixture model to the same alternative models as the previous section.

Table 2.5: Conditional average treatment effect with average income by decile.

Income Decile	Average Income	$\widehat{\tau(x)}$	$\tau(x)_{lwr}$	$\tau(x)_{upr}$
1	-1.137	0.629	0.404	0.857
2	-0.831	0.567	0.374	0.761
3	-0.638	0.558	0.381	0.734
4	-0.472	0.459	0.298	0.620
5	-0.310	0.425	0.270	0.578
6	-0.114	0.396	0.245	0.546
7	0.103	0.343	0.190	0.490
8	0.397	0.272	0.097	0.441
9	0.848	0.172	-0.050	0.389
10	2.143	-0.125	-0.513	0.251

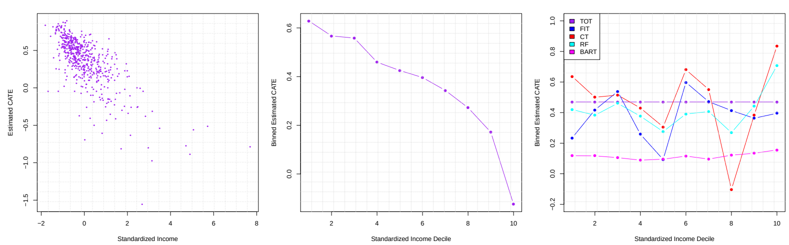
We start with a presentation of the CATE under our model against income in Figure 2.7(a).The proposed model estimates an overall downward trend in the effect of owning a debit card, i.e. as the level of income increases, the effect of owning a debit card declines. In order to summarize this effect, we consider the CATE for binned deciles of income for the proposed model in Figure 2.7(b) and the alternative models in Figure 2.7(c). We find that the proposed model detects a statistically meaningful effect for the first eight deciles of income, and this effect is estimated to decline in size. For the final two deciles, the model concludes that there is no statistically meaningful effect of owning a debit card. These results are summarized in Table 2.5. By comparison, the inference from the alternative approaches is not quite as clear. BART and transformed outcome trees, detect minimal heterogeneity, while transformed outcome random forests, transformed outcome trees and causal trees demonstrate the most heterogeneity at the highest two deciles of income. These results are summarized in Table .3 in Appendix .3.

In order to be comprehensive and garner comparability to past work, we have also produced estimates of the average treatment effect in Table 2.6. The proposed Gaussian process mixture detects a statistically meaningful ATE. This result is consistent with the findings of [59]. Furthermore, we also see that the uncertainty interval for the Gaussian process mixture is the tightest of the methods used here, all of which with the exception of BART garner similar inference. This result is consistent with the findings on simulated data presented in the last section since the BART model does not adapt to heterogeneity well in instances where the number of covariates is high with large contributions to the variation in the treatment effects. Again this argues that the GP mixture model may be outperforming the other methods.

Table 2.6: Comparison of average treatment effects.

	Model Type	$\widehat{\tau^p}$	τ^p_{lwr}	τ^p_{upr}
1	Gaussian Process Mixture	0.369	0.220	0.518
2	Transformed Outcome Tree	0.470	0.210	0.555
3	Fit Based Tree	0.378	0.214	0.608
4	Causal Tree	0.475	0.360	0.939
5	Bayesian Additive Regression Tree	0.115	-1.129	1.397
6	Transformed Outcome Random Forest	0.414	0.229	0.604

Based on the economic concepts of *income elasticity of demand*, *consumer choice* and *substitution effects* [87], the heterogeneity identified at the lowest levels of standardized income is a more sensible result relative to the implication of the other approaches. At the lowest levels of income, economic agents are more likely to substitute debit card use for cash in an effort to maximize spending. The debit cards act as an inflator of perceived financial resources and this effect is expected to diminish as the overall income grows. Therefore, the GP mixture model makes a more convincing case for capturing the true nature of how holding a debit card influences spending.



(a) (b) (c)

Figure 2.7: Estimated CATE (a) against Income (b) binned effects against deciles of income (b) binned effects against deciles of income for comparison.

2.5 Discussion and Future Work

We have proposed a novel non-parametric Bayesian model to estimate heterogenous treatment effects. Our approach combines the *transformed response variable* framework with a mixture of Gaussian-processes. The motivation for the GP mixture model was to improve the accuracy of our point estimates as well to better quantify uncertainty relative to other models particularly those from the Bayesian non-parametrics literature. We compared the performance of our model to a single regression tree and random forests model, within the TRV framework as well as two conditional mean regression type tree based methods and BART. We used simulation studies to show instances where our approach is a better estimator with respect to both point estimation and uncertainty quantification. Furthermore, our approach also has the advantage in that we can address the case where treatment assignment probabilities are unknown within our model; other methods require a two-stage process where another model is required to infer the treatment assignment probabilities. This tandem estimation provides better insight into the data generating process and also captures uncertainty from all levels of inference.

In addition, a Bayesian model of treatment effects with a single likelihood for the design and analysis stages creates concerns of feedback since the TRV depends on the assignment mechanism. We demonstrate that our model is robust to this feedback due to both our prior specification as well as individual covariate adjustment via the Gaussian process covariance functions. However, this raises the question of whether there is a weaker condition that can be satisfied and still lead to effective inference of treatment effects which is the first area that we aim to explore in future work.

There are several ways we can extend our model to be more robust and more flexible. The GP priors specified impose smoothness assumptions on the treatment effects that may not be reasonable. Relaxing the smoothness and using non-parametric models that have been developed to model dose-response curves may result in richer and more robust inference. Inference using the TRV is sensitive to the probability of receiving the treatment and misspecification of this probability systematically moves the casual estimate away from the true value, this bias can be large when the assignment probability is either very low or very large. This is the case since the variance of the mixture model is influenced by the reciprocal of the treatment assignment probability. Extending our model to be more robust to these extreme cases is vital.

Furthermore, we are currently fixing the hyper-parameter values within the kernels of the Gaussian-processes since attempting to learn these from the data creates two problems that we need to carefully study. First, learning these parameters is difficult from a sampling perspective since the target distributions are often extremely multi-modal. A promising avenue for addressing this is the use of a combination of sampling and optimization [52] – this is particularly important since Bayesian non-parametric methods are known to be sensitive to prior calibration. This is of particular importance in instances where the degree of heterogeneity in treatment effects is small as we have seen via simulation study. Second, the scalability of Gaussian processes is very limited [51] and hence increasing the number of parameters that we are learning hurts the scalability even more. Both of these are important questions that warrant further study.

Chapter 3

Evaluating the Effect of Residential Care on Self-Concept and Ego-Resilience: A Generalized Propensity Score Analysis with Clustered Data

3.1 Introduction

As a consequence of challenges and adversity in their personal histories, in global mental health research, orphaned children are considered to be a vulnerable, at risk population[9]. This is attributed to the lifestyle of the institutionalized and the social stigma of being orphaned in certain parts of the developing world. Past research in global mental health has provided a theoretical foundation for suggesting that orphan children demonstrate lower levels of *self-concept* and *ego-resiliency* along with greater personal trauma. Furthermore, these children tend to avoid attachment and show symptoms of anxiety [15, 79, 22, 13, 88].

In the developmental psychology literature that is relevant to global mental health, self-concept is characterized as an individuals' measure of confidence in their own abilities and their prospects for the future. Ego-resilience is characterized as an expression of psychological sturdiness, resourcefulness and personal flexibility that takes advantage of an individual's ability to recover in the face of adversity [8]. These concepts are vital in constructing social identity and perseverance in the face of overwhelming personal setbacks.

This dissertation studies the ego-resiliency and self-concept of orphaned children that are residents of the Udayan Ghar Program – a foster-care type surrogate residential care system in India. The system began in 1994 to offer residential care homes for orphaned and abandoned children and currently comprises of fourteen group residences and has served approximately four hundred occupants. We are interested in understanding the effects of staying in a group home on the ego-resiliency and self-concept of the occupants. Subject matter experts suggest that if children reside in the system for longer, their ego-resilience and self-concept outcomes are likely to be very different.

The treatment variable of interest is the number of months that a resident spends in the Udayan Ghar Program. This variable will be interpreted as a continuous *dosage*. Our objective in this analysis is to estimate a *dose-response function* that estimates the average treatment effect for the treated. We want to determine the causal effects of staying in the system on ego-resilience and self-concept as measured by ratings on the standardized Piers-Harris scale [63]. This approach can help uncover and quantify heterogeneities in the impact of the Udayan Ghar Program on its residents including those of specific demographic groups of interest (such as male versus female residents).

We use a parametric estimator of the dose-response function that is rooted in the Bayesian paradigm as necessitated by the complex structure of this data (see section 3.2 for details). The most frequently used methodology in the literature for estimating causal effects where treatments are binary is by the use of *propensity scores* [74]. We extended this to our continuous treatment by the use of the *generalized propensity scores* [42], which behaves as a summary statistic. It helps eliminate selection bias into various treatment levels and helps identify those individuals for whom it is difficult to construct counterfactual outcomes by imposing an overlap condition and to control for observed covariates in a more flexible manner [24].

Our analysis is augmented by a series of recent results [54, 94] on analyzing the type of multistage hierarchical data arising from the studies conducted at the Udayan Ghar Program. We garner insights into the effects of continuous treatments in the setting where the data has multiple stages of clustering. The first stage arises from children being placed into group homes, and the second, from each child being measured multiple times over the course of their stay in the system. By taking the multi-stage hierarchy of the data into account, we can guarantee the consistency of our treatment effect estimates [54] and tackle issues that arise because of variations at the cluster level that create heterogeneities in the exposure to the treatment which is then propagated to the outcome.

The inference of causal effects is founded upon a number of assumptions, the details of which will be provided at length in the subsequent sections [42]. Our key identifying assumption is *unconfoundedness* which implies that receiving the various dosages of the aforementioned treatment is random conditional on the observed covariates. In Chapter 4, We will present a novel strategy for assessing how violations of this assumption impact the inference for multi-stage clustering data with a continuous treatment regime that builds on related work by [14].

The methodology and results presented in this chapter make a number of significant contributions to both statistical methodology and global mental health. Our approach deals with the estimation of causal effects under a continuous treatment regime in a hierarchical setting using the generalized propensity scores. Past work in hierarchical settings has been restricted to binary treatment variables. Furthermore, since this type of analysis is founded upon the assumption of unconfoundedness, we present a novel strategy for assessing impacts of its violations when the generalized propensity score is employed that has hitherto been absent from the causal inference literature. Finally, understanding impacts of systems like Udayan Ghar Program is of vital importance to policy makers, donors as well as non-government organizations and the system itself. The numerous stakeholders in this program are all interested in ensuring the well-being and improved future prospects of orphaned children that call the program home.

The remainder of this chapter is organized as follows: section 3.2 provides a detailed account of the data, along with the sources of the aforementioned complications and some brief exploratory findings. The relevant notation for the causal framework and the aforementioned modeling are presented in section 3.3 along with the main results. Our novel approach to sensitivity analysis using simulated unmeasured confounders is presented in the next chapter in conjunction with its implementation on the Udayan Ghar data. Concluding remarks along with areas for future work are summarized in section 3.5.

3.2 The Udayan Ghar Program Data

The Udayan Ghar Program consists of fourteen homes in New Delhi and two satellite programs piloted in Jaipur and Kurukshetra, India – in all, including New Delhi, the program has served approximately four hundred children since its advent. The data for this analysis is the result of a collaboration between Duke University and the Udayan Ghar Program to help examine the physical and mental health ramifications of children within the fourteen homes across the city of New Delhi.

The data was collected between 2014 and 2018 by Duke University students volunteering with the Udayan Ghar Program. Children in each of the 14 homes in New Delhi were interviewed on various mental health outcomes during the summer months of each year. A total of 185 resident children were interviewed over the time horizon of the study. The homes are segregated by gender and the sample selected was meant to reflect the proportion of males versus females that exist in the population of interest. The children that were interviewed in 2015 comprised of all the children in 2014 that still resided in the program, along with new additions to the program. This collection strategy continued until 2018, which is the last point of data collection for the analysis that has been presented here [9]. This method of data collection induces a complex hierarchical structure. At the first level of the hierarchy, each of the children in the sample are measured multiple times. At the second level, the children are nested within their respective residential homes. Hence, any analysis must account for these two stages of nesting in the data.

We construct the *treatment* as the number of months at which each evaluation is made. This number is approximately near the one-year mark, but demonstrates randomness which allows us to model it as such. The relationship between the deciles of the treatment and the median value of the standardized Piers-Harris score are presented in Figure 3.1. As the treatment dosage increases, we see a gentle upward trend in the median standardized

Piers-Harris score within each dosage decile. Additionally, the male children exhibit a higher self-concept and ego-resilience overtime relative to their female counterparts, which indicates the potentially important gender specific heterogeneity that we want to study.

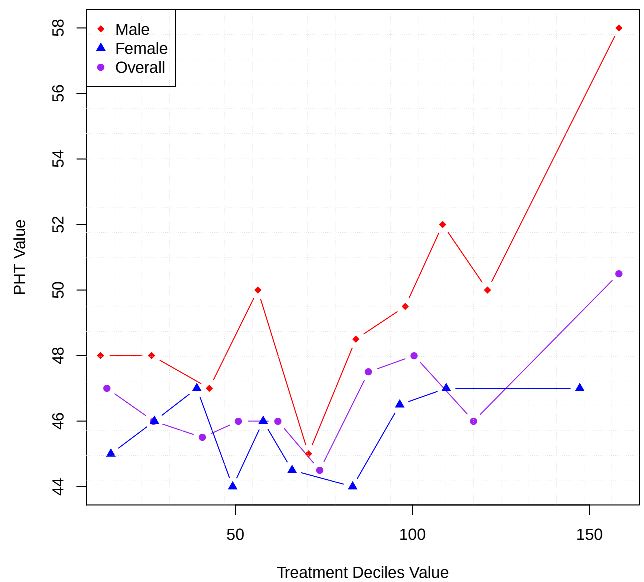


Figure 3.1: Exploratory plot comparing binned treatment deciles against the median standardized Piers-Harris score at each decile of the treatment variable. This plot shows the overall standardized Piers-Harris score as well as the score by gender.

The distribution of the treatment and its breakdown by gender are presented in Figure 3.2 along with its quantiles. These distributions are lightly right skewed. The treatments are evenly distributed between the genders with some important exceptions. There are more women than men in the sample and on average the women are in the system for a smaller period of time than the men (70 months for males versus 61 months for females). Furthermore, the men demonstrate a larger standard deviation in their treatment relative to the women (41 months for men versus 36 months for women). The overall ranges for the treatment are similar (2 months to 147 months for females versus 1 month to 138 months for males). This is an indicator of a larger concern with modeling multistage clustered data that warrants mention: since the homes are segregated by gender, there are likely home specific idiosyncrasies in treatment exposure propagated to the outcome of interest which necessitates accounting for the nesting in any model we use.

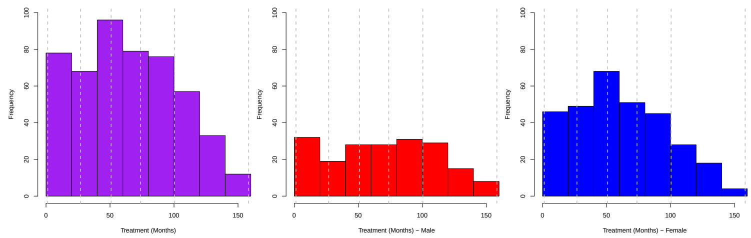


Figure 3.2: Exploratory analysis of the treatment variable . Panel (a) The distribution of the treatments for the entire sample. Panel (b) The distribution of the treatment for the males in the sample. Panel (c) The distribution of the treatment for the females in the sample.

At each psychological evaluation, measurements of the residents' biological features i.e. height and weight are also recorded along with their date of birth, trauma history, and age at entry into the system (in years). A thorough examination of these covariates, along with the single environmental variable measured, the ratio of residents to care-givers in each home is given in appendix .4.

The data has a considerable degree of missingness, particularly with respect to the trauma history and the biological features (see Table .2 in appendix .4 for details). Hence, we use multiple imputation using chained equations (MICE) [2] in an effort to best leverage the available data and note ignore observations due to partial missingness. Summary statistics for the outcome, treatment and covariates (all post-imputation) are given in Table 3.1 with a comparison to the original complete data in Table .3 (also in appendix .4). The data is on average relatively unchanged by imputation.

Table 3.1: Summary table of outcome, treatments and other covariates post imputation.

Variable Name	Mean (Imputed)	SD (Imputed)
Piers Harris (Standardized)	47.17	8.87
Treatment (Months)	64.40	38.24
Age at Entry (Yrs.)	7.78	2.82
Gender	0.62	0.49
Caregiver Ratio	5.20	1.48
Height (Centimeters)	151.23	12.65
Weight (Kilograms)	42.68	12.49
Total Traumatic Events	2.82	1.25

3.3 Estimation Strategy for the Dose Response Function of Staying in the Udayan Ghar System

3.3.1 Setup

The notation that we define here borrows from [42], and extends it to multistage clustering data. Let the homes in the analysis be indexed $k = 1, \ldots K$ with the children in each home being indexed $j = 1, \ldots N_k$ and the units of measurement for the j^{th} child in the k^{th} home be indexed $i = 1, \ldots N_{kj}$. Under this indexing, let $Y_{kji}(\omega)$ denote the potential outcome of a treatment $\omega \in \mathbb{W}$ where $\mathbb{W}$ is an interval corresponding to the continuous treatment. $Y_{kji}(\omega)$ can also be considered an individual dose-response curve. For each unit, we observe, a vector of pre-treatment covariates, X_{kji}, an observed treatment level W_{kji} and the observed outcome for the level of the treatment that is actually received, Y_{kji}. The causal estimand of interest is the average dose-response function: $\mu(\omega) = \mathbb{E}[Y_{kji}(\omega)], \quad \text{for any} \quad \omega > 0.$

Under continuous regimes an unconfoundedness assumption is needed for identification of treatment effects. [47] term this assumption as *weak unconfoundedness*,

Assumption 1.

$$Y_{kji}(\omega) \perp W_{kji} | X_{kj}; \quad \text{for all} \quad \omega \in \mathbb{W} \tag{3.1}$$

This grants that conditional on the observed pre-treatment covariates, the level of the treatment is independent of the potential outcomes. This eliminates any systematic self-selection into levels of the treatment based on unobservable characteristics. Given the pivotal reliance on this assumption, we will provide a novel strategy in section 4.3 for assessing how its violations impact our findings.

In addition to weak unconfoundedness, there is consensus that methods that control for pre-treatment observables perform poorly if there is insufficient overlap in the distribution of covariates by treatment status. We want to ensure that observations at each treatment level have comparable counterparts in *other* treatment levels. This assumption is concisely phrased as the *overlap* or *common-support* assumption.

