A NONPARAMETRIC BAYESIAN APPROACH TO CAusal MODELLING

by

Tim Henry Guimond

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Public Health Sciences University of Toronto

© Copyright 2018 by Tim Henry Guimond
Abstract

A Nonparametric Bayesian Approach to Causal Modelling

Tim Henry Guimond
Doctor of Philosophy
Department of Public Health Sciences
University of Toronto
2018

The Dirichlet process mixture regression (DPMR) method is a technique to produce a very flexible regression model using Bayesian principles based on data clusters. The DPMR method begins by modelling the joint probability density for all variables in a problem. In observational studies, factors which influence treatment assignment (or treatment choice) may also be factors which influence outcomes. In such cases, we refer to these factors as confounders and standard estimates of treatment effects will be biased. Causal modelling approaches allow researchers to make causal inferences from observational data by accounting for confounding variables and thus correcting for the bias in unadjusted models. This thesis develops a fully Bayesian model where the Dirichlet process mixture models the joint distribution of all the variables of interest (confounders, treatment assignment and outcome), and is designed in such a way as to guarantee that this clustering approach adjusts for confounding while also providing a flexible model for outcomes. A local assumption of ignorability is required, as contrasted with the usual global assumption of strong ignorability, and the meaning and consequences of this alternate assumption are explored. The resulting model allows for inferences which are in accordance with causal model principles.

In addition to estimating the overall average treatment effect (mean difference between two treatments), it also provides for the determination of conditional outcomes, hence can predict a region of the covariate space where one treatment dominates. Furthermore, the technique’s capacity to examine the strongly ignorable assumption is demonstrated. This method can be harnessed to recreate the underlying counterfactual distributions that produce observational data and this is demonstrated with a simulated data set and its results are compared to other common approaches. Finally, the method is applied to a real life data set of an observational study of two possible methods of integrating mental health treatment into the shelter system for homeless men. This analysis of this data demonstrates a situation where treatments have identical outcomes for a subset of the covariate space and a subset of the space where one treatment clearly dominates, thereby informing an individualized patient driven approach to treatment selection.
I dedicate this thesis to my family, birth and chosen. My mother, father, sister, husband, and friends have supported me throughout many unusual journeys. My father’s premature death from cancer during my medical school training was simultaneously painful and essential to my growth as a physician. The discussions we had in the last month of his life helped form my professional identity and I am thankful for this. Disrupting my educational plans to complete a MD/PhD earlier, ultimately helped me clarify the values and path. While I miss him on so many occasions, his laugh, his dedication to values and humanity continuously have served as an anchor for me. My mother’s strength to forge her own path despite the struggles of her childhood have served as an example to me throughout my life. She instilled a strong curiosity in me that I will always cherish. This has been a great gift that she has given me, and I am entirely in her debt.
Acknowledgements

First I wish to express my gratitude to my husband, Elliot Alexander. He has now seen me through more degrees and training than any respectable person would be willing to tolerate. I promise that this is a terminal degree and I will seek no further degrees. As previously discussed, courses in new areas are fair game. For many years, Elliot has been witness to the confusion and uncertainty with which I approach my career. The marriage of statistics and psychiatry is unusual and he provides me with a good dose of grounding, regularly asking what I hope to achieve. While I have no clear answer for him and I am certain that this is extremely frustrating for him, his inquiry helps me stay focused on a purpose despite the strange bedfellows that are my interests. Our love, in this burgeoning age of acceptance of homosexuality, has provided me with the strength to weather many storms, where so many have felt isolated and lost. So much I owe to Elliot, he has stood by me through so much, and I feel so much better about my quirks because of him.

This research was supported by a grant from the Ontario HIV Treatment Network (OHTN). I am very much indebted to the community of HIV researchers in general and the OHTN network specifically for supporting me through this PhD. It is quite astonishing and gratifying that HIV researchers were willing to provide me with this opportunity to explore a mathematical discipline with the hopes that I will contribute to a much needed cure for this devastating illness. May I one day repay this gesture and pay forward the hopefulness to a new generation, even if we are not successful in eradicating HIV.

I also wish to acknowledge the assistance and companionship of my fellow students, Sudipta, Katherine, Osvaldo, Kuan, Konstantin, and Mohsen. I also wish to thank my patients who have bared the brunt of the disruptions of my availability. Prof. Dionne Gesink has been an incredible mentor and friend throughout this process, and I thank her immensely for all the panic moments that she was me through. Dr. Yvonne Bergmans was a key colleague and friend who I had promised to complete my PhD along with her, while I delayed, she did not. She has kept me focused and accountable at key moments.

Finally, Prof. Michael Escobar has given me an incredible gift through his supervision. Exploring at the edges of an applied mathematical discipline is not for the faint of heart, and his unique perspective on statistical problems and insightful visualization capacities have broadened my thinking immensely. He has made another Bayesian statistician, and I certainly hope the International Society for Bayesian Analysis has some sort of kick-back program that gives him a nice gift in exchange. I feel I still have much to learn from him and selfishly hope for many years of collaboration.
Contents

Acknowledgements iv

Table of Contents v

List of Tables vii

List of Figures viii

1 Introduction 1

2 Background 4

2.1 Causal Models: Previous theoretical principles 4

2.1.1 Counterfactuals 5

2.1.2 Causal models without counterfactuals 6

2.1.3 Balancing scores and Propensity scores 6

2.1.4 Strongly ignorable and positivity assumptions 7

2.1.5 Single Unit Treatment Value Assumption - SUTVA 8

2.1.6 Deterministic versus stochastic counterfactuals 9

2.2 Causal Models: Previous Applied Methods 10

2.2.1 Rubin causal model 10

2.2.2 Previous Non-Parametric Models 11

2.2.3 Previous Bayesian Causal Models 13

2.3 Non-parametric models: Using the Dirichlet process mixture as a regression model 15

2.3.1 The Dirichlet process prior 15

2.3.2 Stick-breaking implementation of the Dirichlet prior 15

2.3.3 Dirichlet process in joint probability models 16

2.3.4 Regression framework 17

2.4 Integrating a semi-parametric regression modelling approach with causal principles 18

3 Proposed Dirichlet Process Mixture Regression Approach to Causal Modelling 20

3.1 Assumptions for causal models 21

3.1.1 Structural assumption 21

3.2 Linking the DPMR to Causal Modelling principles 23

3.3 Creating causal estimates of the average treatment effect from a DPMR model 23

3.3.1 Impact of the structural assumption of weak ignorability 24
List of Tables

4.1 Simulation 1: Average treatment effect estimates ........................................... 39
4.2 Simulation 1: Region of the covariate X where treatment 1 is preferred .......... 44
4.3 Simulation 2: Average treatment effect estimates ........................................... 54

5.1 Baseline demographic and clinical characteristics ........................................... 72
5.2 Propensity Score Models .................................................................................. 75
5.3 Overall treatment effect ................................................................................... 76
List of Figures

2.1 A schematic diagram of a causal model under strong ignorability. .................................................. 7

3.1 A schematic diagram of the first assumption. Squares are calculated, and circles are randomly generated conditional on the values of the parent values. ............................... 22

3.2 Left: Response curves for each treatment. Right: Propensity and distribution of \( X \). .......... 24

3.3 A schematic diagram of the consequences of both the structural and modelling assumptions. ... 27

4.1 Simulation 1: Left: Response curves for each treatment. Right: Propensity and distribution of \( X \). ................................................................. 34

4.2 Simulation 1: A schematic diagram of the simulated dataset. Squares are calculated, and circles are randomly generated conditional on the values of the parent values. ........ 34

4.3 Left: Response curves under each treatment by gender. Right: Propensity and distribution of \( X \) by gender. ................................................................. 36

4.4 A schematic diagram of the simulated dataset. Squares are calculated, and circles are randomly generated conditional on the values of the parent values. ......................... 36