Under the weak unconfoundedness assumption, the average dose-response function is obtained by estimating the average outcomes in subpopulations defined by the covariates and different levels of the treatment. In applications where there are a large number of covariates, controlling for them simultaneously can be difficult. The dimensionality problem under continuous treatments can be remedied using the *generalized propensity score* which is the conditional density of the treatment given the covariates,

$$r(\omega, x) = f_{W|X}(\omega|X = x) \tag{3.2}$$

and we define,

$$R_{kji} = r(W_{kji}, X_{kji}) \tag{3.3}$$

as the conditional density at a specific treatment receieved. Further, let $R^{\omega}_{kji} = r(\omega, X_{kji})$ – for those observations where $W_{kji} = \omega \rightarrow R_{kji} = R^{\omega}_{kji}$ i.e. the family of random variables indexed by the treatment level.

[42] demonstrate that the generalized propensity score has a *balancing property*,

$$X \perp \mathbf{1}_{W=\omega}|r(W, X) \tag{3.4}$$

which implies that within strata defined by the generalized propensity score, the probability of receiving the treatment does not depend on the level of the observed pre-treatment covariates. This balancing property in conjunction with the assumption of weak unconfoundedness implies that assignment to the level of the treatment is unconfounded given the generalized propensity score,

$$f_{W|X}(\omega|r(\omega,x),Y(\omega)) = f_{W|X}(\omega|r(\omega,x)) \tag{3.5}$$

Controlling for differences in the conditional density of receiving the treatment given the pre-treatment covariates will remove selection bias between various treatment levels conditional on observable features. We only need to adjust for the generalized propensity score, which behaves as a scalar summary of the (possibly high dimensional) pre-treatment covariates observed. This summarization in turn allows for the more flexible estimation of the treatment effects since we are not subject to high-dimensionality constraints.

In addition to its balancing property, using the generalized propensity score, can help determine the extent to which the overlap assumption is satisfied since it behaves as a scalar summary of the covariate distributions at various levels of the treatment. In section 3.3.3, we provide a strategy for utilizing the generalized propensity score in order to assess adherence to this assumption in a computationally tractable manner. This is vital since there are potentially infinite levels of the treatment by definition.

The properties of the generalized propensity score summarized above allow the estimation of the average dose-response function by utilizing it as a means of eliminating selection bias. Controlling for differences in the generalized propensity score will remove selection bias between individuals at various treatment levels if the treatment is based on observable features.

In this procedure, bias removal is attained in two steps, first, by estimating the conditional expectation of the outcome as a function of the observed treatment levels and the generalized propensity score, and second, by estimating the value of the dose-response function by averaging over the values of the generalized propensity score at each treatment level of interest. We express this as the following iterated expectations [42, 24],

$$\mu(\omega) = \mathbb{E}_R[\mathbb{E}_Y[Y_{kji}(\omega)|R^{\omega}_{kji} = r]] \tag{3.6}$$

Intuitively, the inner expectation represents the average potential outcome for each strata defined by the generalized propensity score. However, causal comparisons across different levels of the treatment are not valid, since for other treatment levels, the strata will be different – directly comparing the outcomes across the different treatment levels defines conditional expectations the difference of which is not a valid causal comparison. The outer expectation that averages over the conditional means over the distribution of the generalized propensity score grants valid causal comparisons.

Our procedure, accounting for the assessment of assumptions is summarized as follows:

1. Estimate generalized propensity score as the conditional density of the treatment on the pre-treatment covariates and check (i) overlap (ii) balance.
2. Estimate the conditional expectation of the outcome as a function of the observed treatment and the generalized propensity score.
3. Estimate the value of the dose-response function by averaging over the values of the generalized propensity score at each treatment level.

3.3.2 Model Specification

The iterated expectations in (3.6)[1] suggests the use of a partial means approach [24] to estimate the dose-response function at ω i.e. the average of a regression function over some of its inputs while holding others constant. In the application of interest in this dissertation, the regressor that is held constant in the outer expectation is the treatment dosage ω.

The theoretical constructs presented in the previous section require making a number of decisions regarding parameterizations and functional forms for the estimation of the objects

[1]Note that under the assumptions we have made $\mathbb{E}_Y[Y_{kji}(\omega)|R^{\omega}_{kji} = r] = \mathbb{E}_Y[Y_{kji}|W_{kji} = \omega, R_{kji} = r]$.

necessary for conducting inference. In this dissertation, we follow the implementation discussed in [42] and [24] adapted to the complexities of multi-stage clustered data. Our model specification is comprised of two stages, both using hierarchical Bayesian linear models [26] , first for the treatment stage, to estimate the generalized propensity score, and second for the outcome to estimate the dose-response function.

There is consensus that accounting for the multi-stage nesting of the data for both the stages of the modeling is necessary for reliable inference for three main reasons [54, 53]. First, when both the treatment assignment and the outcome generating mechanisms have a hierarchical structure, the estimator of interest is consistent (large sample unbiasedness) if at least one of the two stages takes the hierarchy into account. Second if this hierarchy is ignored when the clusters implied by it are small relative to the overall sample size, the propensity score may by poorly estimated and suffer from identifiability concerns. Third, there might be both measured and unmeasured confounders at the cluster level that create variations among clusters in the exposure to the treatment that are propagated to the outcome the effects of which have already become apparent in the exploratory analysis in section 3.2.

In order to estimate the generalized propensity score, we specify,

$$\begin{gathered} W_{kji} = \beta_0 + X_{kji}\boldsymbol{\gamma} + \tau_k + \nu_{kj} + \epsilon_{kji}; \\ \epsilon_{kji} \sim \mathrm{N}(0, \sigma^2_\epsilon);\;\; \nu_{kj} \sim \mathrm{N}(0, \sigma^2_\nu);\;\; \tau_k \sim \mathrm{N}(0, \sigma^2_\kappa) \end{gathered} \tag{3.7}$$

with priors,

$$\beta_{kj}|\beta_k, \sigma^2_\nu \sim \text{Normal}(\beta_k, \sigma^2_\nu); \quad j = 1, \ldots, n_j$$

$$\beta_k|\beta_0, \sigma^2_\kappa \sim \text{Normal}(\beta_0, \sigma^2_\kappa); \quad k = 1, \ldots, K \tag{3.8}$$

$$\beta_0 \sim \text{Normal}(0, \sigma^2_{\beta_0})$$

$$\gamma_{1,\ldots,p} \overset{\text{IID}}{\sim} \text{Normal}(0, \sigma^2_\gamma)$$

$$\sigma^2_\nu, \sigma^2_\kappa, \sigma^2_\epsilon, \sigma^2_{\beta_0}, \sigma^2_\gamma \sim \text{Inverse} - \text{Gamma}(a_W, b_W)$$

From (3.7), we can estimate $\hat{R}_{kji}$ and $\hat{R}^\omega_{kji}$ as follows,

$$\hat{R}_{kji} = \exp\{-\frac{\hat{\epsilon}^2_{kji}}{2\hat{\sigma}^2_\epsilon}\} \tag{3.9}$$

$$\hat{\epsilon}_{kji} = W_{kji} - \hat{W}_{kji}$$

$$\hat{R}^\omega_{kji} = \exp\{-\frac{(\omega - \hat{W}_{kji})^2}{2\hat{\sigma}^2_\epsilon}\} \tag{3.10}$$

We also specify a flexible parametric form for the the regression function of Y on the regressors W and the estimated generalized propensity scores $\hat{R}_{kji}$ with cubic polynomial terms [24],

$$Y_{kji} = \alpha_0 + W_{kji}\psi_1 + W^2_{kji}\psi_2 + W^3_{kji}\psi_3 + R_{kji}\psi_4 + R^2_{kji}\psi_5 + R^3_{kji}\psi_6 + W \cdot R_{kji}\psi_7 + \delta_k + \xi_{kj} + \varepsilon_{kji};$$

$$\varepsilon_{kji} \sim \text{N}(0, \sigma^2_\varepsilon);\;\; \xi_{kj} \sim \text{N}(0, \sigma^2_\xi);\;\; \delta_k \sim \text{N}(0, \sigma^2_\delta) \tag{3.11}$$

with priors,

$$\alpha_{kj}|\alpha_k, \sigma^2_\xi \sim \text{Normal}(\alpha_k, \sigma^2_\xi); \quad j = 1, \ldots, n_j$$

$$\alpha_k|\alpha_0, \sigma^2_\delta \sim \text{Normal}(\alpha_0, \sigma^2_\delta); \quad k = 1, \ldots, K \tag{3.12}$$

$$\alpha_0 \sim \text{Normal}(0, \sigma^2_{\alpha_0})$$

$$\psi_{1,\ldots,p} \overset{\text{IID}}{\sim} \text{Normal}(0, \sigma^2_\psi)$$

$$\sigma^2_\delta, \sigma^2_\xi, \sigma^2_\varepsilon, \sigma^2_{\alpha_0}, \sigma^2_\psi \sim \text{Gamma}(a_Y, b_Y)$$

The parameters of this model $\psi_1, \ldots \psi_7$ serve to help determine the shape of the dose-response function, the importance of which is paramount in assessing the impacts of violations of the unconfoundedness assumption. This approach to the model is a hierarchical extension of the parametric technique used in [42]. The joint distributions implied by the two models in (3.7) and (3.11) are given in appendix .5.

From the two models specified in (3.7) and (3.11), the dose-response function at ω is arrived at by averaging (marginalizing) (3.11) over the distribution of $\hat{R}^\omega_{kji}$ using the predictions from the outcome model at various values of the treatment,

$$\widehat{\mu(\omega)} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{j=1}^{N_k}\frac{1}{N_{kj}}\sum_{i=1}^{N_{kj}}\widehat{\mathbb{E}_R}[Y_{kji}|W = w, R^\omega_{kji}] \tag{3.13}$$

Since inference is conducted in the Bayesian paradigm, uncertainty quantification arises directly from the estimation procedure. We summarize the MCMC algorithm as follows:

Set initial values: $\beta_0^{(0)}, \boldsymbol{\gamma}^{(0)}, \alpha_0^{(0)}$, and $\boldsymbol{\psi}^{(0)}$

For $t = 1, \ldots, T$:

- S1 Sample $\beta_0^{(t)}$ from $f(\beta_0|\boldsymbol{\gamma}^{(t-1)}, \alpha_0^{(t-1)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y})$ and $\boldsymbol{\gamma}^{(t)}$ from $f(\boldsymbol{\gamma}|\beta_0^{(t)}, \alpha_0^{(t-1)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y})$
  - S1(a) Sample the GPS$^{(t)}$
- S2 Sample $\alpha_0^{(t)}$ from $f(\alpha_0|\boldsymbol{\gamma}^{(t-1)}, \beta_0^{(t)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y})$ and $\boldsymbol{\psi}^{(t)}$ from $f(\boldsymbol{\psi}|\beta_0^{(t)}, \alpha_0^{(t)}, \boldsymbol{\gamma}^{(t)}, \mathbf{X}, \mathbf{W}, \mathbf{Y})$
- S3 Estimate the dose-response function $\widehat{\mu(\omega)}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{j=1}^{N_k}\frac{1}{N_{kj}}\sum_{i=1}^{N_{kj}}\widehat{\mathbb{E}_R}[Y_{kji}|W = w, R^\omega_{kji}]^{(t)}$

3.3.3 Estimation of the Generalized Propensity Score

The estimated generalized propensity score model is the basis for adjusting for selection bias into different dosages of the treatment. The coefficients for treatment level models are given in Table 3.2 for the three cases that we investigate - the overall effect of residing in the Udayan Ghar system, and the effect by gender.

Table 3.2: Point estimates (posterior means) and 95% credible intervals for all the unit level predictors in the propensity score model.

Analysis Type	Variable Name	Point Estimate	Lower	Upper
	Height	0.907	0.717	1.098
Overall	Weight	0.438	0.218	0.660
	Age at entry	-7.724	-8.879	-6.594
	Gender	-5.381	-17.881	2.885
	Total traumatic events	1.024	-1.285	3.389
	Height	1.125	0.841	1.417
Male	Weight	0.187	-0.147	0.520
	Age at entry	-7.913	-10.201	-5.582
	Total traumatic events	1.3363	-2.647	5.622
	Height	0.754	0.517	1.001
Female	Weight	0.650	0.370	0.928
	Age at entry	-7.569	-8.856	-6.221
	Total traumatic events	0.157	-2.466	2.802

Under our model, we can see that some of the observed features significantly contribute to treatment self-selection. Overall, we see that the taller, and heaver children are expected to be in the system for longer and women are expected to be in the system for shorter periods than men along with those children that enter the system at a later age. In addition, the model suggests that children that have been through more traumatic events are more likely to remain in the system for longer. The gender and and traumatic events in the data are not statistically significant predictors. it also warrants mention that both height and weight for men are more impactful as predictors of treatment than for women.

From the discussion presented in 3.3.1 an important characteristic of the estimated generalized propensity scores that needs to be assessed is its *balancing* property. The balancing property can be leveraged as an empirical assessment of the adequacy of our estimates of the generalized propensity score – this is in the same vein as balance assessments performed for binary treatments. The approach we employ is methodologically derivative of ideas used in [42, 46].

The balancing property of the generalized propensity score is assessed separately for each covariate. We evaluate the explanatory power of the treatment variable on each covariate above and beyond the generalized propensity score. We employ hierarchical regressions[2]

[2]Hierarchical regressions are used to ensure that we correctly infer the standard errors of the coefficients which are crucial in the assessment of significance.

of each covariate on the treatment with and without the generalized propensity score and compare the significance of the coefficient of the treatment variable. If the generalized propensity score sufficiently balances the covariate of interest, then conditional on it, the treatment should not have any added explanatory power, and should therefore yield a non-significant coefficient. The results of the balancing, before and after controlling for the generalized propensity score are summarized in Table 3.3, using 95% credible intervals based on the models in section 3.3.2.

Table 3.3: Results of balancing before and after the accounting for the generalized propensity score with 95% credible intervals for the coefficient of the treatment variables for each covariate in the sample.

Analysis Type		Height	Weight	Age at entry	Gender	TTE	CGR
Overall	lower	0.1819	0.1238	-1.5e-05	-1.7e-04	-1.7e-03	-5.7e-03
	upper	0.2520	0.1921	1.7e-05	1.7e-05	1.4e-03	2.3e-03
Overall (with GPS)	lower	0.1911	0.1307	-1.8e-05	-1.8e-05	-1.6e-03	-4.3e-03
	upper	0.2625	0.2013	1.8e-05	1.7e-05	1.7e-05	4.3e-03
Female	lower	0.1646	0.1297	-2.6e-05		-3.4e-03	-2.9e-03
	upper	0.2547	0.2142	2.7e-05		1.3e-03	9.4e-03
Female (with GPS)	lower	0.1654	0.1295	-3.4e-05		-3.9e-03	-6.0e-03
	upper	0.2545	0.2178	3.1e-05		1.8e-03	5.8e-03
Male	lower	0.1638	0.0679	-4.5e-05		-4.2e-05	-0.0148
	upper	0.2869	0.1834	4.2e-05		4.2e-05	-0.0014
Male (with GPS)	lower	0.1548	0.0581	-5.1e-05		-5.1e-05	-0.0165
	upper	0.2665	0.1794	4.6e-05		4.9e-05	-0.0021

The results presented in Table 3.3 summarize that prior to adjustment for the generalized propensity score, the height and weight are unbalanced across the board, while the age at entry into the system, gender (in the overall model), total number of traumatic events and the caregiver ratios are balanced other than for the male children in the sample. Unfortunately, despite controlling for the generalized propensity score, the height and weight remain unbalanced in all three cases while the caregiver ratio remains unbalanced for the male subgroup only.

There are two possible methods that can be used to help correct this lack of balance. First, we considered a more aggressive trimming strategy of the type that has been employed in the related literature [24]. However, this is not a viable option since complete data is already limited. Instead as a hedge against confounding created by ignoring these variables we opted to add the height, weight and the caregiver ratio into the outcome model (3.11) therefore explicitly controlling for it in the estimation of the average dose-response function.

Along with verifying covariate balance, it is vital that no values of the generalized propensity score are so extreme that observations with comparable values are impossible to find in the sample i.e. assessing the common support or overlap assumption. For observations that are not comparable in terms of their generalized propensity score, the inference will be unreliable as no comparable observations are available to help infer causal comparisons. We restrict the final analysis to those observations for which the common support condition is satisfied i.e. our results focus on a subsample of the observations where the values of the generalized propensity score are within a comparable range of values[19, 29, 48].

To assess the overlap condition, the classic approach comes from analyzing binary treatments using the propensity score. The overlap in the covariate distributions is gauged by looking at the distributions of the estimated propensity score in each treatment group and restricting the analysis to regions where the propensity score is commonly supported. In contrast, extensions to continuous treatments, checking the overlap conditions is not as straight forward, since there are a large (possibly infinite) number of treatment groups and therefore generalized propensity scores to compare. A somewhat simple extension of this idea is presented in [24], and the approach in this dissertation is closely related.

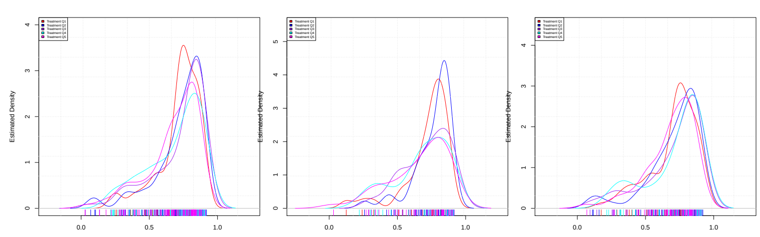


Figure 3.3: Density estimates with rugs for estimated generalized propensity score by binned treatment quintiles. Panel (a) The distribution of the generalized propensity score for the entire sample. Panel (b) The distribution of the generalized propensity score for the male children in the sample. Panel (c) The distribution of the generalized propensity for the female children in the sample.

We coarsen the range of the treatment into quintiles [75]. For each of these treatment quintiles that we denote $Q_1, \ldots Q_5$, we consider the ranges of the generalized propensity score that are estimated and *trim* observations that are not commonly supported. The ranges of the generalized propensity score within each treatment quintile are summarized in 3.4 and their estimated distributions are presented in Figure 3.3. In spirit, the approach we have presented here is similar to past work [24], albeit our approach yields a less aggressive trimming strategy which is necessary given the limited amount of complete data that we are working with.

Table 3.4: Ranges of the generalized propensity score within each of the treatment quintiles. The ranges overall and the ranges by gender are both tabulated here.

Analysis Type		Q_1	Q_2	Q_3	Q_4	Q_5
Overall	lower	0.214	0.073	0.032	0.217	0.028
	upper	0.912	0.917	0.915	0.912	0.918
Male	lower	0.140	0.276	0.276	0.218	0.029
	upper	0.902	0.913	0.916	0.912	0.904
Female	lower	0.171	0.114	0.091	0.215	0.072
	upper	0.910	0.918	0.917	0.913	0.899

It warrants mention here that individuals were not removed from the sample unless they either had missing data (as noted in section 3.2) or had a single measurement of the covariates that was deemed too extreme. In the trimming we have conducted, the latter has not been the case. Our trimming is restricted to observations that are not comparable across all five quintiles simultaneously. The resulting common-support restricted samples are summarized in Table 3.5. The results presented in this dissertation will be based on this smaller, trimmed version of the data.

Table 3.5: Summary of analysis of retained observations after overlap checks by trimming.

Analysis Type	Observation Trimming Summary	Individual Trimming Summary
Overall	482 observations from 499 observations	180 retained from 180 individuals
Male	176 retained from 190 observations	67 retained from 67 individuals
Female	278 retained from 309 observations	113 retained from 113 individuals

3.4 Inference on the Dose Response Function

3.4.1 Results

Using the model in the previous section, the dose-response function is estimated at 100 different values of the length of stay in the Udayan Ghar Program i.e. ω that correspond to the percentiles of the observed distribution of the treatment variable given in Figure 3.1.

We generate a series of plots in Figure 3.4 that illustrate the dose-response functions that we have estimated for three cases: (1) the overall dose response function (2) for male occupants only and (3) for female occupants only. The results that we have presented are accompanied by 95% credible intervals obtained as a byproduct of the Bayesian specification of the models employed. These are based on 10,000 posterior draws for the value of the average dose-response function at each treatment value. The intervals that we have provided account for uncertainty that arises from all stages of the analysis, including the estimation of the generalized propensity score.

Our results provide important insights into the efficacy of the Udayan Ghar Program. The overall average dose-response function for all the children residing in the Udayan Ghar Program is given in Figure 3.4(a). This dose-response function provides a few features of interest that warrant mention. First, over the window of observed treatments, the Piers-Harris score is expected to increase from ≈ 45 units to ≈ 51.5 units. From Table 3.6, we can see how this growth rate is not linear – there is higher growth in the standardized Piers-Harris score at the highest treatment quantiles (3, 4 and 5) moving from ≈ 46 to ≈ 48 units relative to the first two (1, 2 and 3) quantiles where we are stable at ≈ 45 units. We next decompose this behavior by gender which is a prime question for various stakeholders in the Udayan Ghar Program.

The model implies that males in the sample follow a more more monotone trajectory than the women. They start off at a higher point on the Piers-Harris scale ≈ 45 and finish at a higher level ≈ 52 as illustrated in Figure 3.4(b) relative to the females. Looking at the summary of growth provided in Table 3.6, we can see that the expected increments are

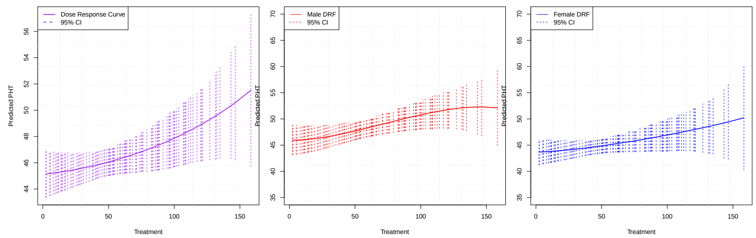


Figure 3.4: Panel (a) The overall dose-response function. Panel (b) The dose-response function accounting for men. Panel (c) The dose-response function accounting for females.

very consistent with time spent in the Udayan Ghar Program. The males are expected to show growth in ego resilience and self concept faster than the females in the sample since we start to see higher Piers-Harris scores. After the first two quantiles, we see a consistent growth of ≈ 1.5 units on the standard Piers-Harris scale in each treatment quantile.

By comparison, the female children in the sample are expected to have a slower response to the treatment. They experience improvement in the standardized Piers-Harris moving from ≈ 43 units to ≈ 51 units over the entire treatment range that we evaluate in this analysis. Like the male children in the sample they do experience consistent increments, albeit these are generally ≈ 1.5. This is clear from the average dose-response function in Figure 3.4(c). We see that the females experience *burn-in*, which aligns with past work in global mental health research in self-concept and ego-resilience for women. Considering this at a less fine grained level, in the first three treatment quantiles, the effects of the treatment on the averages of the Piers-Harris values are minor (a growth of approximately ≈ 0.7 units), while in the final two quantiles, the predicted Piers-Harris values grow more sharply.

We conclude that the Udayan Ghar Program has a net positive impact on the children that reside in their homes. There are differences in this effect that arise by gender in how the children respond to the residential care provided in terms of their ego-resilience and self-concept. However, these results are built upon an assumption of unconfoundedness that needs to be assessed in order to determine the reliability of our findings. The final

Table 3.6: Summary of averaged dose-response by treatment grid bins at average treatments in each quintile of the grid values. This clearly indicates the variation between the male and female children in the sample.

Quintile	Average Treatment	Piers-Harris$_{\text{Overall}}$	Piers-Harris$_{\text{Male}}$	Piers-Harris$_{\text{Female}}$
1	14.172	45.280	46.148	43.878
2	40.344	45.829	47.177	44.559
3	61.608	46.442	48.419	45.301
4	88.345	47.445	50.095	46.405
5	121.392	49.107	51.653	48.053

chapter of this disseratation, presents a strategy for accomplishing exactly this.

3.5 Discussion and Future Work

The objective of this dissertation is to determine whether there is some effect on self-concept and ego-resilience as measured by the Piers-Harris score that can be associated with residing in the Udayan Ghar Program. Our analysis concludes that there are positive psychological returns that can be attributed to spending time in the Udayan Ghar Program. Furthermore, we identify heterogeneities associated with the different lengths of stay in the Udayan Ghar Program across the male and female subpopulations. We find that the female children in the sample, compared to male children, experience a burn-in period in terms of their ego-resilience and self-concept as it is impacted by the program. For the male children, the dose-response function tends to be somewhat more linear, tending to show very consistent increments in their outcome of interest. There is however a large degree of overlap in the behavior of the two genders. Our findings are further validated by concrete theory in developmental psychology and global mental health [9] where similar trends have been proposed.

We employed a parametric estimator for the dose-response function based on the generalized propensity score that uses the continuous nature of the treatment. In our model specification, we controlled for features that are measured for each child as they begin their tenure at the Udayan Ghar Program including their age at entry and trauma history as well biological measurements such as height and weight in conjunction with environmental factors.

These results have two important policy implications. First, the estimated dose-response function indicates that there are clear benefits to children residing in the system for longer. Hence, retention efforts are mostly likely beneficial, particularly if the children are brought on at early ages into the program. Second, given the distinctive behavior of male and female children in the sample, a more targeted effort to ensure that female children are able to integrate better into the program is undoubtedly necessary, in order to minimize their experienced burn-in. Collectively with other on the ground efforts, this analysis can help understand and contribute to how NGO and government support of vulnerable and at risk populations in developing economies can be further supported and improved[9].

Despite the promise of our analysis, there are some important limitations that warrant mention, ordered here in terms of priority. First, we need to note that thus far we are accounting for a single household (i.e. environmental) level variable, the ratio of caregivers to pupils; Accounting for environmental factors outside of this might be a treasure-trove of statistical importance – these might potentially change which observations are considered to be too extreme and are therefore trimmed from the data. We could potentially improve the power of our analysis by incorporating this data especially if it uncovers other sources of treatment effect heterogeneity. Second, we need to attempt more flexible modeling. Parametric models such as the partial means via linear regression are limited in how much they can adapt to the data. At the core of the analysis we are still assuming two linear relationships – the first between the treatment and the pre-treatment covariates and second between the generalized propensity score and the outcome. This is a very strong assumption particularly in the context of the models at the outcome stage. Third, we need to develop more sophisticated flexible priors that can be adapted for observed data to help model unmeasured confounding. Finally, there are other subpopulations of interest to us that we currently do not have sufficient data for such as children that enter the program at various stages of psychological development proxied by age. These are some of the avenues that we aim to explore in future work.

Chapter 4

Sensitivity Analysis for Unmeasured Confounding with Continuous Treatments for Multistage-Clustered Data

4.1 Introduction

The assumption of unconfoundedness or ignorable treatment assignments is the key identifying assumption in causal inference under both binary and continuous treatment regimes. Colloquially speaking, it states that all potential confounders are measured and can therefore be controlled for. This assumption always holds in randomized experiments, but in observational studies, its violations can leave the robustness of the analysis open to criticism. This is an inherently untestable assumption, but rather than abandoning causal inference, we can assess the robustness of causal results to substantive violations by means of a sensitivity analysis as introduced in [73]. Our goal is to build confidence in the findings of observational studies by delineating between studies where the results are relatively robust against potential unmeasured confounders and instances when this is not the case.

Sensitivity analyses are generally approached in two ways: parametrically as exemplified in work by [14, 73] or non-parametrically as in [82, 32, 69]. Both strategies have their own merits and drawbacks, and we begin with a discussion of each of these.

Parametric sensitivity analysis techniques, suffer from concerns regarding their flexibility. By design, they are limited in that they only permit the exploration of sensitivity based on a range of assumptions built into the model that ties the unmeasured confounder to the treatment and the outcome stages of the analysis. This simplification of the relationship, may present an inaccurate depiction of how the unmeasured confounder impacts the two stages of causal inference. As a result, to a high extent, the sensitivity analysis conducted in this fashion will be subject to concerns of model misspecification. Furthermore, this may lead researchers to be more concerned with violations of ignorability than is warranted.

Non-parametric techniques on the other hand, rely on matching between control and treatment groups followed by non-parametric testing on the matched samples as exemplified in work by [82] to assess the sensitivity to unmeasured confounding. While these approaches avoid the strong assumptions of parametric methods for sensitivity analysis, they have their own limitations [14]. An overarching concept under this umbrella is the creation of matched samples that has the penchant for being an extremely laborious task, particularly in larger datasets with many features measured over many observations. Furthermore [40] assert that non-parametric techniques are sensitive to the choice of test statistic employed. Finally, there is an absence of work on non-parametric techniques for continuous treatment regimes where the matching step is not clearly defined.

Our data as introduced in chapter 3, has three complications that merit the proposal of a new technique. First, our application deals with a continuous treatment variable. Classic methods of sensitivity analysis such as those presented in [73] as well as Bayesian variants [58, 34] study the effects of unmeasured confounding in the context of binary treatments only. Second, our motivating application has multiple stages of clustering due to repeated measurements nested within homes which makes the specification of the model likelihood more complex (see appendix .5). While a basic framework for sensitivity analysis under continuous regimes is presented in [14] it neither utilizes the generalized propensity score, nor does it extend to hierarchical settings such as ours. Third, since the estimand of interest is the dose-response function rather than a single parameter, understanding the impact of violations of unconfoundedness requires a more involved evaluation strategy. Our technique in this chapter aims to address all three concerns.

Our focus in this chapter is on a parametric method that draws on the ideas of Bayesian sensitivity analysis presented in [58, 34, 14] which can be naturally extended to continuous treatment regimes in hierarchical settings. We propose the simulation of the unmeasured confounder from its conditional distribution and utilize its imputed values in the estimation of the treatment and outcome stages of the model. By using the *conditional* values of the unmeasured confounder, under various sensitivity parameters, we can capture the effects of unmeasured confounding on the shape of the dose-response curve. The impacts of unmeasured confounding can then be assessed first by visual comparison, and second by means of formal testing. For the latter, we propose a two fold strategy that combines Bayesian posterior simulation with formal non-parametric testing as well as ideas from posterior model comparison. This helps determine the impact of unmeasured confounding on the distribution of the causal estimand – the dose response function.

The approach that we present makes four important contributions to the sensitivity analysis literature. First, our focus is in this chapter is on continuous treatments where the estimand of interest the average dose-response curve which has hitherto been absent from the relevant literature. Second, there is very limited work thus far on sensitivity analyses in hierarchical settings, where the unmeasured confounder can manifest at various stages of the data. Third, our method is used in conjunction with the generalized propensity score which is also hierarchically estimated. Finally, we propose a new non-parametric evaluation strategy where the effects of the unmeasured confounding can be assessed by simulation from the posterior distribution of the dose-response function.

The remainder of this chapter is organized as follows: section 4.2 discusses the unconfoundedness assumption for continuous treatments and section 4.3 presents the setup for conducting the sensitivity analysis under our proposed technique. Using a simple simulation study, we demonstrate how simulation from the conditional distribution of the unmeasured confounder can be used to model its impacts on inference in section 4.4. Finally, we present results on the Udayan Care Data from Chapter 3, in section 4.5. Our conclusions and areas that we would like to explore in the future are discussed at length in section 4.6.

4.2 Unconfoundedness for Continuous Treatments

In a myriad of studies aimed at understanding some *outcome*, the focus is typically on one cause, i.e. the *exposure*. Along with the variables that measure the outcome and the primary exposure of interest, investigators collect information on other suspected causes or potential *confounders* to rule them out. We can characterize this relationship as follows, where $\mathbf{W}$ denotes the vector of treatment levels, $\mathbf{Y}$ denotes the vector of outcomes and $\mathbf{X}$ the measured confounders,

$$\begin{tikzcd} \mathbf{W} \arrow[rr] & & \mathbf{Y} \\ & \mathbf{X} \arrow[ul] \arrow[ur] & \end{tikzcd}$$

When analysis is conducted, this *common cause* i.e. $\mathbf{X}$ can be controlled for, thereby making causal inference possible since $\mathbf{Y}(\omega) \perp \mathbf{W}|\mathbf{X};\ \text{ for all }\ \omega \in \mathbb{W}$. This eliminates any systematic self-selection into levels of the treatment based on unobservable characteristics. In the language of directed graphs as shown above [62], conditioning on $\mathbf{X}$ indicates that the backdoor path from the exposure to the outcome that went through the common cause is blocked. Since this backdoor path is blocked, the confounding effect has been removed.

However, by its very nature, the process of data collection is extremely challenging since it is never possible to identify and measure all potential confounders, and some of these remain unmeasured, which gives the following alternative characterization,

$$\begin{tikzcd} \mathbf{W} \arrow[rr] & & \mathbf{Y} \\ & \mathbf{X}, \mathbf{U} \arrow[ul] \arrow[ur] & \end{tikzcd}$$

where $\mathbf{U}$ denotes an unobserved or unmeasured confounder. Therefore, the assumption rephrased for this setting is that $\mathbf{Y}(\omega) \perp \mathbf{W}|\mathbf{X}, \mathbf{U};\ \text{ for all }\ \omega \in \mathbb{W}$. By design however, while $\mathbf{X}$ can be controlled for in the analysis, $\mathbf{U}$ cannot, which means the backdoor path from exposure to outcome is not completely blocked. Failure to adequately account for $\mathbf{U}$ means that there is self-selection into levels of the treatment meaning that $\mathbf{Y}(\omega) \not\perp$

$\mathbf{W}|\mathbf{X}$; for $\text{all}\,\omega \in \mathbb{W}$. However, this assumption cannot be directly tested which necessitates sensitivity analysis and the next section details this technique.

4.3 Assessing Unconfoundedness

4.3.1 Setup

We propose a likelihood based technique akin to [50, 14], that we supplement with a substantive *prior* on the unmeasured confounder. This allows us to simulate its values from its full conditional distribution. This idea, termed as, Bayesian sensitivity analysis has been explored in the past in [58]. The remainder of the parameters can then be sampled conditional on the simulated values of the unmeasured confounder thereby capturing its influence on the inference we garner for the dose-response function. This concept is closely related to the data augmentation strategies for imputation discussed in [2]. Since this method relies on the likelihood, it can easily be generalized to the type of multi-level data that we are dealing with in our application to the Udayan Ghar Program.

This setup relies on specifying a joint distribution of $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{U}$ as defined in the previous section. We define and factorize this distribution as follows,

$$f(\mathbf{Y}, \mathbf{W}, \mathbf{U}|\mathbf{X}) = f(\mathbf{Y}|\mathbf{W}, \mathbf{U}, \mathbf{X}) f(W|\mathbf{U}, \mathbf{X}) f(\mathbf{U}|\mathbf{X}) \tag{4.1}$$

where $f(\mathbf{Y}|\mathbf{W}, \mathbf{U}, \mathbf{X})$ and $f(\mathbf{W}|\mathbf{U}, \mathbf{X})$ denote the likelihood over the two observed quantities $\mathbf{Y}$ and $\mathbf{W}$ incorporating the unmeasured confounding. We can, characterize two types of relationships for the unmeasured and measured confounders with $\mathbf{U}$, first, that $\mathbf{U} \not\perp \mathbf{X}$, meaning we specify $f(\mathbf{U}|\mathbf{X})$ which can be used to meaningfully characterize how the unmeasured confounder may be related to the observed $\mathbf{X}$ and second where $\mathbf{U} \perp \mathbf{X}$ for which we specify a marginal distribution $f(\mathbf{U})$. Under both distributions that act as *priors*, using the joint likelihood an implied full conditional distribution over the unmeasured confounder can be derived allowing us to simulate its values. Earlier work, such as that of [73] rely on integration to bridge the observed data likelihood to the full likelihood – we use conditional simulation in its stead. Once the values of the unmeasured confounder have been sampled, the remainder of the parameters can be sampled conditional on it. We will summarize this algorithm in context later.

The simplicity of parametric approaches can be advantageous since the effects of unmeasured confounding are specified by how $\mathbf{U}$ relates to $\mathbf{Y}$ and $\mathbf{W}$ in the likelihood. A choice has to be made in terms of the form in which $\mathbf{U}$ enters the likelihood, and second in terms of *sensitivity parameters* that are easy to interpret, and easy to calibrate. For our application, the effect of the unmeasured confounder is characterized by two sensitivity parameters: $\mathcal{L}_W$ which describes how the unmeasured confounder relates to the treatment assignment stage and $\mathcal{L}_Y$ that analogously describes the relationship of the unmeasured confounder to the outcome stage. We will consider two functional forms of $\mathbf{U}$ for our analysis along with a grid of sensitivity parameters.

The notation we use here is based on the data presented in the previous chapter, which we restate here for completeness. Let the homes in the analysis be indexed $k = 1, \ldots K$ with the children in each home being indexed $j = 1, \ldots N_k$ and the units of measurement for the j^{th} child in the k^{th} home be indexed $i = 1, \ldots N_{kj}$. Under this indexing, let $Y_{kji}(\omega)$ denote the potential outcome of a treatment $\omega \in \mathbb{W}$ where $\mathbb{W}$ is an interval corresponding to the continuous treatment. $Y_{kji}(\omega)$ can also be considered an individual dose-response curve. For each unit, we observe, a vector of pre-treatment covariates, X_{kji}, an observed treatment level W_{kji} and the observed outcome for the level of the treatment that is actually received, Y_{kji}. The causal estimand of interest is the average dose-response function: $\mu(\omega) = \mathbb{E}[Y_{kji}(\omega)], \text{ for any } \omega > 0$.