4.5 Simulation 1: Average Treatment Effect (ATE) Estimates. The solid red line represents the true treatment effect, and the dashed red line represents the treatment effect in this particular sample. All methods produce very similar effect estimates with similar credible regions and confidence intervals that include both the true and sample treatment effects. 40

4.6 Simulation 1: Cluster occupancy - the number of clusters that are assigned at least one data point in an iteration is calculated and plotted in a histogram, the mean occupancy of each chain is calculated and noted in each plot. The histograms range from \( C=10 \) clusters (left side) to \( C=40 \) clusters (right side). ........................................ 41

4.7 Simulation 1: Propensity score - \( pr(Z \mid X) \) estimates ranging from \( C=10 \) clusters (left side) to \( C=40 \) clusters (right side). The pink line represents the true propensity score, the blue line is the fitted estimated propensity, and the black lines represent the 95% credible region of this. The raw data used for this simulation is divided into groups and plotted as those assigned to treatment 1 (green points above the graph) and treatment 0 (salmon points below the graph). The estimated propensity well follows the true propensity and is contained within the credible region throughout. ........................................ 42
4.8 Simulation 1: Counterfactual curves for $R_0$ (top row) and $R_1$ (bottom row) ranging from 10 clusters (left side) to 40 clusters (right side). The pink line represents the true conditional mean effect, the blue line is the fitted conditional expectation value and the black lines represent the 95% credible region. The raw data used for this simulation is plotted as grey points. The $R_0$ credible region estimate nearly contains the true values at all regions in $X$ except for a small region between 70 and 74 years of age. The $R_1$ credible region contains the true curve throughout. It is interesting to note that in areas with low amounts of data, the $R_0$ estimate is further away from the true value, but has larger credible regions.

4.9 Simulation 1: **Left**: Response curves under both treatments. **Right**: Expected difference in treatments by $X$.

4.10 Simulation 1: Distribution of $X$ (age) density estimates ranging from $C = 10$ clusters (left side) to $C = 40$ clusters (right side). The pink line represents the true density, the blue line is the estimated density, and the black lines represent the 95% credible region of the height of the density at each value of $X$. The raw data is plotted below the graph with random noise added to aid in the differentiation of regions of high and low density of observations. The estimated density is quite close to the true density and is contained within the credible region throughout. The lack of smoothness in the density estimate is likely induced by the need of the model also to capture the response variables accurately.

4.11 Simulation 1: Distribution of $R_0 \mid X$ (top row) and $R_1 \mid X$ (bottom row) from $C = 10$ clusters (leftmost two) and $C = 40$ clusters (rightmost two) plotted both as contour plots and density slices. In the plots of the slices of the density function, the black line represents the model estimate and the red line the true underlying conditional density. Increasing the number of clusters does not appear to improve the fit of the density functions significantly. However, these plots give a clue to how one might be able to provide a clinician with the potential responses for a particular age superimposed from each curve $R_1$ and $R_0$ on the same graph to aid in decision making and to provide a more realistic view of what the possible outcomes may be.

4.12 Simulation 1: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using each iteration’s density estimate. The true mode is plotted with the blue line, and the darkness of points represent the number of times that point was selected as the mode across all the iterations. The cloud of points generally includes the true mode, however, is less smooth and is therefore likely following the simulated more closely.

4.13 Simulation 1: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using the mean density estimate from $C=10$ clusters (left side) through $C=40$ clusters (right side). The true mode is plotted with the blue line. Due to the use of the grid to search for the mode, we see that the estimate is jagged; however, we can see that the mode is tracking along the expected path.
4.14 Simulation 1: Pointillism plots of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) from $C = 10$ clusters (left side) through $C = 40$ clusters (right side). The true conditional mean is plotted with the blue line. The cluster means $(\mu^X_1, \mu^R_1)$ are plotted using a grey scale so that more highly probable clusters (ones likely to have more membership) are darker and lower probability clusters are lighter. These plots are capturing information about the distribution of $X$ with lower ages having more membership due to the density of $X$, balanced with the probabilities of treatment assignment. Since the estimates arise from smoothing over many clusters, we see that the estimates derived from the models with more clusters ($C = 40$) appear to be closer to the true values.

4.15 Simulation 1: Pointillism plots of $Z \mid X$ from $C=10$ clusters (left side) to $C=40$ clusters (right side). The true propensity score is plotted in blue. Darker points represent clusters with a higher probability of group membership. Most of the dark clusters are close to the true propensity, and in high probability regions for age appear to be symmetric around the true propensity curve.

4.16 Simulation 1: Pointillism plots of treatment difference $R_1 - R_0 \mid X$ from $C=10$ clusters (left side) to $C=40$ clusters (right side). The true conditional expectation of treatment difference is plotted in blue. The cluster means appear to be symmetrically distributed around the true difference, the clustering at some values of $X$ may represent how the curves can be well estimated as a straight line between these clusters, for instance between the ages of 64 and 68 years on all the last four plots.

4.17 Simulation 2 - Comparing Average Treatment Effect (ATE) estimates from various methods. The solid red line represents the ‘true’ treatment effect (ignoring gender), and the dashed red line represents the treatment effect in this particular sample. All the estimates from each method have confidence intervals and credible regions that include the true and sample treatment effects. The DPMR estimates produce effect estimates that improve (are more closely centred around the true effect) with larger cluster size. The propensity score and ANCOVA methods produce very similar effect estimates with similar confidence intervals that include both the true and sample treatment effects. These standard frequentist estimates are more similar to the small cluster results from the DPMR method. Finally, the IPTW methods produce estimates with substantial confidence intervals.

4.18 Simulation 2: Cluster occupancy - the number of clusters that are assigned at least one data point in an iteration is calculated and plotted in a histogram, the mean occupancy of each chain is calculated and noted in each plot. The histograms range from $C=10$ clusters (left side) to $C=40$ clusters (right side).

4.19 Simulation 2: Propensity score - $pr(Z \mid X)$. The pink line represents the true propensity score for women and the blue line the true propensity score for men, the green line is the fitted estimated propensity from the model (which cannot distinguish these two groups), and the black lines represent the 95% credible region of this estimate. The raw data used for this simulation is divided into groups and plotted as those assigned to treatment 1 (green points above the graph) and treatment 0 (salmon points below the graph). Since the true propensity cannot be determined, it would seem that with increasing clusters we find more fluctuations in the estimate with a broadening credible region.
4.20 Simulation 2: Counterfactual curves for $R_0$ (top) and $R_1$ (bottom) beginning with $C = 10$ clusters on the left side through $C = 40$ clusters on the right. The green line represents the true conditional mean response for women and the red line the true response for men, the blue line is the fitted conditional mean response from the model (which cannot distinguish these two groups) and the black lines represent the 95% credible region of this estimate. The raw data is plotted in grey. Similar to the propensity score estimate, the estimate begins fluctuating more (vacillating between the two true curves) and the credible region widens with a larger number of clusters.

4.21 Expected difference between treatments by the covariate $X$. The true difference for men is plotted in red, the true difference for women in green and the fitted model in blue. The counterfactual differences (given that both were simulated initially, but only one was pretended to be known) are plotted in grey.

4.22 Simulation 2: Distribution of $X$ (age).

4.23 Simulation 2: Distribution of $R_0 \mid X$ (top row) and $R_1 \mid X$ (bottom row) from $C = 10$ clusters (leftmost two) and $C = 40$ clusters (rightmost two) plotted both as contour plots and density slices (conditional distributions of the outcome at a specific covariate value plotted vertically to the right of the covariate value). In the plots of the slices of the density function, the black line represents the model estimate and the red line the true underlying conditional density. In contrast with simulation 1, here increasing the number of clusters appears to improve the fit of the density functions significantly. Again, these plots could be very useful for a clinician to demonstrate the potential responses at a particular age from each treatment if we superimposed the $R_1$ curve and $R_0$ curve on the same graph. To a researcher, it would also clearly signal that there is a bimodal response for which it would be important to identify a further predictor.