Based on the factorization presented in (4.1) and the sensitivity parameters, the impact of the unmeasured confounder on the treatment stage is specified as,

$$W_{kji} = \beta_0 + X_{kj}\boldsymbol{\gamma} + \mathbf{g}_W(U_{kji})\mathcal{L}_W + \tau_k + \nu_{kj} + \epsilon_{kji};$$

$$\epsilon_{kji} \sim \mathrm{N}(0, \sigma^2_\epsilon);\;\; \nu_{kj} \sim \mathrm{N}(0, \sigma^2_\nu);\;\; \tau_k \sim \mathrm{N}(0, \sigma^2_\kappa) \tag{4.2}$$

and impacts the outcome stage of the analysis via,

$$Y_{kji} = \alpha_0 + W_{kji}\psi_1 + W^2_{kji}\psi_2 + W^3_{kji}\psi_3 + R_{kji}\psi_4 + R^2_{kji}\psi_5 + R^3_{kji}\psi_6 + W \cdot R_{kji}\psi_7 + \mathbf{g}_Y(U_{kji})\mathcal{L}_Y + \delta_k + \xi_{kj} + \varepsilon_{kji};$$

$$\varepsilon_{kji} \sim \mathrm{N}(0, \sigma^2_\varepsilon);\;\; \xi_{kj} \sim \mathrm{N}(0, \sigma^2_\xi); \delta_k \sim \mathrm{N}(0, \sigma^2_\delta) \tag{4.3}$$

where $\mathbf{g}_Y$ and $\mathbf{g}_W$ are the functional forms of the unmeasured confounder that enter the likelihood under the hierarchical model in the previous chapter. The sensitivity parameters in our specified framework are very clearly interpretable as the conditional associations between the unmeasured confounder and regression function of the two stages of the analysis i.e. these are regression coefficients measuring how the treatment and control are expected to change for unit changes in $\mathbf{g}_W$ and $\mathbf{g}_Y$. When the two sensitivity parameters ($\mathcal{L}_W$ in (4.2) and $\mathcal{L}_Y$ in (4.3) are set to zero, we return to the original models in (3.7), and (3.11) – the case that we will use as a baseline. The sensitivity parameters we use will be selected based on the estimated coefficients for height and weight given in Table 3.2. The priors on the regression parameters are the same as specified in section 3.3.2 in Chapter 3.

We have considered two types of priors to characterize unmeasured confounding. We start first, assuming a single continuous unmeasured confounder at the observation level[1], for which we place a Gaussian distribution on $\mathbf{U}$,

$$U_{kji} \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2_U) \tag{4.4}$$

with the functions $\mathbf{g}_W(\cdot)$ and $\mathbf{g}_Y(\cdot)$ assumed to be cubic polynomials,

$$\mathbf{g}_W(U_{kji}) = a_1 U_{kji} + a_2(U_{kji})^2 + a_3(U_{kji})^3$$

[1]We have assumed that this unmeasured confounding is at the measurement level, but our approach can also be used for assessing the impacts of unmeasured confounding at the household or observation level.

$$\mathbf{g}_Y(U_{kji}) = b_1 U_{kji} + b_2(U_{kji})^2 + b_3(U_{kji})^3$$

This form was selected based on subject matter expertise from global health research where growth curves are used to characterize developmental and risk metrics for mental disorders. For instance, in our application of interest, neural development is an important feature that researchers have expressed an interest in understanding with respect to its effects on ego-resilience and self concept. It is anticipated to demonstrate rapid growth during youth, followed by pruning at later age, followed by degeneration in later years. Another example of similar but negatively related behavior is expected from the development of mental health disorders.

Second, we also study a single binary unmeasured confounder, again at the observation level, for which we place a logistic prior on $\mathbf{U}$,

$$\begin{aligned} U_{kji}|X_{kji} &\overset{\text{IID}}{\sim} \text{Bern}(\pi_{U_{kji}}) \\ \pi_{U_{kji}} &= \text{logit}^{-1}\{\mathbf{h}(X_{kji})\} \end{aligned} \tag{4.5}$$

where $\mathbf{g}_W(U_{kji}) = U_{kji} \cdot X_{kji}$ and $\mathbf{g}_Y(U_{kji}) = C \cdot U_{kji} \cdot W_{kji}$ and $\mathbf{h}(X_{kji})$ is a linear function. These forms are motivated by a key demographic variable that is currently missing in the data: psychological development subpopulations. Much like gender, these are anticipated to have impacts to the measures of self-concept and ego-resilience in conjunction with the treatment and the biometric data measured. Therefore, we have formulated the prior to be dependent on the developmental features. The impact of the unmeasured confounder at treatment stage is in conjunction with height and weight via an interaction term while its own behavior is determined by height and weight as well. At the outcome stage, the unmeasured confounder interacts with the treatment itself.

Under both of these approaches, the algorithm for simulating the dose-response function accounting for unmeasured confounding can be summarized as follows, Set initial values: $\beta_0^{(0)}, \boldsymbol{\gamma}^{(0)}, \alpha_0^{(0)}, \boldsymbol{\psi}^{(0)}$ and $\mathbf{U}^{(0)}$

For $t = 1, \ldots, T$:

S1 Sample $\mathbf{U}^{(t)}$ from $f(\mathbf{U}|\beta_0^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \alpha_0^{(t-1)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y})$

S2 Sample $\beta_0^{(t)}$ from $f(\beta_0|\boldsymbol{\gamma}^{(t-1)}, \alpha_0^{(t-1)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{U}^{(t)})$ and $\boldsymbol{\gamma}^{(t)}$ from $f(\boldsymbol{\gamma}|\beta_0^{(t)}, \alpha_0^{(t-1)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{U}^{(t)})$

S2(a) Sample the GPS$^{(t)}$

S3 Sample $\alpha_0^{(t)}$ from $f(\alpha_0|\boldsymbol{\gamma}^{(t-1)}, \beta_0^{(t)}, \boldsymbol{\psi}^{(t-1)}, \mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{U}^{(t)})$ and $\boldsymbol{\psi}^{(t)}$ from $f(\boldsymbol{\psi}|\beta_0^{(t)}, \alpha_0^{(t)}, \boldsymbol{\gamma}^{(t)}, \mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{U}^{(t)})$

S4 Estimate the dose-response function $\widehat{\mu(\omega)}^{(t)} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{N_k}\sum_{j=1}^{N_k} \frac{1}{N_{kj}}\sum_{i=1}^{N_{kj}} \widehat{\mathbb{E}_R}[Y_{kji}|W = w, R^{\omega}_{kji}]^{(t)}$

Appendix .6 shows the density of the full conditional distribution (up to a normalizing constant) for both cases represented by equations (4.4) and (4.5) that can be incorporated into the the sampling scheme to be used for inference.

Due to the complexity of assessing sensitivity to unmeasured confounding with continuous treatments we employ three means of evaluation. First, we visually compare how the unmeasured confounding impacts our causal estimand of interest i.e. the dose response curve. Our assessment looks at the baseline where there is no unmeasured confounding and super-imposes this estimated curve with versions where the impacts of unmeasured confounding is varied at the treatment and outcome stages. Second, we augment this visual comparison using KS-testing [30]. We assess the impact to the posterior distributions of the parameters of the outcome model relative to the baseline as the impact of unmeasured confounding varies at the two stages. Finally, we develop a new Bayesian summarization strategy using simulation from the posterior distribution of the dose-response function.

We formalize the KS-testing scheme as follows: Assume that for any of the parameters of the outcome model denoted $\theta_Y = (\alpha_0, \psi_1, \ldots, \psi_7)$, we want to make pairwise comparisons between the posterior distributions of each parameter at the baseline versus at each unmeasured confounding case. Assume that we generally denote $\mathbf{F}^{\cdot}_{\mathcal{L}_{(W,Y)}}$ to be the posterior distribution of any of these parameters, where the subscript denotes what pair of sensitivity parameters being considered at the treatment and outcome stages respectively and the superscript, the parameter being compared. Therefore we test the following hypothesis,

$$
\begin{aligned}
H_0: &\quad \mathbf{F}^{\theta_Y^{\cdot}}_{\mathcal{L}_{(0,0)}} = \mathbf{F}^{\theta_Y^{\cdot}}_{\mathcal{L}_{(W,Y)}} \\
H_A: &\quad \mathbf{F}^{\theta_Y^{\cdot}}_{\mathcal{L}_{(0,0)}} \neq \mathbf{F}^{\theta_Y^{\cdot}}_{\mathcal{L}_{(W,Y)}}
\end{aligned}
$$

In the next section, we present a discussion on the evaluation of the dose-response function directly to quantify the effects of violations of unconfoundedness, for which we develop a new technique.

4.3.2 Summarizing Differences in Dose-Response Functions

Recycling the notation from above, assume that we simulate $\mu(\omega)$, the average dose-response curve discussed in Chapter 3, from its posterior distribution $\mathbf{H}_{\mathcal{L}_{(W,Y)}}$, where the subscript, as before, denotes the pair of sensitivity parameters being considered. We are interested in comparing overlap of the distributions of the dose-response function under various degrees of unmeasured confounding controlled by the sensitivity parameters relative to the baseline case.

Assume that we simulate T draws from the distribution $\mathbf{H}_{\mathcal{L}_{(W,Y)}}$ that we denote $\{\mu(\omega)_{\mathcal{L}_{(W,Y)}}\}_{t=1}^{T}$ and analogous draws from the baseline unconfounded distribution $\mathbf{H}_{\mathcal{L}_{(0,0)}}$ that we denote $\{\mu(\omega)_{\mathcal{L}_{(0,0)}}\}_{t=1}^{T}$. Each draw from the baseline distribution has a natural pairing in the confounded distribution. Therefore, we can define the differences of the draws,

$$
\mu(\omega)^t_{\mathcal{D}} = \mu(\omega)^t_{\mathcal{L}_{(W,Y)}} - \mu(\omega)^t_{\mathcal{L}_{(0,0)}}
$$

and the mean difference as,

$$
\bar{\mu(\omega)}_{\mathcal{D}} = \frac{1}{T}\sum_{t=1}^{T} \mu(\omega)^t_{\mathcal{D}}
$$

and the radius, $\mathcal{R}$ as being the distance between the mean difference and the origin,

$$\mathcal{R} = \text{dist}(\bar{\mu(\omega)}_{\mathcal{D}}, \mathbf{0})$$

The radius $\mathcal{R}$, defined here acts as a threshold for assessing how likely the two dose-response functions are to have come from the same distribution and summarizes their average overlap. We threshold the average difference against the origin since intuitively that is the *ideal* difference if confounded and baseline dose-response functions are on average the same. Using this summary, we can estimate the probability with which the two distributions are different as,

$$\mathcal{P} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(\mathcal{R} < \text{dist}(\bar{\mu(\omega)}_{\mathcal{D}}, \mu(\omega)^{t}_{\mathcal{D}}))$$

In the expression above, we are determining how many simulated differences $\{\mu(\omega)_{\mathcal{D}}\}_{t=1}^{T}$ lie within the ball defined by the distance between the origin and the mean difference of the dose-response functions. If the two distributions are not vastly different from each other, we expect that more of the distances should be outside the ball defined by $\mathcal{R}$. On the other hand, if the two distributions are substantially different from each other, more of the differences should lie within the ball indicating that on average the differences are larger than zero.

In the representation above, we have not specified what type of distance measure to use. In our application, we use the euclidean distance, and therefore, $\mathcal{R}$ is the L_2 norm of the average difference of the dose-response function. However, this approach can be generalized to other distance measures as well, albeit this distance should be well defined for functions. Validation for this technique in the form of visual diagnostics is given in .8.

Our proposed technique offers four advantages over past work on sensitivity analysis in causal inference. First, our approach is tailored to assessments of unmeasured confounding for hierarchical dose-response functions using the generalized propensity score. Second, using posterior simulation under the Bayesian paradigm, we assess the effects of unmeasured confounding at a distributional level in two ways: indirectly via the parameters of the outcome model and directly by using the proposed scheme on the draws of the dose-response function. Therefore our understanding of unmeasured confounding is more holistic than the assessment of point estimates and standard errors alone. Finally, this approach has built in flexibility – by using suitable priors, we assess unmeasured confounding of various types i.e. binary or continuous.

The causal analysis is considered relatively robust to unmeasured confounding if the posterior distributions for the parameters of interest and the estimated dose-response curve are not significantly different from the baseline both visually and via our proposed tests. It is pertinent to mention that sensitivity analyses showing favorable results are not evidence implying that the unconfoundedness assumption is satisfied. However, unfavorable sensitivity analyses can cast considerable doubt on the findings of a causal analysis. It is therefore vital to consider substantive ways in which unmeasured confounding can effect the findings of an analysis and attempt to conduct sensitivity analyses in a fashion that reflects that.

4.4 Simulation Study to Demonstrate Conditional Sampling of the Unmeasured Confounder

The objective of this simulation study, derived from [14], is to help establish the usefulness of conditional simulation in connecting the observed and full data likelihoods. For our method to be a useful representation of the effects of unmeasured confounding, controlling for values of the confounder directly and controlling for its values by imputation via conditional simulation need to yield similar results, particularly when compared against the case where the unmeasured confounder is completely ignored. We will consider the continuous case of unmeasured confounding denoted U_{C} in the following setup.

Data Generation Scheme: We start by simulating the following $p = 5$ covariates as independent standard normal covariates.

$$X_k \sim \text{Normal}(0, 1); \quad k = 1, \ldots, 5$$

and the unmeasured confounder,

$$U_{\text{C}} \sim \text{Normal}(0, 0.5)$$

Next, we simulate the treatments based on the covariates and the unmeasured confounders as follows,

$$\mu_W = 0.25 \sum_{k=1}^{5} X_k + \mathcal{L}_{\text{C}}^{W} U_{\text{C}}$$
$$W|X_k \sim \text{Normal}(\mu_W, 1); \quad k = 1, \ldots, 5$$

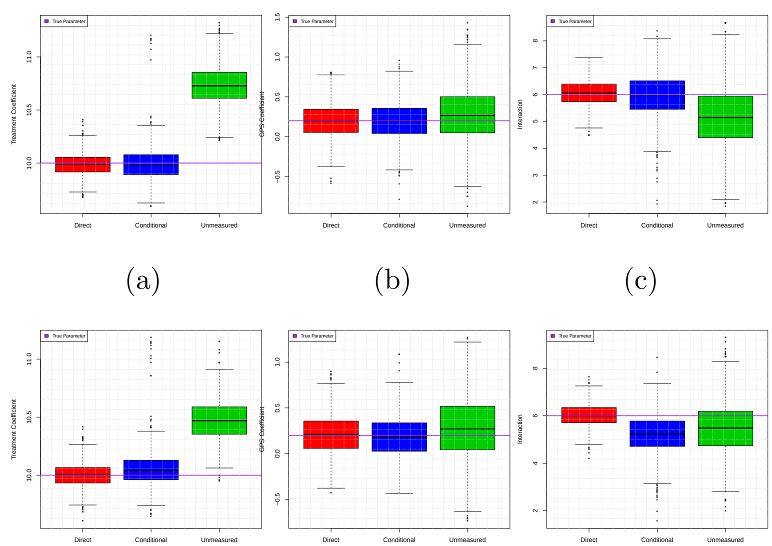
where the parameter $\mathcal{L}_{\text{C}}^{W}$ is the *sensitivity parameter* that control the impact of the unmeasured confounder at the treatment level. Finally we generate the outcomes based on the simulated covariates (via the true generalized propensity score, denoted as R) and treatments,

$$R = \frac{1}{\sqrt{2\pi}} \exp \frac{(W - \mu_W)^2}{-2}$$
$$\mu_Y = 10 \cdot W + 0.2 \cdot R + 6 \cdot W \cdot R + \mathcal{L}_{\text{C}}^{Y} U_{\text{C}}$$
$$Y|W, X_k \sim \text{Normal}(\mu_Y, 1)$$

where the parameter $\mathcal{L}_{\text{C}}^{Y}$ denotes the sensitivity parameter that control the impact of the unmeasured confounders at the outcome level. The use of the generalized propensity score delineates this simulation study from the work presented in [14] since its use is at the core of our application.

Assessment We aim to apply our proposed method of sensitivity analysis to the estimation of dose-response functions. Therefore, we want to assess the variation in the estimation of the outcome model parameters that determine the shape of the dose-response function, under various values of the sensitivity parameters at the two stages of the data generating process.

We generate outcome model parameter estimates from 1000 simulations of size $n = 250$, with the sensitivity parameters set to $\mathcal{L}_{\mathrm{C}}^{Y} = (0, 1, 2)$ and $\mathcal{L}_{\mathrm{C}}^{W} = (-2, -1, 0, 1, 2)$ for the unmeasured confounder. These estimates are compared using side-by-side box and whisker plots to determine how the result vary when (1) directly controlling for the unmeasured confounder, (2) controlling for the unmeasured confounder using our proposed simulation technique and (3) ignoring the unmeasured confounder. An example of these comparative plots is given in Figure 4.1.



(a) (b) (c)

(d) (e) (f)

Figure 4.1: Selection of plots for assessment of introduced method using simulated data with $\mathcal{L}_{\mathrm{C}}^{Y} = 1$ and $\mathcal{L}_{\mathrm{C}}^{W} = -2$.

Based on the plots presented in 4.1, we are can understand a few different properties of unmeasured confounding as modeled by conditional simulation. First, in this very controlled setting, we observe that the behavior of the estimated parameters is very similar between the direct control of the unmeasured confounder, and controlling for it via conditional simulation. Second, when the confounder is ignored in the analysis, the parameter estimates are drastically altered. The conditional values of the unmeasured confounder lead to estimates that are somewhere between the estimates when the confounder is ignored versus direct control. Finally, what we notice is that the effects are most prominent when the parameter that we are interested in gauging the impact on is large. Therefore, we can conclude that the method of conditional simulation is an effective one in bridging the complete and observed data likelihoods. In the next section, we use this as the backbone for assessing the impacts of unmeasured confounding on our motivating dataset.

4.5 Results

4.5.1 Assessing the Impacts of Unmeasured Confounding for the Udayan Ghar Data

The results that we present in this section are based on the data from the Udayan Ghar Program. The assessment is focused on the dose-response curve and the parameters of the outcome stage of the model in equation (3.11), under various values of $\mathcal{L}_W$ and $\mathcal{L}_Y$ relative to the baseline $\mathcal{L}_W = \mathcal{L}_Y = 0$. The objective is to determine how the inference we garner is changed in terms of the estimated dose-response curve and the parameters of the outcome model as the impact of unmeasured confounding is varied. As discussed in section 4.3 the two types of functional forms for the unmeasured confounding that we have considered are both representations of how the dose-response function can be impacted by binary and continuous unmeasured confounding.

The values of the sensitivity parameters are set to be $\mathcal{L}_W$ = (-2, -1, -0.5, 0.5, 1, 2) and $\mathcal{L}_Y$ = (0.25, 0.5, 1, 2). These values were derived from the coefficients of the height and weight in the baseline setting. We evaluate all possible pairings of the sensitivity parameters under the settings defined above for each type of unmeasured confounding that we want to investigate.

First, we visually inspect the impact of unmeasured confounding on the dose-response function in Figure 4.2 if the continuous unmeasured confounder behaves like the developmental indices common in the motivating literature. Relative to the baseline, there are two important findings that we highlight. First, we see that as the sensitivity parameters are varied at the outcome stage, the dose response curve originates at a lower level and terminates at a lower level as well therefore we observe a shift. However, as the sensitivity parameters are varied at the treatment level, we see minor changes in the slope of the curve at various points relative to the baseline. In general we observe a flattening at the ends, since this is the region where the dose-response function is most likely to be susceptible to concerns of confounding due to limited data.

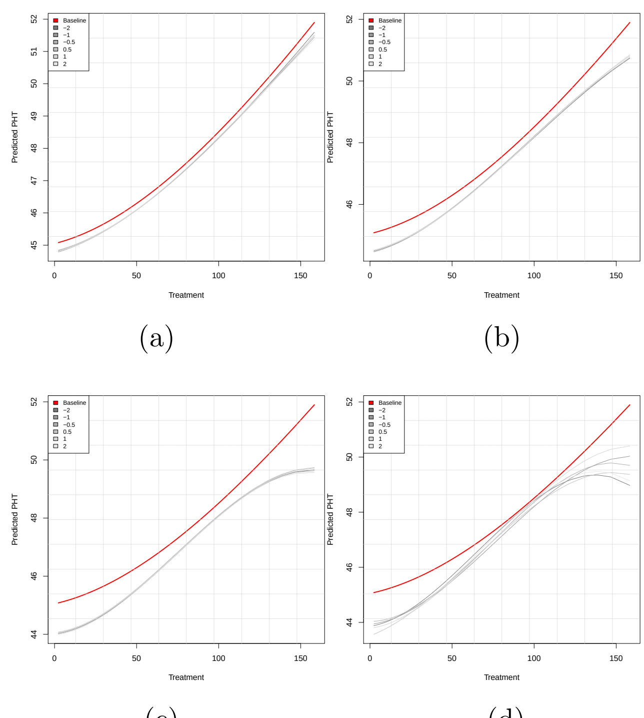


Figure 4.2: Impacts of unmeasured confounding under the polynomial case. Each panel corresponds from left to right and top to bottom values of $\mathcal{L}_Y = (0.25, 0.5, 1, 2)$.

Next, we consider the impact of unmeasured confounding if the confounder is binary.

Our functional form was chosen to represent a variable that interacts with observed covariates (thereby defining subpopulation like behavior) and the treatments – this was akin to the gender specific heterogeneity identified in Chapter 3. We aimed to mimic the likely fashion in which the developmental group data that the researchers were interested in investigating, could impact the analysis. The corresponding dose-response functions are given in Figure 4.3. In general, we find that the dose-response functions are dampened in this instance with the baseline gains in ego-resilience and self-concept being more optimistic than what is likely if developmental unmeasured confounders exist.

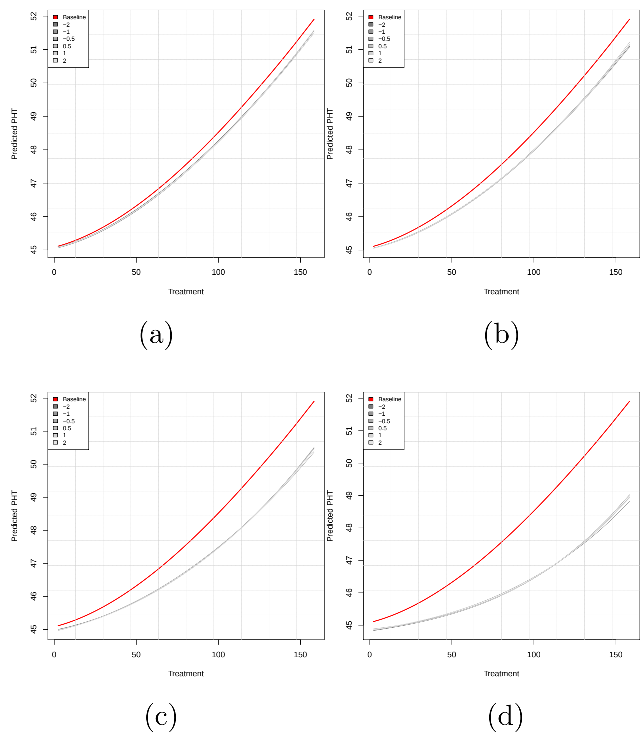


Figure 4.3: Impacts of unmeasured confounding under the binary case. Each panel corresponds from left to right and top to bottom values of $\mathcal{L}_Y = (0.25, 0.5, 1, 2)$.

To supplement the visual analysis, the results of the KS-testing for both cases are summarized in Tables 4.1 and 4.2 to help assess nuances not easily noticeable in the visualizations. What we see is that the posterior distributions of the parameters are significantly different from the baseline, albeit there is a clearer impact of the continuous unmeasured confounder relative to the binary case as expected from the two dose-response curves presented earlier. However, since these parameters are marginalized out, we will use technique we proposed in the last section as a means of considering an alternative summarization scheme that assess variation in the posterior distributions of the dose-response curves directly.

Table 4.1: KS-testing p-values for polynomial unmeasured confounding for the posterior distributions of the parameters of the outcome model relative to the baseline case.

$\mathcal{L}_Y$	$\mathcal{L}_W$	α_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9
	-2	0.15	0.06	0.33	0.66	0.38	0.10	1.00	0.45	0.63	0.29
0.25	-1	0.38	0.29	0.57	0.38	0.08	0.03	0.08	0.15	0.08	0.00
	-0.5	0.63	0.02	0.29	0.72	0.72	0.10	0.21	0.38	0.45	0.06
	0.5	0.45	0.88	0.63	0.48	0.25	0.21	0.05	0.88	0.07	0.01
	1	0.63	0.23	0.29	0.33	0.40	0.17	0.81	0.78	0.18	0.00
	2	0.40	0.38	0.29	0.57	0.43	1.00	0.12	0.04	0.20	0.69
	-2	0.00	0.21	0.00	0.00	0.97	0.03	0.84	0.75	0.06	0.08
0.5	-1	0.00	0.02	0.20	0.15	0.91	0.60	0.29	0.86	0.23	0.00
	-0.5	0.25	0.43	0.07	0.01	0.88	0.10	0.69	0.75	0.23	0.01
	0.5	0.00	0.00	0.51	0.05	0.45	0.66	0.86	0.94	0.03	0.48
	1	0.31	0.05	0.08	0.03	0.94	0.08	0.45	0.40	0.97	0.04
	2	0.45	0.33	0.33	0.03	0.33	0.13	0.78	0.75	0.01	0.00
	-2	0.00	0.54	0.00	0.00	0.11	0.01	0.01	0.08	0.00	0.00
1	-1	0.00	0.02	0.00	0.00	0.63	0.25	0.38	0.63	0.02	0.00
	-0.5	0.09	0.29	0.00	0.00	0.02	0.01	0.21	0.02	0.00	0.00
	0.5	0.01	0.40	0.00	0.00	0.23	0.02	0.21	0.45	0.00	0.00
	1	0.00	0.25	0.00	0.00	0.17	0.05	0.43	0.25	0.01	0.00
	2	0.13	0.78	0.00	0.00	0.54	0.02	0.05	0.05	0.02	0.00
	-2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
2	-1	0.00	0.01	0.00	0.00	0.08	0.00	0.01	0.03	0.00	0.00
	-0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00
	0.5	0.00	0.00	0.00	0.00	0.01	0.00	0.12	0.05	0.00	0.00
	1	0.00	0.00	0.00	0.00	0.02	0.01	0.09	0.04	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Visualizations of these posterior draws using kernel density estimates [80] are given in given in Appendix .7. The goal of presenting these is to grant the reader a visual aid as to what drives the variation in the KS-testing results in the cases considered. A recurring feature to notice is that the impact on the dose-response functions is more substantial in both cases considered, when there is a higher degree of unmeasured confounding at the outcome stage particularly at the levels of the sensitivity parameters that are nearly twice as large as the coefficients estimated for height and weight in Table 3.2

Table 4.2: KS-testing p-values for binary unmeasured confounding for the posterior distributions of the parameters of the outcome model relative to the baseline case.

$\mathcal{L}_Y$	$\mathcal{L}_W$	α_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9
	-2	0.15	0.65	0.44	0.07	0.85	0.35	0.20	0.41	0.19	0.67
0.25	-1	0.32	0.78	0.28	0.19	0.98	0.61	0.89	0.99	0.89	0.55
	-0.5	0.15	0.38	0.95	0.99	0.57	0.07	0.98	0.41	0.14	0.46
	0.5	0.38	0.92	0.08	0.38	0.70	0.55	0.03	0.63	0.65	0.49
	1	0.01	0.70	0.49	0.92	0.15	0.70	0.44	0.72	0.02	0.33
	2	0.01	0.51	0.70	0.57	0.44	0.65	0.23	0.06	0.00	0.11
	-2	0.03	0.98	0.08	0.49	0.93	0.39	0.57	0.72	0.00	0.65
0.5	-1	0.03	0.21	0.53	0.24	0.78	0.36	0.53	0.76	0.18	0.63
	-0.5	0.63	0.99	0.07	0.20	0.98	0.48	0.99	0.70	0.63	0.95
	0.5	0.15	0.57	0.05	0.01	0.44	0.35	0.91	0.28	0.08	0.22
	1	0.28	0.63	0.15	0.06	0.85	0.49	0.81	0.49	0.05	0.89
	2	0.27	0.78	0.02	0.03	0.33	0.38	0.24	0.29	0.33	0.35
	-2	0.27	0.16	0.00	0.04	0.32	0.63	0.88	0.13	0.32	0.42
1	-1	0.06	0.03	0.01	0.03	0.96	0.70	0.23	0.80	0.03	0.72
	-0.5	0.25	0.48	0.00	0.01	0.12	0.41	0.12	0.89	0.81	0.25
	0.5	0.27	0.28	0.00	0.00	0.67	0.83	0.74	0.63	0.83	0.48
	1	0.00	0.05	0.04	0.05	0.91	0.44	0.48	0.13	0.01	0.85
	2	0.05	0.25	0.00	0.04	0.63	0.89	0.78	0.78	0.19	0.20
	-2	0.00	0.12	0.00	0.00	0.17	0.49	0.49	0.41	0.02	0.46
2	-1	0.00	0.14	0.00	0.00	0.68	0.30	0.25	0.44	0.01	0.36
	-0.5	0.35	0.19	0.00	0.00	0.97	0.35	0.65	0.86	0.44	0.23
	0.5	0.00	0.70	0.00	0.00	0.22	0.25	0.98	0.70	0.15	0.51
	1	0.06	0.03	0.00	0.00	0.57	0.78	0.08	0.44	0.30	0.83
	2	0.08	0.14	0.00	0.00	0.38	0.30	0.28	0.12	0.80	0.00

The last stage in the assessment of unmeasured confounding attempts to determine how the distribution of the actual dose-response function changes as we vary the degree of unmeasured confounding. We provide summarizations of these findings in Tables 4.3 and 4.4 for the polynomial and binary cases respectively.

By design if the draws from the distributions are comparable, the estimated probabilities should be small i.e. closer to zero. In the case of the continuous unmeasured confounder, if the $\mathcal{L}_Y$ is small, then the distributions are not very dissimilar – however as this value becomes larger, the unmeasured confounding starts to have a more substantial

Table 4.3: Analysis of polynomial unmeasured confounding using the proposed summary scheme.

	$\mathcal{L}_Y = +0.25$	$\mathcal{L}_Y = +0.50$	$\mathcal{L}_Y = +1.0$	$\mathcal{L}_Y = +2.0$
$\mathcal{L}_W = -2.0$	0.00	0.07	0.30	0.28
$\mathcal{L}_W = -1.0$	0.00	0.06	0.27	0.27
$\mathcal{L}_W = -0.5$	0.00	0.06	0.28	0.22
$\mathcal{L}_W = +0.5$	0.00	0.05	0.24	0.37
$\mathcal{L}_W = +1.0$	0.00	0.06	0.30	0.24
$\mathcal{L}_W = +2.0$	0.00	0.06	0.28	0.31

effect on the inference even at lower values of $\mathcal{L}_W$. These results align with the results of the visual analysis: when the sensitivity parameters are large, the overlap between the uncertainty intervals of the confounded curves versus the baseline start to show less and less overlap, which indicates that the results are becoming substantially different. However these probabilities are still relatively low, i.e. even at the highest levels of the sensitivity parameters, the underlying distributions of the two dose-response functions are not likely (less than 50%) to be different.

Table 4.4: Analysis of binary unmeasured confounding using the proposed summary scheme.

	$\mathcal{L}_Y = +0.25$	$\mathcal{L}_Y = +0.50$	$\mathcal{L}_Y = +1.0$	$\mathcal{L}_Y = +2.0$
$\mathcal{L}_W = -2.0$	0.00	0.03	0.20	0.74
$\mathcal{L}_W = -1.0$	0.01	0.03	0.21	0.73
$\mathcal{L}_W = -0.5$	0.00	0.03	0.21	0.73
$\mathcal{L}_W = +0.5$	0.01	0.03	0.22	0.73
$\mathcal{L}_W = -1.0$	0.01	0.03	0.22	0.74
$\mathcal{L}_W = -2.0$	0.01	0.03	0.23	0.73

In the case of the binary unmeasured confounder, our analysis indicates that our causal results are less robust particularly at higher levels of the outcome sensitivity parameters. However, while the confounded curves appear shifted from the baseline their overall behavior does not change – there is still an expected increase in the predicted Piers-Harris scores. We see these results through our proposed testing strategy as well. Supplementing the visual analysis, at higher levels of $\mathcal{L}_Y$, our summarization indicates that distributions for the curves are likely to be different. Our findings are also supported by the last panel in Table 4.2. The next section provides some closing remarks on future work we are interested in exploring.

4.6 Discussion and Future Work

The analysis that we have conducted in Chapter 3 relies on the identifying assumption that the treatment assignment is unconfounded. Motivated by the complex hierarhical structure of the data and by the causal estimand of interest, we developed and applied a new sensitivity analysis technique and a new summarization strategy based on posterior simulation to assess the robustness of our findings against violations of this key identifying assumption.

Using visual evidence and the two stages of testing we have a number of insights into the violations of the unconfoundedness assumption in our application. We find that if the unmeasured confounder is continuous, and behaves in a fashion similar to other physical and psychological growth metrics, the dose response curve is susceptible to changes at the extremities where data is limited. If on the other hand, the unmeasured confounder is binary, its effect appears to be more to dampen the effectiveness of the residential care program. We find that in both instances, the overall trajectory of the dose-response function remains largely positive – therefore our conclusion remains that the Udayan Ghar Program does provide positive psychological returns to its residents. Furthermore, the extent of the variation as a result of induced unmeasured confounding leaves the results in the same categorization of the Piers-Harris scale which means that practically, our results are unchanged.

There are however, several areas where this technique will require further development. First, we need to develop more sophisticated flexible distributions that can be adapted for observed data to substantively represent unmeasured confounding. We have only considered two possible cases which are largely motivated by the application. Second, there are other subpopulations of interest to us that we currently do not have sufficient data for, such as children that enter the program at various stages of psychological development proxied by age – a confounder that we considered as motivation for our sensitivity analysis. Third, under the summarization framework we propose, we considered the analysis under one possible distance measure that we were also able to validate. We are interested in investigating more flexible methods to select the distance measure used for various applications. These are concepts we aim to explore in our future work.

Chapter 5

Conclusion

Causal inference using statistical methods is arguably a Bayesian problem, particularly when considered under the potential outcomes framework posited by the Rubin-Neyman Causal Model. Naturally, statistical inference is most useful when our goal is understanding the effects of various policy or treatment interventions i.e. causal effect estimation where the aforementioned model grants both an intuitive definition of causal effects as well as a high degree of model flexibility. Furthermore, given the importance of uncertainty quantification in said problems, working under the Bayesian paradigm is doubly suitable.

While the potential outcomes framework does provide an intuitive formulation of treatment effects, it can mask underlying complications that warrant more sophisticated statistical thinking, particularly in the realm of observational causal inference problems. This dissertation has attempted to tackle three distinct issues that underlie questions in observational causal inference and extend the potential outcomes framework to help address these.

In a slew of causal inference problems, the focus has been the estimation of average treatment effects that ignore covariate level driven idiosyncrasies. However, based on recent developments in the literature, we conclude that heterogeneous treatment effects are more the norm than the exception. In chapter 2, we start by developing a flexible technique for heterogeneous treatment effect estimation that uses Bayesian non-parametric methods. The method we propose uses a two component mixture of Gaussian processes to capture covariate driven treatment effect heterogeneity through the conditional average treatment effect.

Addressing non-binary treatment effects, in Chapter 3, we proposed a methodology for estimating the effects of a continuous treatment in a hierarchical data setting where the estimand of interest is a dose-response function. This method was motivated by the analysis of data that aimed to assess the effects of a new foster care system – units of observation are measured multiple times, and these are nested within residential cohorts. Our work is distinct from past work since it attempts to take the hierarchical nature of a dose-response curve into account; earlier work has focused on either dose-response function estimation in non-hierarchical settings, or estimating binary treatment effects from multistage clustered data. Our focus is on combining ideas from both.

The final methodological development that we posit in Chapter 4 is the development of a new Bayesian sensitivity analysis technique for assessing how violations of unconfoundedness can impact causal findings under continuous treatment regimes. The technique that we have proposed can be integrated into any MCMC sampling scheme where the unmeasured confounder can be sampled from its full conditional distribution subject to sensitivity parameter settings. Using this framework, we can explore the impacts of unmeasured confounding of various types e.g. binary or continuous. We were able to demonstrate that this technique, subject to model specification considerations, is able to generate draws of the unmeasured confounder that behave much in the same way as directly controlling for the it. We supplemented this with a new means of posterior comparison to summarize the overlap between the confounded and the unconfounded dose-response function distributions providing a novel assessment scheme for unconfoundedness.

However, it is important, in the spirit of ongoing research to recognize the limitations of the work that we have presented in this dissertation. First, in modeling treatment effect heterogeneity, we have showed how we can use the flexibility of Bayesian non-parametric regression to estimate and identify causal estimands that vary with covariate levels. In doing so, we found that prior calibration is of paramount importance and while we developed some heuristic strategies to help reduce prior sensitivity, it does remain a non-negligible concern. As a result, we feel that a concrete mechanism to help calibrate the priors for our proposed model is vital.

A second concern develops with respect to the sensitivity analysis method for unconfoundedness that we have proposed. While our method for simulating unmeasured confounders from their conditional distributions performs well in the specified setting, it does place emphasis on a correctly specified model. As a next step, we are interested in exploring unconfoundedness under flexible models. Furthermore, the comparisons that we have presented in this dissertation are not formal Bayesian hypothesis tests. For our findings to be stated as such, we need to reformulate these as such using ideas of Bayes factors. These areas present the concrete directions for future work that we are interested in exploring.