4.24 Simulation 2: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using the mean density estimate from $C=10$ clusters (left side) through $C=40$ clusters (right side). The pink line represents the true counterfactual response for women $E(R_1 \mid X, X_g = w)$ and the blue line the true counterfactual response for men $E(R_1 \mid X, X_g = m)$. Here many small ridges appear as artifacts of attempts by the model to find clusters of data at various values of $X$ that may still create ‘ripples’ in the density. The number of these artifactual modes increase with increased cluster size and suggest that visually weighting the modes by the overall height might be useful.

4.25 Simulation 2: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using each iteration’s density estimate from $C=10$ clusters (left side) through $C=40$ clusters (right side). The pink line represents the true counterfactual response for women $E(R_1 \mid X, X_g = w)$ and the blue line the true counterfactual response for men $E(R_1 \mid X, X_g = m)$. In this estimate, we can more clearly see that the modes at each iteration are most often finding the true underlying modes to this joint distribution. With greater cluster size, these modes are clustering much more closely to the true values.
4.26 Simulation 2: Pointillism plots of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) from $C = 10$ clusters (left side) through $C = 40$ clusters (right side). The true conditional mean for men is plotted with the blue line and for women with pink. The cluster means $(\mu_{Xj}, \mu_{Rj})$ are plotted using a grey scale so that more highly probable clusters (ones likely to have more membership) are darker and lower probability clusters are lighter. These plots are demonstrating that the model is capturing information about the joint distribution and its bimodal nature. At small cluster sizes the distinction between the modes is much less clear; however, we see that the estimates derived from the models with the most clusters ($C = 40$) that the model is identifying clusters close to the true curves.

4.27 Simulation 2: Pointillism plots of $Z \mid X$ from $C=10$ clusters (left side) to $C=40$ clusters (right side). The true propensity score for men is plotted in blue and for women in pink. Darker points represent clusters with higher probability of group membership. In this simulation, the cluster centres are very diffuse, and high probability clusters can be found throughout. This diffusion can help to identify the possibility of a missing confounder.

4.28 Simulation 2: Pointillism plots of treatment difference $R_1 - R_0 \mid X$ from $C=10$ clusters (left side) to $C=40$ clusters (right side). The true difference for men is plotted in blue and for women in pink. The model cannot distinguish which subgroups match and as the cluster size increases, we can see four distinct lines beginning to form as male and female subgroups for treatment 0 and 1 match and cross with each other.

5.1 Estimated density of the logarithm of years homeless by treatment. The red line is the estimate for men in the IMCC arm and the blue line for men receiving SOCC.

5.2 Outcome, days homeless, by treatment and years homeless. The red points represent data from men assigned to the IMCC arm and the blue points for men receiving SOCC. The points plotted below the solid line in the diagram represent observations where outcomes were not available, but the initial baseline lifetime homelessness data was measured.

5.3 These plots contain three elements: propensity score estimates in strata by quintiles of lifetime homelessness (represented by points with error bars centred within each stratum), observed treatment assignment (represented by tick marks either at the top (IMCC) or bottom (SOCC) of the plot, and predicted propensity score models are plotted as curves. Each plot uses progressively higher order covariates ($x, x^2, x^3$) for these propensity models, from a linear model on the left to a cubic model on the right.

5.4 Cluster occupancy - the number of clusters that are assigned at least one data point in an iteration is calculated and plotted in a histogram, the mean occupancy of each chain is calculated and noted in the plot.

5.5 Left: Distribution of X. Right: Propensity score - $Pr(Z \mid X)$.

5.6 Left: Response curve under SOCC, $R_0$. Right: Response curve under IMCC, $R_1$.

5.7 Left: Predicting $R_0$ and $R_1$. Right: Predicting Treatment difference $E[R_1 - R_0 \mid X]$.

5.8 Left: Local neighbourhood centres. Right: Density estimate. Top: SOCC. Bottom: IMCC.

5.9 DPMR Propensity Probability Clusters.
Chapter 1

Introduction

There has been a considerable amount of research on causal modelling. Considerable attention has been given to the Rubin causal model (RCM) since the original papers 40 years ago ([47, 48, 43]); however, few articles have used a fully Bayesian approach to the problem with some notable exceptions, which will be reviewed in Chapter 2. This thesis will introduce a nonparametric Bayesian model of the joint distribution of all variables considered in the causal relationship (response, treatment assignment and confounders) with a Dirichlet process mixture prior. This thesis explores this model by considering only a very rudimentary form of the problem with the minimal number of covariates required. While recognizing that many approaches are capable of dealing with multiple confounders and predictors, this thesis attempts only to demonstrate another approach built up rigorously from basic principles to demonstrate possible advantages of an alternative framework in certain situations. As a note, we are considering a modelling environment where the results are intended to inform clinical decision making; ones where a small set of covariates are available to a clinician and where a treatment decision will be made for a particular individual between a small subset of possible options. It builds on the intentions of patient-centred health care. Health care providers and patients are faced with decisions between alternative treatments in settings where the measurements on predictive covariates are presumably available before making a treatment decision and models which can differentiate the potential benefits for the specific patient seeking treatment are preferable. Note that in these settings, including a plethora of covariates may not be as desirable as including only a small set of variables that could reasonably be measured and considered by a practitioner. We thus aim to produce both conditional response models for this type of predictive problem, in addition to estimating the average treatment effect (ATE) which could be compared to other (potentially more inclusive) models. Further, the particular tools used in this approach (Dirichlet process priors, conditional models for regression) have been used in other settings and with large datasets, and so it is also reasonable to expect that extensions of this approach may be possible to include a much larger set of covariates.

Hence, for this thesis, we will consider a study with observational data with \( n \) subjects that contains only the basic requirements for demonstrating the validity of the method. Suppose that for each subject one observes the response \( Y \) and one wishes to provide evidence that treatment \( Z \) causes a difference in the response \( Y \) in the presence of a confounding variable \( X \). Here, the treatment choices \( Z \) are binary with values 0 or 1. Now consider that for each subject, there are random variables \((R_0, R_1)\) referred to as counterfactuals which are the possible responses for a subject dependent on whether the subject receives
treatment 0 or 1, respectively. We proceed with a standard assumption that the treatment assignment of one subject shall not have any impact on the outcomes of another subject. This assumption is typically referred to as the stable unit treatment value assumption, and in this model will be denoted as a stable unit treatment distribution assumption. One observes \( Y = R_0 \cdot (1 - Z) + R_1 \cdot Z \); that is, one measures a response from the density of \( R_0 \) if the subject receives treatment 0 and one measures a response from the density of \( R_1 \) if treatment 1 is given. Note that one only observes either a realization from \( R_0 \) or \( R_1 \) but not both. The expected value of the marginal difference between the treatments is then \( \mathbb{E}(R_1 - R_0) \) and represents the difference between treating the average subject with treatment 1 versus treatment 0, most commonly referred to as the average treatment effect (ATE). As noted in the literature for causal modelling, this value is not the same as \( \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) \) when \( Z \) is assigned in an unbalanced manner, and the ATE can be estimated through the proposed model. The assumption of strongly ignorable treatment assignment is an essential component of most causal models in the estimation of causal effects from observational data, see section 2.1.4 for details, and a similar assumption will be used in this model.

The Dirichlet process mixture regression (DPMR) model uses a Dirichlet process mixture to model the joint distribution of \((R_0, R_1, X, Z)\). The DPMR then uses conditional distributions to estimate regression functions. The Dirichlet process mixture is used as a prior on the family of distributions for the model parameters. The space of distributions which are sampled from the DPM is dense within the space of distributions. That is, we can assume that there exists a distribution from the DPM which is arbitrarily close to any other distribution. Therefore, methods which use DPM result in an extremely flexible fit and so are considered nonparametric Bayesian methods. There is also a great deal of flexibility in the specification of the structure of the model through particular choices of distributions for the observables, and relationships between parameters. This flexibility allows for the specification of a DPMR which integrates with the principles of causal modelling. The DPM and DPMR are discussed further in chapter 2.

This thesis, in Chapter 2, begins by providing a background of common assumptions and methods used in causal modelling and then reviews basic concepts of the Dirichlet process prior and its utilization in regression through Dirichlet Process Mixture Regression. In Chapter 3, we propose an approach to using causal modelling by linking the key assumptions of causal approaches through a fully Bayesian Dirichlet process mixture regression. In Chapter 4, we describe the details of two simulations that use simulated datasets under two conditions (one where the model assumptions are correct and another where the assumptions are violated) and implements the proposed model. The first simulation has only the minimum number of variables: an outcome, a treatment assignment with binary choices, and a single covariate which is confounded with both outcome and treatment assignment. The various possible outputs of the model are demonstrated. A second simulation is conducted with two confounding covariates; however, for the purposes of the analysis, it is presumed that one is not known and the results are examined in a similar manner to the first simulation. It is through this exploration that the clues of an unmeasured confounder are demonstrated. Hence, the flexibility of results produced by the models are explained, the nuances of what this new method can accommodate and the differences it can detect are examined.

Chapter 5 presents the results from fitting a simple model to a real-life data set from a non-randomized trial of homelessness where the outcomes are not normally distributed, and the treatment assignment is not well modelled by logistic regression. As in the previous chapter, the marginal average treatment
effect is compared to other standard methods, and conditional effects are plotted. This exercise also reveals the possibility of an unmeasured confounder. Chapter 6 discusses some of the implications that this approach can have on the utilization of research results on treatment decision making by health care providers, limitations of the model, and areas for further research. An appendix includes example code from R and WinBUGS that were used in these analyses and can be adapted for the analysis of simulated or actual research data.

A comment about notation, to simplify expressions throughout this thesis, functions of distributions (densities for continuous measures and probability mass functions for discrete distributions) are denoted by $f$ without subscripts. Thus the specific function being referred to is denoted by the argument of the function. For example, $f_{X|Y}(X \mid Y)$ will be written $f(X \mid Y)$, and $p_{Z|Y}(Z \mid Y)$ as $f(Z \mid Y)$. 
Chapter 2

Background

2.1 Causal Models: Previous theoretical principles

Several definitions and theorems have been proposed, and are generally deemed necessary, for causal modelling and some that are shared across many methods and thus appear essential to causal models. An abbreviated history and some critical models are reviewed here. Causal modelling concepts developed in several fields nearly simultaneously, for instance, the field of econometrics was struggling with the costs of implementing randomized trials to evaluate the effectiveness of social programs, and econometricians were aware of the difficulties in using non-random samples as controls for comparison. For instance, Lalonde [32] compared various methods used in econometrics using randomized trial results versus creating other comparison groups from observational data obtained elsewhere. Heckman et al. [20] expanded this research, comparing varied approaches to correct for ‘selection bias’ and developed a semiparametric approach to matching. They too begin with a counterfactual model. Donald Rubin [47] made early contributions beginning in 1974 that also developed an approach to causal modelling with observational data in the statistical literature.

From a statistical theory point of view, the first critical concept that is now a common element of most causal inference approaches is that of a counterfactual. This conceptualization posits that while we observe one result with one choice of treatment, there exists a potential outcome if a different treatment had been assigned. The reasoning continues that if we wish to make inferences about the difference between these treatments, then we must take into account that we have only one of these observations available. While random assignment is a potent assistant in the evaluation of causal effects, the theory outlined by Rubin allows us to develop an approach in the presence of non-random treatment assignment, in conjunction with covariates. These approaches have been argued on the basis of a balancing score generally, and are generally implemented using a propensity score. The theory supporting the use of balancing scores makes use of the assumption of a ‘strongly ignorable treatment assignment,’ along with assumptions of the constancy of effects and non-interference between units (sometimes referred to as the single unit treatment value assumption). The single unit treatment value assumption is described, and its implications are explored. Countervailing views are outlined regarding the need for counterfactuals in section 2.1.2 and regarding alternative descriptions of counterfactuals in section 2.1.6 however such models have not been widely adopted.
2.1.1 Counterfactuals

The concept of comparisons of potential outcomes has historical roots in philosophy, and amongst experimenters, however, a formal expression was not put forward until Neyman first introduced an approach to considering randomized experiments in 1923 [37]. The analysis makes use of a potential outcomes framework. In his description of a thought experiment to determine the average yield of a field from an agricultural experiment, he describes a system of urns containing balls to denote the yield on \( m \) plots (subdivisions of a field) with \( \nu \) different varieties of seeds that could be planted. He states “Let us take \( \nu \) urns, as many as the number of varieties to be compared, so that each variety is associated with exactly one urn. In the \( i \)th urn, let us put \( m \) balls (as many balls as plots of the field), with labels indicating the unknown potential yield of the \( i \)th variety on the respective plot, along with the label of the plot.”

He then clarifies that only one of these values can be observed, “Further suppose that our urns have the property that if one ball is taken from one of them, then balls having the same (plot) label disappear from all the other urns” [37]. That is, only one variety can be planted in a plot. As Imbens and Rubin 2015 [27], comment on page 25, “Throughout, the collection of potential outcomes... is considered a priori fixed but unknown.” This thought experiment formalizes the concept of counterfactuals, allowing us to imagine a table of possible observations of a yield from each plot. Similarly then in any observational study, we could infer a table of potential results with columns for each potential outcome and a row for each unit to be subject to study. For this table to be consistent regardless the order of subjects, we will need some additional assumptions that are described in the coming sections, while these were implicit in his presentation they were made more explicit with the developments in non-randomized experiments.

Fisher’s 1925 book “Statistical Methods for Research Workers” [14] is credited with introducing the concept that randomization is a requirement to ensure that the test of significance of an effect will be valid. His work also deals with experimentation on plots with field experiments in agricultural studies, where he compares different ways to assign varieties of plants or fertilizers to blocks of a field - systematically versus randomly. The combination of the counterfactual framework with randomization germinates several different experimental designs and new statistical techniques to the analysis of randomized controlled trials. However, it is quickly recognized that non-random assignment presents a difficulty which various authors from several disciplines (econometrics, public health, education, etc.) attempt to contend with.

Imbens and Rubin [27] outline a history of the development of counterfactual reasoning in observational studies, and cite important work by two economists, Tinbergen and Haavelmo. These economists made early forays into counterfactual reasoning but then seem to abandon this approach. In 1974, Rubin [47] described a model for both randomized and non-randomized experiments that uses reasoning about the difference between counterfactuals, recognizing that the bias is minimized in randomized trials, but may be balanced by matching. This reasoning forms the groundwork of his later work with Rosenbaum [48] where they connect the principles of a balancing score to this idea.

Heckman et al. [21], who contributed to developments in causal modelling approaches in the economic literature of labour market programs, posit a broader history for causal models development. He points out that individuals in various fields have developed approaches to causal modelling that use counterfactual reasoning and this development is described as “differentially credited” to various authors, including: Fisher [13], Neyman [37], Roy [44], Quandt [41] or Rubin [47]. For instance, in 1951 Roy [44] describes a thought experiment in economics where the actors in the economy can choose to be either hunters or fishers, and their income depends on this choice. He describes various possibilities for the
differing incomes in some imaginary currency based on the skill of the worker in the chosen profession and the impacts on the economy in terms of pricing of the goods (fish and rabbits). In this example, however, the matter of treatment assignment, which corresponds best to the choice of profession, is neither random nor haphazard but assumed to follow some principles based on individuals having a sense of their competence at the skill required for their profession. It is also clear from the work in Heckman, that they envision that participation in programs may have direct effects on those participating in programs and indirect effects on individuals who did not participate in a social program but who live in a community where such a program is offered and may be impacted positively or negatively by its presence. Such an indirect effect would be a violation of the non-interference assumption which will be outlined later.