Appendices

.1 Proof of Equivalence

We now show that the transformation presented in section 2.2 in expectation recovers the CATE i.e.

$$\mathbb{E}_Y[Y_i^* \mid X_i = x] = \tau(x).$$

Proof. First observe that $Y_i = Y_i(W_i) = W_i Y_i(1) + (1 - W_i)Y_i(0)$.

By the definition of the TRV

$$\begin{aligned} A = \mathbb{E}_Y[Y_i^* \mid X_i = x, \mathcal{D}] &= \mathbb{E}_Y\left[\frac{W_i - e_i}{e_i(1-e_i)} Y_i \mid X_i = x, \mathcal{D}\right], \\ &= \frac{1}{e_i(1-e_i)}\left(\mathbb{E}_Y[Y_i W_i \mid X_i = x, \mathcal{D}] - e_i \mathbb{E}_Y[Y_i \mid X_i = x, \mathcal{D}]\right). \end{aligned}$$

Due to the ignorability of the treatment assignment the following holds

$$\begin{aligned} A &= \frac{1}{e_i(1-e_i)}(e_i \mathbb{E}_Y[Y_i \mid W_i = 1, X_i = x, \mathcal{D}] - e_i \mathbb{E}_Y[Y_i \mid X_i = x, \mathcal{D}]) \\ &= \frac{1}{1-e_i}\mathbb{E}_Y[Y_i \mid X_i = x, W_i = 1, \mathcal{D}] - \frac{1}{1-e_i}\mathbb{E}_Y[Y_i \mid X_i = x, \mathcal{D}]. \end{aligned}$$

By iterating expectations the following holds:

$$\begin{aligned} A &= \frac{1}{1-e_i}\mathbb{E}_Y[Y_i \mid W_i = 1, X_i = x, \mathcal{D}] - \frac{1}{1-e_i}\mathbb{E}_W[\mathbb{E}_Y[Y_i \mid W_i = 1, X_i = x, \mathcal{D}]], \\ &= \frac{1}{1-e_i}\mathbb{E}_Y[Y_i \mid W_i = 1, X_i = x, \mathcal{D}] - \frac{e_i}{1-e_i}\mathbb{E}_Y[Y_i \mid W_i = 1, X_i = x, \mathcal{D}] - \\ &\quad \mathbb{E}_Y[Y_i \mid W = 0, X_i = x, \mathcal{D}]. \end{aligned}$$

Collecting the first two terms provides the desired result

$$A = \mathbb{E}_Y[Y_i \mid W_i = 1, X_i = x, \mathcal{D}] - \mathbb{E}_Y[Y_i \mid W_i = 0, X_i = x, \mathcal{D}].$$

□

.2 Derivation of Model

The derivation of the model presented in the dissertation begins with the transformation of interest given as follows, with Y_i denoting the observed response, W_i the assigned treatment and $e_i = P(W_i = 1|X_i = x)$

$$Y_i^* = \frac{W_i - e_i}{e_i(1-e_i)} Y_i$$

In addition, we define the two regression functions for the outcome, one under the treatment and one under the control,

$$(Y_i|W_i = 0) = f_0(X_i) + \epsilon_i(0)$$
$$(Y_i|W_i = 1) = f_1(X_i) + \epsilon_i(1)$$

Using the transformation, and substituting the regression functions under the two cases i.e. when $W_i = 1$ and when $W_i = 0$ and assuming further that $\epsilon(1), \epsilon(0) \overset{IID}{\sim} \mathrm{N}(0, \sigma^2)$, we can define with probability e_i,

$$\begin{aligned}(Y_i^*|W_i = 1) &= \frac{f_1(X_i) - e_i f_1(X_i) + e_i f_0(X_i)}{e_i} + f_1(X_i) - f_0(X_i) + \frac{\epsilon_i(1)}{e_i} \\ &= f_1(X_i) - f_0(X_i) + (1-e_i)\left(\frac{f_1(X_i)}{e_i} + \frac{f_0(X_i)}{1-e_i}\right) + \frac{\epsilon_i(1)}{e_i} \\ &= g(X_i) + (1-e_i)h(X_i) + \frac{\epsilon_i(1)}{e_i}\end{aligned}$$

and similarly, with probability $1 - e_i$ that,

$$\begin{aligned}(Y_i^*|W_i = 0) &= \frac{-(1-e_i)f_1(X_i) + (1-e_i)f_0(X_i) - f_0(X_i)}{e_i} + f_1(X_i) - f_0(X_i) - \frac{\epsilon_i(0)}{1-e_i} \\ &= f_1(X_i) - f_0(X_i) + (-e_i)\left(\frac{f_1(X_i)}{e_i} + \frac{f_0(X_i)}{1-e_i}\right) - \frac{\epsilon_i(0)}{1-e_i} \\ &= g(X_i) + (-e_i)h(X_i) + \frac{\epsilon_i(0)}{e_i - 1}\end{aligned}$$

This yields the mixture model model that we have presented in Chapter 2,

$$Y_i^* = g(X_i) + \varepsilon_i$$
$$\varepsilon_i \sim (e_i)\mathrm{Normal}((1-e_i)h(X_i), \frac{\sigma^2}{e_i^2}) + (1-e_i)\mathrm{Normal}(-e_i h(X_i), \frac{\sigma^2}{(1-e_i)^2})$$

Table .1: Comparison of conditional average treatment effects by decile of standardized income, along with 95% uncertainty intervals using alternative models.

$Avg.Income_{Decile}$	tot	fit	ct	$BART$	RF	lwr_{tot}	upr_{tot}	lwr_{fit}	upr_{fit}	lwr_{ct}	upr_{ct}	lwr_{BART}	upr_{BART}	lwr_{RF}	upr_{RF}
-1.137	0.470	0.234	0.637	0.118	0.420	0.089	0.678	0.087	0.741	0.202	1.103	-1.166	1.394	0.082	0.772
-0.831	0.470	0.417	0.500	0.117	0.383	0.087	0.672	0.096	0.730	0.191	1.117	-1.211	1.397	0.083	0.770
-0.638	0.470	0.538	0.515	0.105	0.461	0.067	0.679	0.083	0.733	0.183	1.094	-1.204	1.398	0.082	0.772
-0.472	0.470	0.258	0.430	0.089	0.376	0.076	0.676	0.094	0.725	0.193	1.105	-1.224	1.393	0.085	0.744
-0.310	0.470	0.093	0.307	0.095	0.276	0.077	0.670	0.082	0.740	0.198	1.112	-1.191	1.398	0.075	0.766
-0.114	0.470	0.598	0.681	0.116	0.391	0.074	0.676	0.097	0.732	0.204	1.103	-1.107	1.397	0.083	0.772
0.103	0.470	0.471	0.552	0.096	0.407	0.081	0.660	0.086	0.748	0.212	1.112	-1.136	1.370	0.087	0.760
0.397	0.470	0.414	-0.103	0.122	0.269	0.086	0.682	0.092	0.744	0.192	1.104	-1.120	1.385	0.082	0.760
0.848	0.470	0.363	0.384	0.134	0.442	0.082	0.667	0.094	0.744	0.178	1.108	-1.119	1.400	0.087	0.776
2.143	0.470	0.396	0.835	0.156	0.706	0.075	0.673	0.107	0.735	0.190	1.108	-1.068	1.406	0.081	0.760

.3 Comparison of SHIW Data

This section presents comparative analysis using various methods for the CATE estimation for the SHIW data using the Gaussian process mixture in section 2.4.3. Point estimates of the CATE along with 95% uncertainty intervals for each decile of income, along with the average value of income in that decile are presented in Table .3.

.4 Exploratory Data Analysis

The missingness in each of the covariates is summarized in Table .2. The height, weight and total number of traumatic events contain the most missingness, although the extent of this is a third in the number of traumatic events relative to the other two variables.

Table .2: Percentage of Missingness in the Data from the Total of 180 Observations.

Variable Name	**Percentage of Missingness**
Piers Harris (Standardized)	<5%
Homes	<1%
Age at Entry (Yrs.)	<1%
Caregiver Ratio	<1%
Height (Centimeters)	63%
Weight (Kilograms)	63%
Total Traumatic Events	22%

The summary statistics of the covariates before and after imputation are given in Table .3 with the mean of 50 iterations of the algorithm used. We can surmise that the imputation leaves the data relatively unchanged.

We examine the post-imputation covariates considered in this analysis for three cases: (1) Jointly for all occupants, (2) for male occupants only and (3) for female occupants only.

Table .3: Comparison of covariate summary measurements before and after imputation using MICE.

Variable Name	Mean (Complete)	SD (Complete)	Mean (Imputed)	SD (Imputed)
Piers Harris (Standardized)	47.09	8.86	47.17	8.87
Treatment (Months)	64.40	38.24	64.40	38.24
Age at Entry (Yrs.)	7.76	2.82	7.78	2.82
Gender	0.62	0.49	0.62	0.49
Caregiver Ratio	5.20	1.48	5.20	1.48
Height (Centimeters)	150.00	14.38	151.23	12.65
Weight (Kilograms)	42.66	13.11	42.68	12.49
Total Traumatic Events	2.91	1.25	2.82	1.25

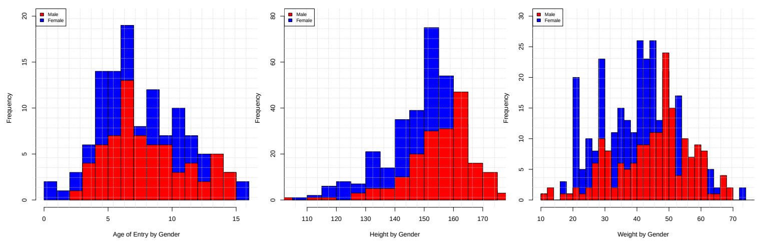


Figure .1: Exploratory analysis of the age and biometric data measured for each child in the Udayan Ghar Program. Panel (a) Age of entry (in days). Panel (b) Height at entry (in centimeters). Panel (c) Weight at entry (in kilograms).

With respect to the age of entry (Figure .1(a)), we can see that the data ranges from 0 years to 16 years. When we consider the variation in age of entry between male and female children, we can see that there is considerable overlap with the range of values for male children and female children – 2 years to 15 years for males, versus $<$ 1 year to 16 years for females. Note that the male children have a higher mean (7.916 years versus 7.689 years) but a marginally lower standard deviation than the female children (2.738 years versus 2.872 years).

For the height measured (Figure .1(b)), we see the same type of summaries by gender. The female children in the sample tend to be shorter (ranging between 105 cm and 173 cm) relative to their male counterparts (ranging between 97 cm to 180 cm), with a mean of 146 cm for the females and 156 cm for the males. Finally, females have a lower standard deviation than the males (13 cm versus 14 cm). It is worth noting that the distribution is somewhat left skewed – more children in the sample tend to be larger in height than

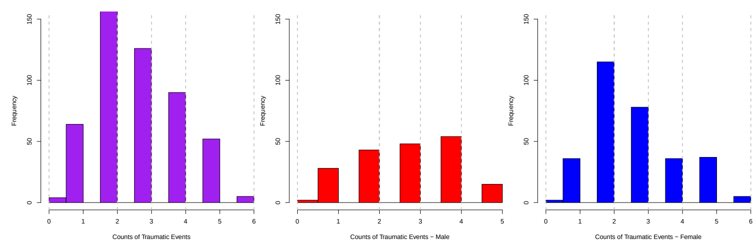


Figure .2: Exploratory analysis of traumatic events. Panel (a) The distribution of number of traumatic events for the entire sample. Panel (b) Counts of traumatic events for males Panel (c) Counts of traumatic events for females.

shorter. The weight of the children (Figure .1(c)) is where the male and female children tend to be the most similar. On average the males in the sample are heavier than the women (mean of 47 kgs versus 39 kgs), albeit not by a large amount. The range of values that we observe have overlap (11 kgs to 73 kgs for the females versus 11 kgs to 88 kgs for the males) and the standard deviations of the weight are very similar (14 kgs for for males versus 12 for females). Finally, in Figure .2, we summarize the total number of traumatic events experienced by the individual students. The female and male children on average experience approximately 3 traumatic events in their lifetimes, with a standard deviation that is also very similar (1.25 events). However, the distribution of these events for the two genders is somewhat different, with the men in the sample more frequently experiencing a higher number of traumatic events relative to the women. The ranges for men and women are also similar (0 to 5 events for men and 0 to 6 events for women).

The data also includes a single environmental variable – the ratio of the residents in a home to the total number of caregivers present (Figure .3. This variable is meant to summarize the quality of the residence, with a small ratio indicating a higher quality of living. We see that on average the ratio is slightly higher for the female homes (5.31) relative to the male homes (5.01) with similar variation; the standard deviation for the female homes is 1.6 relative to the males homes at 1.2. This higher variation is also reflected in observed ranges (2 to 10 for females) and (3 to 7 for males). Only the difference in heights is statistically significantly different between the men and women in the sample; the women

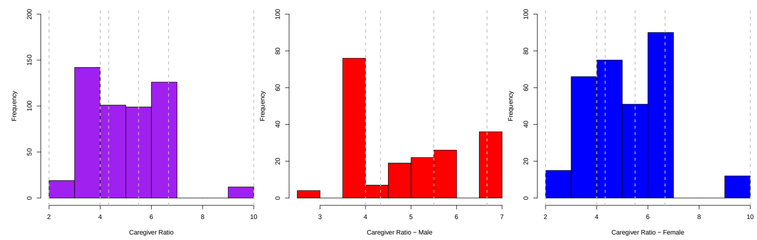


Figure .3: Exploratory analysis of caregiver ratios. Panel (a) The distribution of the ratio of caregivers to house residents for the entire sample. Panel (b) Ratio for males Panel (c) Ratio for females.

in the sample are significantly shorter.

.5 Implied Joint Distributions of the Hierarchical Model Specification

We assumed that W_{kji} denotes the i^{th} measurement for the j^{th} observation in the k^{th} home. Here, we will jointly denote this as the vector $\mathbf{W}$. Therefore, the joint distribution for $\mathbf{W}$ conditional on the model given in section 3.3.2 in (3.7), where β_0 denotes the intercept, $\mathbf{X}$ denotes the design matrix, and $\boldsymbol{\gamma}$ denotes the regression coefficients is given as,

$$
\begin{aligned}
\mathbf{W}|\boldsymbol{\beta}, \mathbf{X}, \sigma^2_\epsilon, \sigma^2_\nu, \sigma^2_\kappa &\sim \text{Normal}_N(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W) \\
\boldsymbol{\mu}_W &= \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\gamma} \qquad (.1)\\
\boldsymbol{\Sigma}_W &= \begin{cases} \sigma^2_\epsilon + \sigma^2_\nu + \sigma^2_\kappa & \text{Measurement Level} \\ \sigma^2_\kappa + \sigma^2_\nu & \text{Observation Level} \\ \sigma^2_\kappa & \text{Home Level} \end{cases}
\end{aligned}
$$

Therefore, the dependencies arising due to the two-stage clustering of the data, are addressed via the block-diagonal matrix $\boldsymbol{\Sigma}_W$. Analogously, at the outcome level, we jointly represent the vector of responses as $\mathbf{Y}$ and $\mathbf{R}$ the generalized propensity score, with the same indices over home, observation and measurement level. The joint distribution under the model in (3.11) with α_0 denoting intercept and $\boldsymbol{\psi} = \{\psi\}_{i=1}^{7}$ denoting the regression coefficients is given as,

$$\mathbf{Y}|\alpha_0, \boldsymbol{\psi}, \mathbf{W}, \mathbf{X}, \mathbf{R}, \sigma^2_\varepsilon, \sigma^2_\xi, \sigma^2_\delta \sim \text{Normal}_N(\boldsymbol{\mu}_Y, \Sigma_Y)$$

$$\mu_Y = \alpha_0\mathbf{1} + \mathbf{W}\psi_1 + \mathbf{W}^2\psi_2 + \mathbf{W}^3\psi_3 + \mathbf{R}\alpha_4 + \mathbf{R}^2\psi_5 + \mathbf{R}^3\psi_6 + \mathbf{W}\cdot\mathbf{R}\psi_7 \tag{.2}$$

$$\boldsymbol{\Sigma}_W = \begin{cases} \sigma^2_\varepsilon + \sigma^2_\xi + \sigma^2_\delta & \text{Measurement Level} \\ \sigma^2_\xi + \sigma^2_\delta & \text{Observation Level} \\ \sigma^2_\delta & \text{Home Level} \end{cases}$$

These two joint distributions as given in (.1) and (.2) are implied by the hierarchical structures of the two models based on how the random effects have been specified. Appendix .6 uses these to develop a full-conditional distribution for the values of an unmeasured confounder $\mathbf{U}$ that we are interested in studying the impact of.

.6 Sampling the Unmeasured Confounder U

The sensitivity analysis framework that we presented in section 4.3, relies on the simulation of the unmeasured confounder from its full conditional posterior distribution. Based on the model framework that we have presented for the analysis of our data, we can specify the following hierarchy,

$$\begin{gathered} \mathbf{Y}|\theta_Y, \mathbf{W}, \mathbf{X}, \mathbf{U}, \mathcal{L}_Y \sim \text{Normal}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y) \\ \mathbf{W}|\theta_W, \mathbf{X}, \mathbf{U}, \mathcal{L}_W \sim \text{Normal}(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W) \end{gathered} \tag{.3}$$

with two possible prior specifications, based on the type of unmeasured confounding that we are interested in studying that are given as follows,

$$\mathbf{U} \sim \text{Normal}(0, \sigma_U^2 \mathbb{I}) \tag{.4}$$

$$U_{kji} \sim \text{Bernoulli}(\pi_{U_{kji}}) \tag{.5}$$
$$\pi_{U_{kji}} = \text{logit}^{-1}(\mathbf{h}(X_{kji}))$$

where $\mathbf{U}$ denotes the vector of unmeasured confounders, and $\mathbf{Y}$, $\mathbf{W}$ and $\mathbf{X}$, are as defined earlier. We use vector notation here again for ease and for consistency with appendix .5. The parameter vectors $\boldsymbol{\mu}_Y$, and $\boldsymbol{\mu}_W$ incorporate $\mathbf{U}$, under (4.2) and (4.3) and are defined as,

$$\boldsymbol{\mu}_Y = \alpha_0 \mathbf{1} + \mathbf{W}\psi_1 + \mathbf{W}^2\psi_2 + \mathbf{W}^3\psi_3 + \mathbf{R}\psi_4 + \mathbf{R}^2\psi_5 + \mathbf{R}^3\psi_6 + \mathbf{W} \cdot \mathbf{R}\psi_7 + \mathbf{g}_Y(\mathbf{U})\mathcal{L}_Y$$

$$\boldsymbol{\mu}_W = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{g}_W(\mathbf{U})\mathcal{L}_W$$

and the two covariance matrices $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_W$ along with parameter vector $\boldsymbol{\psi}$ are defined as in appendix .5.