2.1.2 Causal models without counterfactuals

The use of counterfactuals, while almost always underlying causal modelling approaches, is not ubiquitous. For instance, Dawid [7, 8] has proposed a decision-theoretic approach which does not require counterfactuals and instead proposes expressing a full joint model for baseline covariates, actions taken (interventions) and outcomes. His work has developed in the setting of treatment strategies that evolve over time (for instance, initiation of HIV antiretroviral treatment, or adjustment of medication in response to blood levels) and the question to be answered is often regarding the causal effects of various possible regimes (treatment strategies).

Specifically, he goes on to write a more philosophical treatise on the use of counterfactuals in 2000 [8] where he addresses the use of counterfactuals in experimental research (while making some connections to observational study). His argument is built by beginning with the creation of a counterfactual model including a term for correlation between the counterfactuals which he refers to as a “metaphysical model” (since this can never be observed) and then comparing this to a purely “physical model” of observed data. He builds up towards a contradiction by suggesting that one must always posit a correlation term between the counterfactuals, and while this can never be measured (since we can never observe both) our dependency on it creates contradictions in estimating approaches. He argues that each common causal approach in current use induces an assumption at the level of the correlation through its other assumptions and modelling tasks even if we do not always appreciate how this correlation is induced. He further asserts that some assumptions (such as treatment-unit additivity) seem more likely to be erroneous under certain situations. For instance, he suggests that when we know covariates about our data, we have additional data that might relate to the correlation in the outcome. He then proposes that a decision-theoretic approach can address this problem. This proposal faced intense opposition, and several countervailing views were written in response to Dawid’s arguments against the use of counterfactuals. Since he restricted his arguments to experimental situations; the implications for observational data are less clear from his article.

2.1.3 Balancing scores and Propensity scores

The second concept, a balancing score, complements a counterfactual model and is introduced as an intermediary to the propensity score. It is used to demonstrate and prove the unbiased nature of a family of estimators that can be implemented with observational data. A balancing score is defined as a function of covariates, which when conditioned on, the distribution of \( X \) is independent of treatment
assignment (that is the distribution of \( X \) is identical for treatment 0 and treatment 1 at identical values of the balancing score). This property can be written as:

\[
X \perp Z \mid b(X) \\
f\{X, Z \mid b(X)\} = f\{X \mid Z, b(X)\}f\{Z \mid b(X)\} = f\{X \mid b(X)\}f\{Z \mid b(X)\}
\]

Rosenbaum and Rubin’s 1983 paper [43] advances several critical theorems: first that the propensity score, \( e(X) = pr(Z = 1 \mid X) \), is the ‘coarsest’ balancing score and \( X \) itself the ‘finest’, and second that a function \( b(X) \) is a balancing score if and only if there exists a function, say \( g \), of \( b(X) \) that equals the propensity score, also denoted by there exists a \( g \) such that \( g\{b(X)\} = e(X) \).

2.1.4 Strongly ignorable and positivity assumptions

The third concept often required is an assumption regarding conditional independence of the counterfactuals and the treatment assignment. This assumption allows the balancing score, and hence propensity score, to be used to create unbiased estimators. The theorems in [43] rely on an assumption about the covariates available to the analysis; this assumption requires us to assert that the counterfactual responses (\( R_0/R_1 \)) are conditionally independent of the treatment assignment (\( Z \)) given the measured covariates and that the probability of treatment assignment to any treatment must be non-zero at all values of the covariates. The first assumption regarding conditional independence has been described by some authors as the strong ignorability assumption and by others as the condition of no unmeasured confounders ([10]), and appears in many authors works on causal modelling. This assumption can be written in these equivalent ways, or be represented in this directed acyclic graph, in figure 2.1:

\[
(R_0/R_1) \perp Z \mid X \\
f((R_0/R_1), Z \mid X) = f(R_0/R_1 \mid Z, X)f(Z \mid X) = f(R_0/R_1 \mid X)f(Z \mid X)
\]

Figure 2.1: A schematic diagram of a causal model under strong ignorability.

The further assumption is proposed that there must be non-zero treatment assignment probability in the range of covariates under study. This assumption is referred to as the positivity assumption.

The other essential theorem that was proved in Rosenbaum and Rubin’s paper is that the expected difference between two treatments conditioned on a balancing score will be an unbiased estimate of the treatment difference at that value of the balancing score, so long as the balancing score is based on covariates for which treatment assignment is strongly ignorable.

\[
E\{R_1 \mid b(X), Z = 1\} - E\{R_0 \mid b(X), Z = 0\} = E\{R_1 - R_0 \mid b(X)\}
\]
Hence, by taking expectation over \( b(x) \) we find:

\[
E_{b(x)}[E\{R_1 \mid b(X), Z = 1\} - E\{R_0 \mid b(X), Z = 0\}] = E_{b(x)}[E\{R_1 - R_0 \mid b(X)\}]
\]

\[
= E(R_1 - R_0)
\]

Heckman et al. [20] argue that Rosenbaum & Rubin’s use of a known propensity score ignores the impact of estimating the propensity score from data, and they argue this relies on an assumption about the counterfactual conditional mean, namely that \( B(f(X)) = E[Y_0 \mid f(X), Z = 1] - E[Y_0 \mid f(X), Z = 0] = 0 \). This expression serves to more directly convey that a function of the covariates, \( f(X) \), is used to model the propensity score and itself may not be a true balancing score, the difference between the expectations of this estimated propensity for those assigned to one treatment versus the other measures the balance achieved through this estimate at various levels of \( X \). This assumption regarding \( B(f(X)) \) can replace the usual strong ignorability assumption; they go on to argue that this condition is testable and in their particular problem, is erroneous.

### 2.1.5 Single Unit Treatment Value Assumption - SUTVA

David Cox in his 1958 book[6] on the design of experiments outlined a series of assumptions that he considered necessary for experimentation. The first was the concept of additivity; that each unit’s outcome was the sum of an effect based on the unit and an effect based on the treatment. He goes on to note that these effects are “to be unaffected by the particular assignment of treatments to the other units.” He posits these three key results of this assumption (or three ways this assumption could be violated): 1) additivity of effects (although allowing for the possibility that some effects are multiplicative and hence additive on a log scale), 2) constancy of effects, and 3) non-interference between units. Rubin[49] describes these principles again in a 1980 commentary and refers to them as the stable unit-treatment value assumption when he proposes that from an experiment one could envision a table of outcomes \( Y_{ij} \) which represents “the response of the \( i \)th unit (\( i = 1, \ldots, 2n \)) if exposed to treatment \( j \), (\( j = 1, 2 \)).” Here, he again emphasized that in this set up one assumes that by assigning treatment to one unit, it has no impact on the outcome of another unit. In this setup he envisions a balanced experiment with paired comparisons and \( 2n \) units being exposed to 2 different treatments. One could imagine this being violated in situations where there is a scarcity of treatment providers, where later treatments involve a tired provider who offers a treatment which is less effective and poorer in quality, or the dose is decreased to treat more individuals.

Imbens and Rubin[27] also describe the single unit treatment value assumption as it relates to non-randomized experiments in their book on causal analysis, where they describe it as: “The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.” They add the concept of “no hidden variations of treatments,” giving the example that a mixture of new appropriate strengths/dose of medications and older medications that no longer contain an effective dose from which treatments were selected would be an example of a violation of this assumption. This change in dose over time would violate Cox’s constancy of effects assumption. So this assumption is shared with both experimental and observational approaches and is also assumed in our approach.
2.1.6 Deterministic versus stochastic counterfactuals

Returning to Neyman’s urn metaphor/thought experiment, one could also imagine rather than selecting a ball upon which the potential yield/outcome is written instead one draws from the urn a random variable generator. It is from this random variable that a specific outcome will be realized when the outcome is measured on this unit. This framework is similar to the use of probability densities in quantum mechanics; where it is assumed that the location, velocity, and momentum of quantum particles exist as a probability density function until operated upon (for example, through measurement) by an outside force, at which point the particle ‘snaps’ into a specific state.