The unmeasured confounder $\mathbf{U}$, appears in $\boldsymbol{\mu}_Y$ directly via $\mathcal{L}_Y$ and indirectly through the generalized propensity score via $\mathcal{L}_W$ as follows for each entry in the vector $\mathbf{R}$,

$$R_{kji} = \exp\left\{ \frac{(e_{kji}^{(W)} - \mathcal{L}_W \mathbf{g}_W(U_{kji}))^2}{-2} \right\}$$

where $e_{kji}^{(W)} = W_{kji} - \beta_0 - X_{kji}\boldsymbol{\gamma}$ is the partial residual at the treatment stage. The vector of partial residuals is denoted as $\boldsymbol{e}^{(W)}$. At the outcome stage, we define two matrices to help concisely rephrase the model,

$$\bar{\mathbf{R}} = [\mathbf{R}, \mathbf{R}^2, \mathbf{R}^3]$$

$$\bar{\mathbf{W}} = [\mathbf{W}, \mathbf{W}^2, \mathbf{W}^3]$$

The partial residual at the outcome stage is therefore defined as $e_{kji}^{(Y)} = Y_{kji} - \alpha_0 - [W_{kji}, W_{kji}^2, W_{kji}^3]^T \boldsymbol{\psi}_{1:3}$, and its vector form $\boldsymbol{e}^{(Y)}$. The two collections of partial residuals are meant to identify the parts of the treatment and outcome models that are not directly tied to the unmeasured confounder $\mathbf{U}$.

In order to derive the full conditional posterior distribution we utilize the notation defined above to first express the joint distribution, first assuming that unmeasured confounding is of the form in (.4),

$$
\begin{aligned}
f(\mathbf{Y}, \mathbf{W}, \mathbf{U}|\theta_Y, \theta_W, \mathbf{X}) \propto \exp\left\{\frac{(\boldsymbol{e}^{(Y)} - \bar{\mathbf{R}}\boldsymbol{\psi}_{4:6} - \psi_7 \mathbf{W}\cdot\mathbf{R} - \mathcal{L}_Y \mathbf{g}_Y(\mathbf{U}))^T \boldsymbol{\Sigma}_Y^{-1} (\boldsymbol{e}^{(Y)} - \bar{\mathbf{R}}\boldsymbol{\psi}_{4:6} - \psi_7 \mathbf{W}\cdot\mathbf{R} - \mathcal{L}_Y \mathbf{g}_Y(\mathbf{U}))}{-2}\right\}\times \\
\exp\left\{\frac{(\boldsymbol{e}^{(W)} - \mathcal{L}_W \mathbf{g}_W(\mathbf{U}))^T \boldsymbol{\Sigma}_W^{-1} (\boldsymbol{e}^{(W)} - \mathcal{L}_W \mathbf{g}_W(\mathbf{U}))}{-2}\right\}\times \\
\exp\left\{\frac{\mathbf{U}^T\mathbf{U}}{-2\sigma_U^2}\right\}
\end{aligned}
\tag{.6}
$$

We can use .8 to derive the density for the full conditional distribution (up to a normalizing constant) for $\mathbf{U}$, that can be used to sample its values, conditional on the model parameters θ_Y and θ_W and the two sensitivity parameters. This yields,

$$
\begin{aligned}
f(\mathbf{U}|\theta_Y, \theta_W, \mathbf{X}, \mathbf{Y}, \mathbf{W}) \propto \exp\Bigg\{ & \frac{\mathbf{g}_Y(\mathbf{U})^T(\mathcal{L}_Y^2 \boldsymbol{\Sigma}_Y^{-1})\mathbf{g}_Y(\mathbf{U}) + \mathbf{g}_W(\mathbf{U})^T(\mathcal{L}_W^2 \boldsymbol{\Sigma}_W^{-1})\mathbf{g}_W(\mathbf{U}) + \mathbf{U}^T(\frac{1}{\sigma_U^2}\mathbb{I})\mathbf{U}}{} \\
& \frac{-2(\mathbf{g}_Y(\mathbf{U})^T(\mathcal{L}_Y \boldsymbol{\Sigma}_Y^{-1}\boldsymbol{e}^{(Y)} - \mathcal{L}_Y \boldsymbol{\Sigma}_Y^{-1}\bar{\mathbf{R}}\boldsymbol{\psi}_{4:6} - \mathcal{L}_Y \boldsymbol{\Sigma}_Y^{-1}\psi_7 \mathbf{W}\cdot\mathbf{R}) + \mathbf{g}_W(\mathbf{U})^T \mathcal{L}_W \boldsymbol{\Sigma}_W^{-1}\boldsymbol{e}^{(W)}) +}{-2} \\
& \frac{(\bar{\mathbf{R}}\boldsymbol{\psi}_{4:6})^T \boldsymbol{\Sigma}_Y^{-1}(\bar{\mathbf{R}}\boldsymbol{\psi}_{4:6}) - 2(\bar{\mathbf{R}}\boldsymbol{\psi}_{4:6})^T \Sigma_Y^{-1}(\boldsymbol{e}^{(Y)} - \psi_7 \mathbf{W}\cdot\mathbf{R}) + \psi_7^2(\mathbf{W}\cdot\mathbf{R})^T \Sigma_Y^{-1}(\mathbf{W}\cdot\mathbf{R}) - 2\psi_7(\mathbf{W}\cdot\mathbf{R})^T \Sigma_Y^{-1}\boldsymbol{e}^{(Y)}}{}\Bigg\}
\end{aligned}
\tag{.7}
$$

Similarly, under the unmeasured confounding in (.5), the joint distribution is given as,

$$
\begin{aligned}
f(\mathbf{Y}, \mathbf{W}, \mathbf{U}|\theta_Y, \theta_W, \mathbf{X}) \propto \exp\left\{\frac{(\boldsymbol{e}^{(Y)} - \bar{\mathbf{R}}\boldsymbol{\psi}_{4:6} - \psi_7 \mathbf{W}\cdot\mathbf{R} - \mathcal{L}_Y \mathbf{g}_Y(\mathbf{U}))^T \boldsymbol{\Sigma}_Y^{-1} (\boldsymbol{e}^{(Y)} - \bar{\mathbf{R}}\boldsymbol{\psi}_{4:6} - \psi_7 \mathbf{W}\cdot\mathbf{R} - \mathcal{L}_Y \mathbf{g}_Y(\mathbf{U}))}{-2}\right\}\times \\
\exp\left\{\frac{(\boldsymbol{e}^{(W)} - \mathcal{L}_W \mathbf{g}_W(\mathbf{U}))^T \boldsymbol{\Sigma}_W^{-1} (\boldsymbol{e}^{(W)} - \mathcal{L}_W \mathbf{g}_W(\mathbf{U}))}{-2}\right\}\times
\end{aligned}
\tag{.8}
$$

$$\prod_{k=1}^{K}\prod_{j=1}^{N_k}\prod_{i=1}^{N_{kj}} \mathrm{logit}^{-1}(\mathbf{h}(X_{kji}))^{U_{kji}} \cdot (1-\mathrm{logit}^{-1}(\mathbf{h}(X_{kji})))^{(1-U_{kji})}$$

which implies a similar density of the full conditional distribution (again, up to a normalizing constant) as (.7),

$$f(\mathbf{U}|\theta_Y, \theta_W, \mathbf{X}, \mathbf{Y}, \mathbf{W}) \propto \exp\Bigg\{ \frac{\mathbf{g}_Y(\mathbf{U})^T(\mathcal{L}_Y^2\mathbf{\Sigma}_Y^{-1})\mathbf{g}_Y(\mathbf{U}) + \mathbf{g}_W(\mathbf{U})^T(\mathcal{L}_W^2\mathbf{\Sigma}_W^{-1})\mathbf{g}_W(\mathbf{U})}{}$$

$$\frac{-2(\mathbf{g}_Y(\mathbf{U})^T(\mathcal{L}_Y\mathbf{\Sigma}_Y^{-1}\boldsymbol{e}^{(Y)} - \mathcal{L}_Y\mathbf{\Sigma}_Y^{-1}\bar{\mathbf{R}}\boldsymbol{\psi}_{4:6} - \mathcal{L}_Y\mathbf{\Sigma}_Y^{-1}\psi_7\mathbf{W}\cdot\mathbf{R}) + \mathbf{g}_W(\mathbf{U})^T\mathcal{L}_W\mathbf{\Sigma}_W^{-1}\boldsymbol{e}^{(W)}) +}{-2} \tag{.9}$$

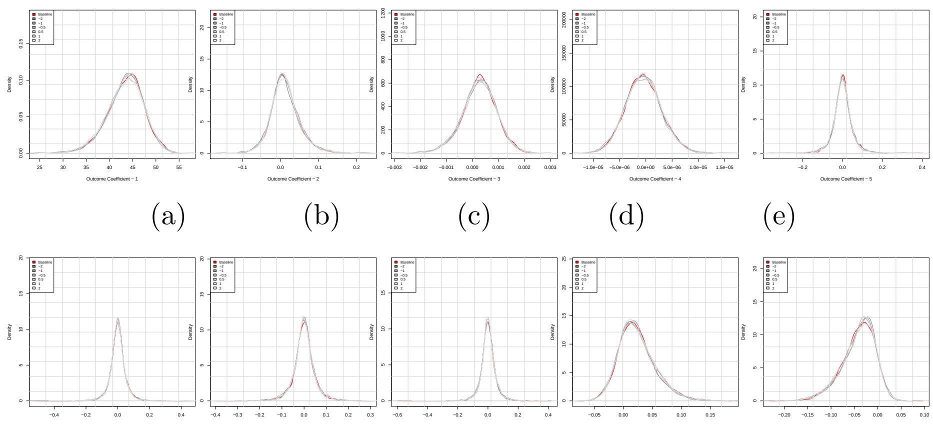
$$\frac{\psi_7^2(\mathbf{W}\cdot\mathbf{R})^T\Sigma_Y^{-1}(\mathbf{W}\cdot\mathbf{R}) - 2\psi_7(\mathbf{W}\cdot\mathbf{R})^T\Sigma_Y^{-1}\boldsymbol{e}^{(Y)}}{}\Bigg\}\times$$

$$\prod_{k=1}^{K}\prod_{j=1}^{N_k}\prod_{i=1}^{N_{kj}} \mathrm{logit}^{-1}(\mathbf{h}(X_{kji}))^{U_{kji}} \cdot (1-\mathrm{logit}^{-1}(\mathbf{h}(X_{kji})))^{(1-U_{kji})}$$

The unnormalized density of the full conditional distributions, given in (.7) and (.9), can be incorporated as an additional sampling step within the overall MCMC algorithm. This can help the researcher assess the impact of unmeasured confounding on the results of an analysis by bridging the observed and full data likelihoods. However, sampling values of the unmeasured confounder cannot be handled using Gibbs steps since the two full conditional distributions are not of standard forms. Hence, we need to use a Metropolis-Hastings step within the overall MCMC sampler to generate these values [16]. The priors that we use for the unmeasured confounder can be used as the proposal distributions. In the case of the Gaussian prior, if we had directly controlled for the covariates rather than the generalized propensity score at the outcome stage, we could leverage conjugacy and directly use Gibbs steps, as in [14].

The analyst has to make two key choices here: the forms used for $\mathbf{g}_{\cdot}(\mathbf{U})$ and the sensitivity parameters selected ($\mathcal{L}_Y$ and $\mathcal{L}_W$). These choices are both crucial since they represent *how* unmeasured confounding factors into the treatment and outcome stages. These must be thoughtfully selected to fully understand the robustness of the analysis to unmeasured confounding in substantive ways. We propose using a flexible form for $\mathbf{g}_{\cdot}(\mathbf{U})$ e.g. as a cubic polynomial function in the case of continuous unmeasured confounding and setting the sensitivity parameters as a grid around a range of values based on the average values of the other coefficients in the model. For instance, in our motivating application, we expected that the response variable would exhibit a slow burn-in during childhood, followed by rapid growth and then a slowdown in adult years as a function of various developmental growth metrics. As a consequence, a cubic polynomial function was a natural choice.

.7 Figures for Sensitivity Analysis



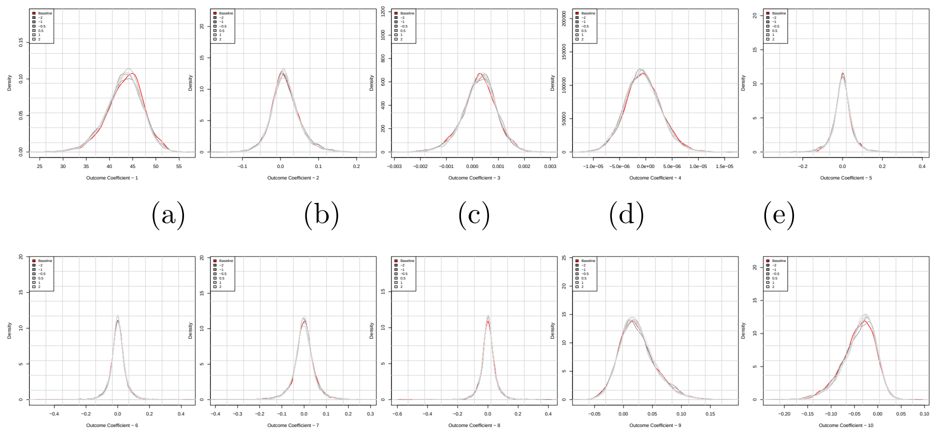
(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure .4: Kernel density estimation of coefficients of the treatment stage model under the polynomial unmeasured confounder with $\mathcal{L}_Y = 0.25$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.

.8 Summarization Diagnostics

In the application of our summarization strategy, we proposed the use of the L_2 norm as the distance measure for our analysis. However, depending on the differences for the dose-response function drawn in general, it may be that this is an inappropriate choice. Therefore, as a means of validation, we want to assess what the *shape* of our differences is.



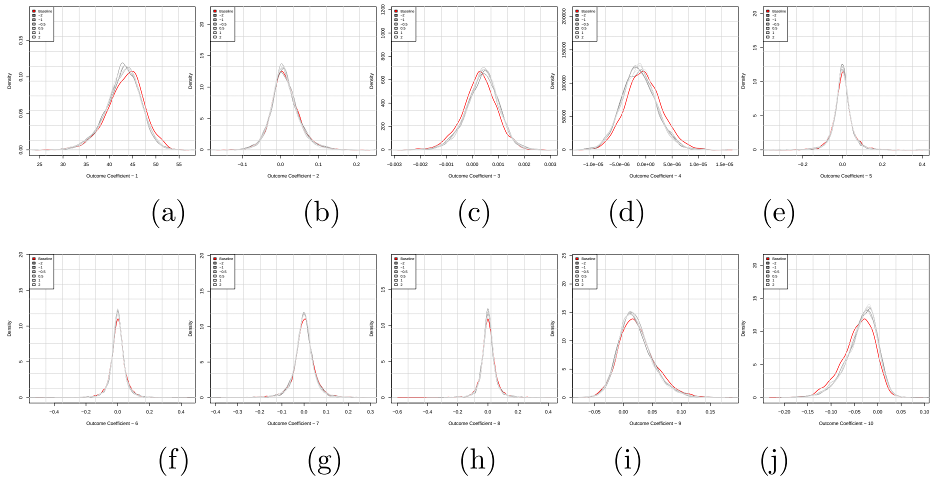
(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure .5: Kernel density estimation of coefficients of the treatment stage model under the polynomial unmeasured confounder with $\mathcal{L}_Y = 0.5$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.

In Figure .12 we present visual diagnostics for $\mathcal{L}_Y = 2$ and $\mathcal{L}_W = 2$, under the binary and continuous unmeasured confounding that we considered. The top panels ((a) and (b)) show the distribution of the distances relative to the radius, $\mathcal{R}$, while the bottom panels ((c) and (d)) show the projection onto the principal components to make the shape of the simulated differences easy to visualize.

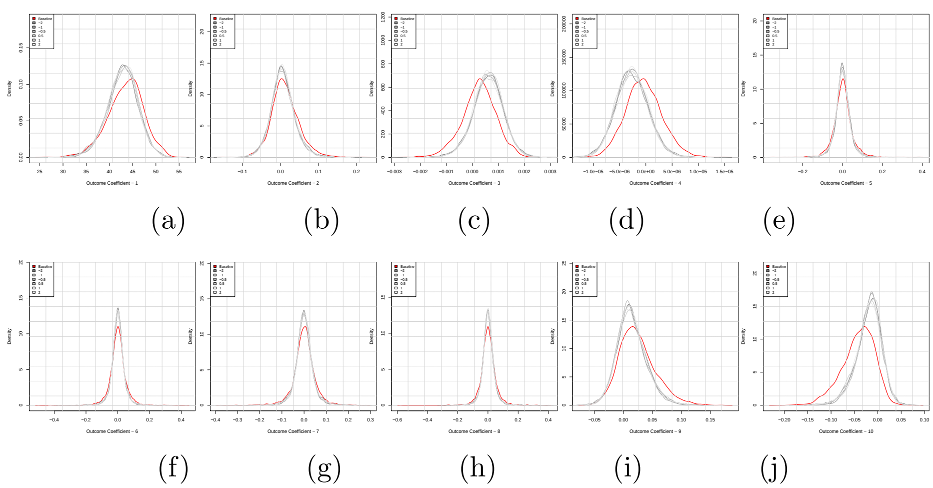
First, using the plots in figure .12(a) and (b), we can see the differences in the extreme case between the binary and continuous cases of unmeasured confounding. Relative to the continuous case, considerably more points are concentrated within the ball, indicating that that there is more impact of unmeasured confounding in the latter. Second, in both the projections presented in figure .12(c) and (d), we can assess the shape of the distances, and determine whether the shape generated by the L_2 norm is appropriate for the application. As we can see, the points are approximately distributed in a ball-like pattern, which validates our use of the L_2 norm in our application.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

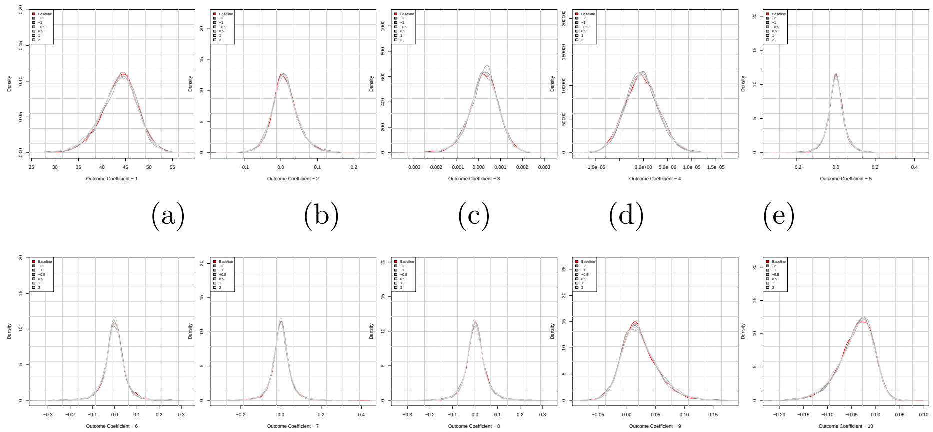
Figure .6: Kernel density estimation of coefficients of the treatment stage model under the polynomial unmeasured confounder with $\mathcal{L}_Y = 1$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