Sander Greenland [17] use the concept of a stochastic counterfactual, first introducing it conceptually in his 1987 paper. This paper explores the use of odds ratios and demonstrates how in the face of a mixture of two populations the odds ratio may be misleading. In this paper, he clearly describes imagining outcomes per unit as arising from probabilistic rather than deterministic processes. This conceptualization is formalized in a 1989 paper [42] that Greenland co-authored with James Robins, in a survival framework where they imagine counterfactual survival functions over time which express the risk of an event at time t, under each possible treatment. They describe this as “a stochastic version of Rubin’s (1978) causal model”. This approach has yielded advances in considering a stochastic sufficient cause framework that can detect the presence of joint causes in a stochastic counterfactual model[55]. These are models where one might envision different pathways that may lead to the development of a response, with or without shared exposures (for instance specific genetic factors with particular environmental exposures). By basing this model on stochastic counterfactuals and cluster-based models, there is the possibility of extending the models to capture this sufficient cause framework in later developments.

This conceptualization of stochastic counterfactuals seems particularly apt in the analysis of observational data on health care outcomes, as there are likely large numbers of factors with influence on outcomes that are driven by biological random processes. This line of reasoning also opens a further parallel to be considered, which is embedded in the quantum mechanical view of physics - the concept that the operation of measurement itself perturbs the system. It seems both reasonable and likely that a similar process could occur in some areas of medicine to a greater or lesser degree. This is typically ignored in medicine; however, we suspect that a similar process factors significantly in mental health and addictions research, where the questions that are used to inquire about a person’s mental state or behaviours induces a state of mind to answer these questions which can, in turn, impact a person’s mental state. While this may be an important factor in the measurement of responses, this is not incorporated into our current model but remains instead as an area for potential future development. By designing our current model using stochastic counterfactuals, it allows for it to be more easily adapted in the future to deal with this impact of measurement. Specifically, the measurement could be treated as an operator on the density function which may vary by measurement technique, as is the practice in physics.

Perhaps a more subtle and less obvious implication of this line of reasoning asks us to consider the situation where we might identify all covariates that influence outcomes. That is, in the idealized situation, being aware of and measuring all known confounders and all known direct covariates (that is factors which influence the outcome but do not influence treatment assignment) within a naturalistic study, would there still be a random element to responses within an individual. We proceed imagining that there will still be randomness even after accounting for all unmeasured direct covariates and these are then separately modelled when we generate a simulated dataset, and as such, they differ by assigned
condition, unlike the direct covariates, which if unmeasured are still assumed to influence outcomes identically for both counterfactuals. Given that the counterfactuals can never be measured, this assumption can never truly be tested and rests in the philosophical perspective of the statistician who analyses data. Other authors may contest this point and proceed with a different approach to simulating data; we do not believe that this has a substantial impact on the findings presented.

2.2 Causal Models: Previous Applied Methods

In order to ease comparisons between different approaches used in the literature, the notation from previously reported research is expressed in the notation used in this thesis rather than the notation used in the papers themselves unless to do so would detract from additional distinctions that their alternate notation would clarify. Specifically, counterfactuals or potential responses are denoted by $R_1$ and $R_0$, the treatment assignment by $Z$ and confounders or covariates with $X$.

2.2.1 Rubin causal model

Peter Austin provided a review of propensity score methods in 2011 outlining many of the pragmatic issues in implementing causal modelling with these methods [2]. He clarifies the difference between two estimates: the average treatment effect (ATE), $E(R_1 - R_0)$, and the average treatment effect for the treated (ATT), $E(R_1 - R_0 | Z = 1)$, crediting Imbens [26] for this distinction. The first estimate, ATE, represents the average effect of switching from treating the entire population from treatment 0 to treatment 1, whereas the second estimate, ATT, represents the average treatment benefit that individuals who accepted the treatment are receiving over the expected effect if they had not. He gives examples where one or the other may be the more important estimand and points out that this is a scientific question, to determine which is more relevant. For instance, if one is concerned with a treatment where there may be many barriers to offering the treatment to a broader set of people, the ATT may be the most relevant, whereas, a public health intervention that could easily be disseminated to a larger population may warrant ATE as the more appropriate estimator. The differing propensity score methods may be more or less useful in estimating each.

Austin outlines several key features: the existence of four standard approaches which use the propensity score to redress confounding, the two-step nature of estimating the propensity score and then creating a treatment effects model adjusted by this propensity score, and practical tasks involved at each step in estimation. The four standard approaches to propensity score use are matching, stratifying, inverse probability of treatment weighting (IPTW) and adjusting the treatment model by inclusion of the score as a covariate. Matching typically creates estimates of the ATT, as one generally creates a sample that retains the overall covariate distribution of the treated sample. Matched pairs can then be compared directly using methods similar to RCTs; however, adjustments are needed to the estimates of standard error and confidence intervals to reflect the lack of independence between treatment and matched controls. He points to simulation studies to argue that approaches that adjust the standard errors accounting for the dependence are more accurate. He also describes additional approaches to improve estimates that include matching on other prognostic factors in addition to the propensity score, or further covariate adjustments. Several practical decisions need to made regarding matching: how close a match must be, when to leave observations unmatched and thus discarded from the analysis, whether to match with or without replacement; and the appropriate model implications of each of these decisions. A distinction
between greedy (matches are done sequentially in random order with the first control match found being kept and thus not available to another unit, even if it is a better match for another treated unit) versus optimal matching (a process is used to find the best set of matches over the entire dataset) is made.

To decide whether a pair constitutes a match, one can use a nearest neighbour approach, or a nearest neighbour within a set distance (referred to as a ‘caliper distance’). Cases may remain unmatched, thus excluded from the analysis, if no match can be identified within this threshold distance. Much work has been done on caliper distance, and Austin cites Rosenbaum and Rubin, and Cochran and Rubin on the use of the logit of propensity score as an important method to constructing the most useful matches. Finally, in some situations one may have access to a large number of potential matches, and while 1:1 matching remains most common, a higher ratio of control to treated observations can also be used, including methods that use a variable number of matches as opposed to a fixed ratio.

Austin continues to describe the other standard methods, noting that most stratification methods typically use ‘5 equal sized groups’, but a larger number of strata results in less bias with declining improvements. This method relies on near similar propensity within each subgroup. He notes that these methods can account for approximately 90% of the bias and cites Cochran 1968. The ATE can then be estimated by using a within-group comparison of effect, summarized by weighting by the group size as a proportion of the total sample. The ATT can be estimated by weighting by the size of the treated within each stratum instead. Variance estimates are calculated using pooled variances from each stratum. He also notes that additional methods can be used to correct for remaining differences and these can be accomplished within each stratum with regression methods and cites the work of Imbens (2004) and Lunceford and Davidian (2004).

IPTW uses survey methods to account for the lack of balance by weighting observations of the ith observation with the propensity \( e_i = P(Z_i = 1 \mid X_i) \) with weights given by \( w_i = Z_i \frac{1}{e_i} + (1-Z_i) \frac{1}{1-e_i} \), that is, the inverse of probability of this particular observation having been selected for this treatment condition \( Z_i \). This weighting allows one to estimate the ATE using various survey methods, and the ATT can also be estimated through the use of alternate weights, \( w_i = Z_i + (1-Z_i) \frac{e_i}{1-e_i} \). Problems emerge from the instability of the estimated weights at the extremes of the propensity score, namely, the very unlikely or the highly probable treatment group assignment. The variance estimates need to be carefully constructed using estimates which also account for these weights, as in other complex survey methods.

Finally, regression methods that include terms for treatment group and propensity score as predictors are referred to as covariate adjustment models, and these have also been studied. Austin concludes that several studies have demonstrated that matching methods outperform stratification and covariate adjustment. He also reports that IPTW and matching have closer results, with some suggestions that in some situations matching may provide more bias correction. He goes on to state that the IPTW and covariate adjustment may be more sensitive to correct specification and estimation of the propensity score. Austin reviews several methods of checking the estimated propensity score with ‘balance diagnostics.’