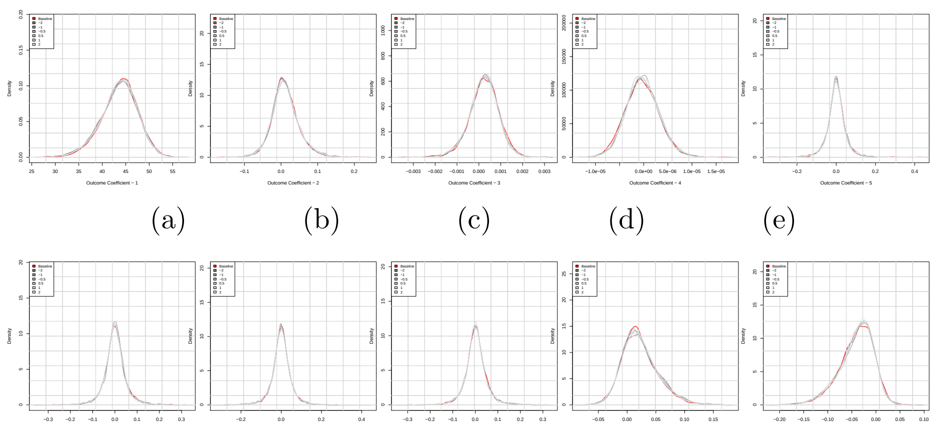
Figure .7: Kernel density estimation of coefficients of the treatment stage model under the polynomial unmeasured confounder with $\mathcal{L}_Y = 2$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

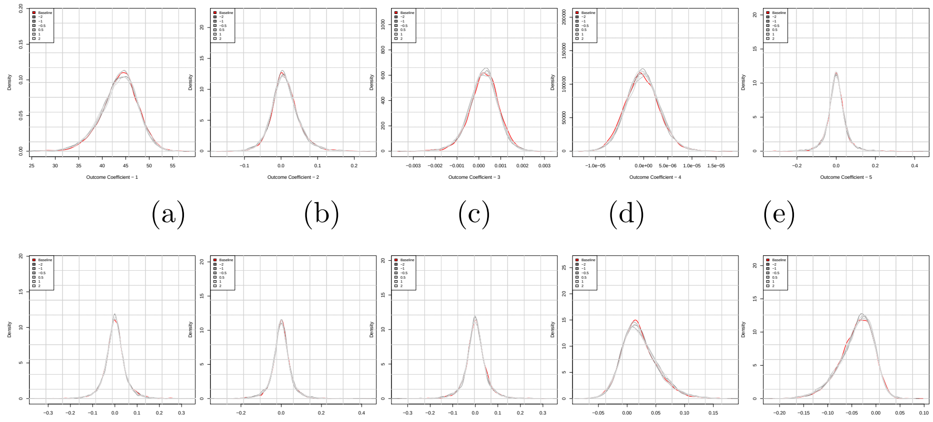
Figure .8: Kernel density estimation of coefficients of the treatment stage model under the binary unmeasured confounder with $\mathcal{L}_Y = 0.25$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

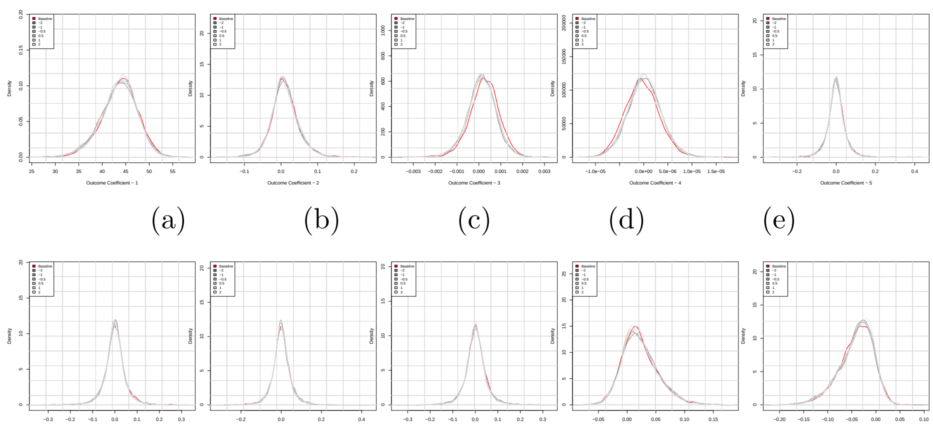
Figure .9: Kernel density estimation of coefficients of the treatment stage model under the binary unmeasured confounder with $\mathcal{L}_Y = 0.5$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure .10: Kernel density estimation of coefficients of the treatment stage model under the binary unmeasured confounder with $\mathcal{L}_Y = 1$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure .11: Kernel density estimation of coefficients of the treatment stage model under the binary unmeasured confounder with $\mathcal{L}_Y = 2$ (a) Intercept (b) Linear treatment term (c) Quadratic treatment term (d) Cubic treatment term (e) Linear GPS term (f) Quadratic GPS term (g) Cubic GPS term (h) Interaction Term (i) Height Term (j) Weight Term.

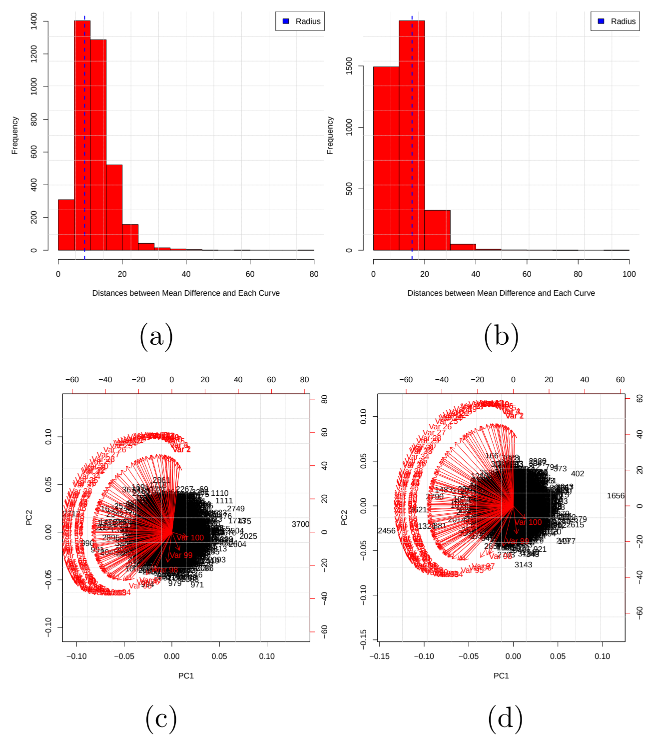


Figure .12: Computing probabilistic differences using simulation (a) relative distances versus the radius in the polynomial case (b) relative distances versus the radius in the binary case (c) projection of sampled onto the principal components in the polynomial case (d) projection of sampled differences onto the principal components in the binomial case.

Bibliography

[1] ABPP ABPN. Female Domestic Violence Offenders: Their Attachment Security, Trauma Symptoms, and Personality Organization. *Violence and Victims*, 22(5):532, 2007.

[2] Paul D Allison. *Missing Data*, volume 136. Sage Publications, 2001.

[3] Susan Athey. Machine Learning and Causal Inference for Policy Evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 5–6, New York, NY, USA, 2015. ACM.

[4] Susan Athey and Guido Imbens. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[5] Susan Athey and Guido W Imbens. Machine Learning Methods for Estimating Heterogeneous Causal Effects. *Stat*, 1050:5, 2015.

[6] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests. *arXiv preprint arXiv:1610.01271*, 2016.

[7] Samuel Ball et al. Reading with Television: An Evaluation of The Electric Company. A Report to the Children's Television Workshop. volumes 1 and 2. 1973.

[8] Roy F Baumeister. *Self, Psychology of.* Wiley Online Library, 1999.

[9] Anna Bensley, Jin Huh, Deepthi Nacharaju, Sri Sridharan, Sumedha Ariely Gupta, Kiran Modi, Shyna Bhattacharjee, Monisha Akhtar, Riti Chandrashekar, Anshita Jindal, and Kashni Channa. Duke SRTP India 2016 Report for Udayan Care. 2016.

[10] Alina Beygelzimer and John Langford. The Offset Tree for Learning with Partial Labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138. ACM, 2009.

[11] Leo Breiman. Classification and Regression Trees. 1984.

[12] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Barbara M Byrne and Richard J Shavelson. On the Structure of Adolescent Self-concept. *Journal of Educational Psychology*, 78(6):474, 1986.

[14] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

[15] James W Chapman, Robert Lambourne, and Phil A Silva. Some Antecedents of Academic Self-concept: A Longitudinal Study. *British Journal of Educational Psychology*, 60(2):142–152, 1990.

[16] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

[17] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[18] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric Tests for Treatment Effect Heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

[19] Rajeev H Dehejia and Sadek Wahba. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

[20] Peng Ding and Fan Li. Causal Inference: A Missing Data Perspective. *arXiv preprint arXiv:1712.06170*, 2017.

[21] Miroslav Dudík, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Learning. *arXiv preprint arXiv:1103.4601*, 2011.

[22] Kathleen M Dwyer, Bridget K Fredstrom, Kenneth H Rubin, Cathryn Booth LaForce, Linda Rose Krasnor, and Kim B Burgess. Attachment, Social Information Processing, and Friendship Quality of Early Adolescent Girls and Boys. *Journal of Social and Personal Relationships*, 27(1):91–116, 2010.

[23] David J Earl and Michael W Deem. Parallel Tempering: Theory, Applications, and New Perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[24] Carlos A Flores, Alfonso Flores Lagunes, Arturo Gonzalez, and Todd C Neumann. Estimating the Effects of Length of Exposure to Instruction in a Training Program: the Case of Job Corps. *Review of Economics and Statistics*, 94(1):153–171, 2012.

[25] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup Identification From Randomized Clinical Trial Data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[26] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel-hierarchical Models*, volume 1. Cambridge University Press New York, NY, USA, 2007.

[27] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multi-level/hierarchical models*, volume Analytical methods for social research. Cambridge University Press, New York, 2007.

[28] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[29] Michael Gerfin and Michael Lechner. A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland. *The Economic Journal*, 112(482):854–893, 2002.

[30] Asghar Ghasemi and Saleh Zahediasl. Normality Tests for Statistical Analysis: A Guide for Non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486, 2012.

[31] Donald P Green and Holger L Kern. Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Aadditive Regression Trees. *Public Opinion Quarterly*, 76(3):491–511, 2012.

[32] Sander Greenland. Basic Methods for Sensitivity Analysis of Biases. *International Journal of Epidemiology*, 25(6):1107–1116, 1996.

[33] Zijian Guo and Dylan S Small. Control Function Instrumental Variable Estimation of Nonlinear Causal Effect Models. *Journal of Machine Learning Research*, 17(100):1–35, 2016.

[34] P Gustafson, LC McCandless, AR Levy, and S Richardson. Simplified Bayesian Sensitivity Analysis for Mismeasured and Unobserved Confounders. *Biometrics*, 66(4):1129–1137, 2010.

[35] P Richard Hahn, Carlos M Carvalho, David Puelz, Jingyu He, and Others. Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, 13(1):163–182, 2018.

[36] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian Regression Tree Models for Causal inference: Regularization, Confounding, and Heterogeneous Effects. 2017.

[37] Tibshirani Hastie and R Tibshirani. & friedman, j.(2008). the Elements of Statistical Learning; Data mining, Inference and Prediction.

[38] James J Heckman and Jeffrey A Smith. The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies. *The Economic Journal*, 109(457):313–348, 1999.

[39] James J Heckman, Sergio Urzua, and Edward Vytlacil. Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006.

[40] Ruth Heller, Shane T Jensen, Paul R Rosenbaum, and Dylan S Small. Sensitivity Analysis for the Cross-match Test, with Applications in Genomics. *Journal of the American Statistical Association*, 105(491):1005–1013, 2010.

[41] Jennifer L Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[42] Keisuke Hirano and Guido W Imbens. The Propensity Score with Continuous Treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, pages 73–84, 2005.

[43] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189, 2003.

[44] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.

[45] Michael G Hudgens and M Elizabeth Halloran. Toward Causal Inference with Interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[46] Kosuke Imai and David A Van Dyk. Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

[47] Guido W Imbens. The Role of the Propensity Score in Estimating Dose-response Functions. *Biometrika*, 87(3):706–710, 2000.

[48] Guido W Imbens. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[49] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[50] Guildo W Imbens. Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review*, 93(2):126–132, 2003.

[51] James E Johndrow, Jonathan C Mattingly, Sayan Mukherjee, and David Dunson. Approximations of Markov Chains and Bayesian Inference. *arXiv preprint arXiv:1508.03387*, 2015.

[52] Richard A Levine and George Casella. Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

[53] Fan Li, Alan M Zaslavsky, and Mary Beth Landrum. Propensity Score Analysis with Hierarchical Data. In *Proceedings of the American Statistical Association Joint Statistical Meetings*, 2007.

[54] Fan Li, Alan M Zaslavsky, and Mary Beth Landrum. Propensity Score Weighting with Multilevel Data. *Statistics in Medicine*, 32(19):3373–3387, 2013.

[55] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[56] Wei-Yin Loh. Classification and Regression Trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

[57] Jared K Lunceford and Marie Davidian. Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: a Comparative Study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[58] Lawrence C McCandless, Paul Gustafson, and Adrian Levy. Bayesian Sensitivity Analysis for Unmeasured Confounding in Observational Studies. *Statistics in Medicine*, 26(11):2331–2347, 2007.

[59] Andrea Mercatanti, Fan Li, et al. Do Debit Cards Increase Household Spending? Evidence from a Semiparametric Causal Analysis of a Survey. *The Annals of Applied Statistics*, 8(4):2485–2508, 2014.

[60] C.Z. Mooney, R.D. Duval, and R. Duvall. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Number nos. 94-95 in Bootstrapping: A Nonparametric Approach to Statistical Inference. SAGE Publications, 1993.

[61] Judea Pearl. Myth, Confusion, and Science in Causal Analysis. 2009.

[62] Judea Pearl et al. Causal Inference in Statistics: an Overview. *Statistics Surveys*, 3:96–146, 2009.

[63] Ellen V Piers and Dale B Harris. Children's Self-concept Scale. *Nashville, Tenn.: Counselor Recordings and Tests*, 1969.

[64] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some Methods for Heterogeneous Treatment Effect Estimation in High-Dimensions. *arXiv preprint arXiv:1707.00102*, 2017.

[65] Matthew T Pratola and Others. Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian Analysis*, 11(3):885–911, 2016.

[66] Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems*, pages 554–560, 2000.

[67] Carl Edward Rasmussen. Gaussian Processes for Machine Learning. 2006.

[68] Robert McCulloch, Rodney Sparapani, Robert Gramacy, Charles Spanbauer, and Matthew Pratola. *BART: Bayesian Additive Regression Trees*, 2018. R package version 1.9.

[69] Paul R Rosenbaum. Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies. *Biometrika*, 74(1):13–26, 1987.

[70] Paul R Rosenbaum. Sensitivity to Hidden Bias. In *Observational Studies*, pages 105–170. Springer, 2002.

[71] Paul R Rosenbaum. An Exact Distribution-free Test Comparing Two Multivariate Distributions Based on Adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

[72] Paul R Rosenbaum. Design Sensitivity and Efficiency in Observational Studies. *Journal of the American Statistical Association*, 105(490):692–702, 2010.

[73] Paul R Rosenbaum and Donald B Rubin. *Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome*. Wisconsin Clinical Cancer Center, Biostatistics, 1982.

[74] Paul R Rosenbaum and Donald B Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

[75] Paul R Rosenbaum and Donald B Rubin. Reducing Bias in Observational Studies using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[76] Donald B Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, pages 34–58, 1978.

[77] Donald B Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[78] Donald B Rubin and Neal Thomas. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, pages 249–264, 1996.

[79] Kenneth H Rubin, Kathleen M Dwyer, Cathryn Booth LaForce, Angel H Kim, Kim B Burgess, and Linda Rose Krasnor. Attachment, Friendship, and Psychosocial Functioning in Early Adolescence. *The Journal of Early Adolescence*, 24(4):326–356, 2004.

[80] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

[81] Susheela Singh. Gaussian Process Regression. 2016.

[82] Kyle Steenland and Sander Greenland. Monte Carlo Sensitivity Analysis and Bayesian Analysis of Smoking as an Unmeasured Confounder in a Study of Silica and Lung Cancer. *American Journal of Epidemiology*, 160(4):384–392, 2004.

[83] Susan Athey, Guido Imbens, and Yanyang Kong. *causalTree: Recursive Partitioning Causal Trees*, 2016. R package version 0.0.

[84] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On Causal Inference in the Presence of Interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[85] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2018. R package version 4.1-13.

[86] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[87] Hal R Varian. *Intermediate Microeconomics: A Modern Approach: Ninth International Student Edition*. WW Norton & Company, 2014.

[88] Michele Vecchione, Guido Alessandri, Claudio Barbaranelli, and Maria Gerbino. Stability and Change of Ego Resiliency From Late Adolescence to Young Adulthood: A Multiperspective Study Using the Er89–R Scale. *Journal of Personality Assessment*, 92(3):212–221, 2010.

[89] Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

[90] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

[91] T Wendling, K Jung, A Callahan, A Schuler, NH Shah, and B Gallego. Comparing Methods for Estimation of Heterogeneous Treatment Effects using Observational Data from Health Care Databases. *Statistics in Medicine*, 2018.

[92] Ian R White, Patrick Royston, and Angela M Wood. Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, 30(4):377–399, 2011.

[93] Yu Xie, Jennie E Brand, and Ben Jann. Estimating Heterogeneous Treatment Effects with Observational data. *Sociological Methodology*, 42(1):314–347, 2012.

[94] Shu Yang. Propensity Score Weighting for Causal Inference with Multi-stage Clustered Data. *arXiv preprint arXiv:1607.07521*, 2016.

[95] Corwin M Zigler, Krista Watts, Robert W Yeh, Yun Wang, Brent A Coull, and Francesca Dominici. Model Feedback in Bayesian Propensity Score Estimation. *Biometrics*, 69(1):263–273, 2013.

Biography

Abbas Zaidi began his undergraduate studies at the University of Pittsburgh in August 2008, where he originally pursued a degree in economics and journalism. During the his sophomore year, he transferred to the economics department at Carnegie-Mellon University where he received his Bachelors of Science in Economics and Statistics with University Honors in 2012. During his time as an undergraduate he received the Luke Edwards Research Fellowship, the Distinguished Services in Economics Award and was granted membership to Omicron Delta Epsilon. The following year, he earned his Masters of Statistical Practice.

In August 2015, he began his graduate studies in the Department of Statistical Science at Duke University. His doctoral research, advised by Drs. Sayan Mukherjee and Fan Li, focused on the development of Bayesian methodology for applications in Causal Inference.