### 2.2.2 Previous Non-Parametric Models

Ho, Imai, King and Stuart describe matching of observations as a pre-processing procedure and as a nonparametric technique that creates “less model-dependent causal inferences,” [24]. While earlier we discussed methods that use matching of treated units with controls based on propensity score, Ho et al. discuss and propose a step-wise approach to matching that uses many approaches that have appeared in
the literature. They are not using a propensity score or balance score reduction but instead the entire covariate $X$ to determine matches (and with potentially multiple matches) to then use this in whatever the scientist feels is the most appropriate, presumably parametric, analysis with which to answer the research question best. They also assume “the absence of ‘omitted variable bias’” which they explain as the term in the political science literature for the ignorability assumption in statistics.

Neugebauer and van der Laan \cite{Neugebauer2001} propose a nonparametric causal approach applied to longitudinal data in which there is a continuous exposure by extending the marginal structural model (as originally forwarded by Robins in 1998). These models have additional assumptions due to their applications to longitudinal data, but similar to the approach in this thesis share an assumption of the existence of counterfactuals. An additional assumption of sequential randomization allows for a factorization of the likelihood into two components, one of which relates to the treatment assignment mechanism (over time in this situation) and the other to the ‘full data process’ which parallels the treatment response model/propensity score conditional independence in point treatment approaches. When this method is applied to data arising from a single time point rather than collected longitudinally, this reduces to a propensity score method with inverse probability weighting. The nonparametric approach was only applied to the full data process component.

Ernest and Buhlmann \cite{Ernest2002} apply a nonparametric approach to marginal integration for causal modelling. This approach builds on the structural equation modelling approach and builds on the directed acyclic graph (DAG) approach developed by Pearl. Their model uses Pearl’s approach to modelling which distinguishes variables that can be manipulated by an experimenter; this invokes a ‘do-operator’ which corresponds to an active decision by a scientist (or policymaker) to intervene on a system through a specific co-variate and then measure an effect elsewhere on the graph. This ‘do-operator’ has an algebra, and determining which variables are required for adjustment, uses both conditional probability expressions and requires a graph that identifies the proposed/identified relations between all intermediate variables and the predictor and outcome variables under study. The conditional probability assumptions are ‘read’ from the directed acyclic graph. In Ernest and Buhlmann’s work an ignorability assumption is included (‘all relevant variables are observed’) and further conditional independence relationships are expressed as embedded through the structure of the DAG. They approach the problem of causal modelling in situations both where the ‘true’ DAG is known and where it is not known, by using a nonparametric regression of the response $Y$ based on the measured covariates $X$ (which includes all predictors, confounders, and the treatment variable), and a subset of $X = X_S$ to be further adjusted for. These additional variables, which satisfy the “backdoor criteria,” are selected from the DAG, and then the treatment effect is estimated by marginal integration over this $X_S$ subset. When the DAG is unknown, they propose that either a parametric or nonparametric process can be used to estimate the DAG.

Other extensions of the marginal integration approach have used nonparametric approaches to longitudinal and time series data. It is important to note that there exists a somewhat distinct field of causal modelling that has developed out of time series data. It emerged from the econometric analysis of markets and attempts to model how policy changes, critical decisions or interventions at a higher level may affect trends in stock prices or other trading outcomes. These causal modelling techniques are often referred to as “Granger” type causal models and the development of changes in variables over time are an essential aspect of these causal models. Here the interventions are also posited to occur at a point in time (or over time in different jurisdictions) and this relationship with time is a critical aspect
of the theory and models. While these “Granger” causal models are not reviewed or treated extensively in this thesis, it is important to recognize that nonparametric methods have been considered in this setting. The aforementioned work in marginal structural models was further extended by Li et al. to develop a nonparametric causal inference framework for time series data. Given that the assumptions of “Granger” causal models are somewhat distinct from data that is measured at only one time-point, we did not consider their methods further. Of note, the fully nonparametric approach they implemented, they claim remains unaffected by the ‘curse of dimensionality so long as smoothness conditions hold.’

Athey and Imbens propose a modification of the classification and regression trees (CART) models for application to causal modelling. They focus on creating conditional treatment effects so that clinicians might personalize treatment recommendations within identified subgroups. They propose an ‘honest’ approach by partitioning data and using a portion to decide on covariate selection (selecting which variables appear in each tree) and a separate partition for estimating the treatment effects and refer to this as a causal forest approach. The approach is developed theoretically in application to randomized settings, then is said to be able to be modified by propensity score weighting. In Wager and Athley, the authors develop a method to construct confidence intervals for the treatment effect. They note that they can use the ignorability assumption (they refer to it as ‘unconfoundedness’) to achieve consistency of estimates “without needing to explicitly estimate the propensity $e(x)$.”

Kennedy et al. also use nonparametric methods in combination with a doubly robust estimator to address the estimation of a continuous-valued treatment. They have a two-stage model; one stage creates a ‘pseudo-outcome’ consistent with the doubly-robust principles they wish to adhere to and then the second stage predicts this outcome based on the treatment using a nonparametric approach, specifically, a kernel density estimator. While they outline three assumptions common in the literature to allow for identification (consistency, positivity and ignorability), they state “even if we are not willing to rely on assumptions [of consistency] and [ignorability] it may often still be of interest to estimate [the effect curve] as an adjusted measure of association, defined purely in terms of observed data.” Alternate assumptions are then proposed in their demonstration of consistency and asymptotic normality, which rely on having the true function of the mean outcome given covariates or the true function of the conditional treatment density given covariates.

Nonparametric approaches have primarily been developed to address difficulties in the common approaches to causal modelling in more sophisticated settings then the early causal approaches had been developed for (applications in longitudinal data or time series data, continuous treatment assignment, misspecification of the propensity score) and often attempt to advance and build on doubly robust principles. We instead seek to develop a more explicit basic causal model that could later be extended. Further, it seems that applications of nonparametric approaches to causal modelling struggle with the ignorability assumption, and this is commonly dealt with by either applying the nonparametric strategies to either the treatment assignment model or the outcome model, or both separately, or adapting/replacing the ignorability assumption to demonstrate other conditions to support the nonparametric approach.

### 2.2.3 Previous Bayesian Causal Models

Other Bayesian approaches to causal modelling have been proposed previously. Rubin first introduced the ideas of Bayesian analysis for causal modelling in 1978 and described the interactions between unit sampling, treatment assignment, and data recording, positing that a Bayesian method must model each of these processes if they cannot be assumed to be ignorable. In 2009, McCandless proposed a
Bayesian method that jointly models propensity score and outcome, by using the model for the propensity score to generate latent class memberships from the propensity scores. Hill proposed using Bayesian additive regression trees (BART) to model the response surface in 2011 [23]. She notes advantages of the approach include the simplicity of method, capacity to include a large number of covariates and flexibility in fitting data; she argues that the flexibility of fit justifies its capacity to create unbiased estimates [22]. In this paper, she described the conditional average treatment effect for the treated (CATT), and the sample average treatment effect for the treated (SATT), making the distinction that the sample analyzed may not be a random sample from the larger population that one wishes to make estimates about.

Hoshino [25] proposed a joint model composed of three submodels \( p(R_1, R_0 \mid \nu)p(x \mid R_1, R_0, \nu)p(Z \mid R_1, R_0, X, \nu) \), where there is inclusion of regressors \( \nu \) which are different than the covariates but also considered important by the researcher. The first two submodels, that is \( p(R_1, R_0 \mid \nu) \) and \( p(x \mid R_1, R_0, \nu) \), are fit using a probit stick-breaking process mixture, which is an extension of a Dirichlet process. Their model assumes a somewhat different conditional independence assumption; it assumes that just one of the counterfactuals \( R_1 \) is conditionally independent of the treatment assignment \( Z \) given the other counterfactual \( R_1 \), the confounders \( X \) and regressors \( \nu \). Specifically:

\[
R_0 \perp Z \mid (R_1, X, \nu)
\]

They argue that this weakens the usual strongly ignorable assumption and contains models that are both parametrically and nonparametrically increasing the flexibility of the fit.

Zigler [57] introduces a Bayesian approach with a joint model including both propensity score and outcome modelled in a single step rather than the usual two-stage process typically used in a frequentist approach, but they note model feedback limits its capacity to create unbiased estimates. In our method, the propensity score per se is not modelled, nor is a balancing score introduced which cannot factor within the joint model specified, thus sidestepping some of the problems implicit in their proposed approach.

More recently, Jason Roy, Michael Daniels and colleagues have proposed causal modelling approaches using a Dirichlet process prior. In a paper introducing a framework for causal inference of mediation, they apply a Bayesian nonparametric approach to data obtained from a randomized control trial [31]. While the treatment was randomly assigned, they apply counterfactual reasoning to the mediating variable, thus have a different framework and set of assumptions specific to this situation than is considered in this thesis. In Roy, Lum and Daniel’s 2017 article [45] they focus their attention on causal models based on marginal structural models, while these models have typically been used for situations that include both time-varying covariates and longitudinal or survival outcomes, in their article they focus on a single treatment, baseline covariates and a continuous or survival outcome measured at a single time point. They express outcomes in the form \( E(Y_z \mid X; \Phi) = h_0(x; \Phi_0) + h_1(z, x; \Phi_1) \) and use a dependent Dirichlet process to model the outcome given confounders \( (h_1) \) and a Gaussian process for the mean model \( (h_0) \). They describe three required assumptions (consistency, positivity and strong ignorability). Finally, in March of 2018, Roy et al., in an electronic publication ahead of print release, presented an enriched Dirichlet process approach to causal inference and then explore how this works for random covariates [46]. While there are some similarities with the approach they propose, they claim a strongly ignorable assumption, but it is not clear from the article where this is used and if a local ignorability assumption may in fact be needed for their computational steps. Further they model the outcome conditional on all covariates (within which treatment assignment is included) and a second Dirichlet process prior for the
parameters of the covariates. They note that they model all covariates as independent. Their inclusion of many variables can provide a framework for the extension of the model presented in this thesis to higher dimension covariate spaces.

2.3 Non-parametric models: Using the Dirichlet process mixture as a regression model

2.3.1 The Dirichlet process prior

The capability of performing Bayesian nonparametric modelling was greatly advanced by Ferguson’s development of the Dirichlet process in 1973 [12]. He extended the Dirichlet distribution to a process by considering the Dirichlet distribution as arising from a partition of the sample space $\mathcal{X}$. He generates this partition by using a $\sigma$-field of subsets $A$ of $\mathcal{X}$. By starting with an arbitrary finite collection of measurable sets $A_1 \cdots A_m \in A$ along with a finite non-null measure $\alpha$ on this space, he demonstrates how a random probability measure $P$ can be created. First he creates a partition $B_1 \cdots B_k, k = 2^m$ using intersections of the $A_i$ sets and their complements $A_i^c, B_{\nu_1} \cdots B_{\nu_m} = \cap_{j=1}^m A_{\nu_j}^c$ where $\nu_j = \{c, 1\}$. By creating a partition, he can invoke the Kolmogorov consistency conditions for the distributions of $P(B_j)$ (the partitions from which each arbitrary set $A_i$ are constructed from), and extend this to ensure that the probabilities $P(A_i)$ also exist and are appropriately defined (have a sigma-additivity). Furthermore, he posits an underlying continuous measure on the sample space $\mathcal{X}$ which generates this measure on the partitions. He defines a Dirichlet process $P$ with parameter $\alpha$, if for every $k = 1, 2, \cdots$ and arbitrary set $A_1 \cdots A_m$ the distribution of the $(P(A_1), \cdots P(A_m))$ is Dirichlet with parameter $(\alpha(A_1) \cdots \alpha(A_m))$. He was able to demonstrate that draws from this process are almost surely discrete, and he also demonstrated that the posterior distribution of $P$ given a set of observations $X_1, \cdots, X_n$ from $P$ is also a Dirichlet process. These random draws from the process can be conceptualized as a draw of a random distribution with an infinite number of discrete jumps, a discrete probability measure.

Blackwell demonstrated an alternate way to prove that realizations of the Dirichlet process are discrete distributions with probability 1. His proof did not rely on a gamma process, which Ferguson had used in his initial set of proofs [3]. This result helped expand how early researchers understood the properties of the Dirichlet process. Blackwell and MacQueen in the same issue of the Annals of Statistics put forward a procedure to sample points from the Dirichlet process related to a Polya urn scheme [4]. They describe setting up an urn with $\alpha(x)$ balls of colour $x$ in an urn, where $x$ is an observation from $X$. They define a Polya sequence $X_n$ with parameter $\alpha$, where each $X_i$ represents a draw with replacement from the urn where a second ball of the same colour is added back to the urn after the draw. They then draw parallels between this set-up and the Dirichlet process demonstrating that they converge to the $P$ described by Ferguson. They also simplified some of the notation and definition of the Dirichlet process by narrowing down on some of the essential components; for instance, defining it in terms of a finite partition of the sample space $\mathcal{X}$ rather than as an arbitrary collection of sets.

2.3.2 Stick-breaking implementation of the Dirichlet prior

Sethuraman in 1994 proposed an alternate construction to the Dirichlet process that has been described as the stick-breaking prior [51]. In his construction, he was able to develop an observation from the
process that could be created step-wise, this allowed the creation of a truncated version of the Dirichlet process, and led to easier implementation in some situations.

The Dirichlet process samples a random discrete distribution $G$ on $\Omega$ (the support for the parameters of the model which are included in the Dirichlet process), and this is parametrized by two components, a distribution $G_0$ which can be thought of as the ‘center’ of this process, and $\alpha$ which acts like a precision parameter ([11]). To elucidate, let us introduce an example, and one of the algorithms used to generate such a $G$ referred to as the stick-breaking construction ([51]). Here we define $G = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$ with point mass at $\theta_j$ and this would represent one draw from the Dirichlet process. The $\theta_j$‘s will have been sampled identically and independently from $G_0$ and the $p_j$ are independently constructed iteratively by ‘breaking off’ a new probability for the $j$th group from the remaining probability not yet accounted for by the previous $(j-1)$ terms. The proportion of the remaining probability used is sampled as $V_j$ using the tuning parameter $\alpha$, specifically: $V_j \sim \text{Beta}(1, \alpha)$; $p_1 = V_1$, leaving $(1-V_1)$ remaining; $p_2 = V_2(1-V_1)$, leaving $(1-V_1)(1-V_2)$ remaining; \cdots; $p_j = V_j \prod_{k=1}^{j-1}(1-V_k)$; \cdots. This process generates a cumulative distribution function of $G$ as a step function with locations $\theta_j$ sampled from $G_0$ and corresponding step heights of $p_j$. Hence as $\alpha$ increases, smaller probabilities are broken off, and the distribution approaches $G_0$, whereas, as $\alpha$ decreases a few points at locations $\theta_j$ predominate with greater probabilities at these points.

### 2.3.3 Dirichlet process in joint probability models

Suppose we are interested in modelling the joint distribution of some multivariate random variable $W$ with support $\mathcal{X}$, which in our example will contain the confounders, counterfactuals and treatment assignments, by this infinite mixture of clusters. The joint distribution of $W$, $f(W)$, is parameterized by several terms, some which will be constant across all clusters (for instance, one can choose a constant within cluster precision) and others that will differ between clusters (for example the cluster means). When we observe a specific draw from the Dirichlet process, $G$, which models the cluster dependent parameters, it will be defined by the sets $\Theta_{1:\infty} = \{\theta_1, \theta_2, \ldots\}$ (the location of the parameters) and $P_{1:\infty} = \{p_1, p_2, \ldots\}$ (the cluster membership probability). Let us then imagine that we set up the following nested distribution: a distribution $G$ representing parameters of the data model is sampled from the Dirichlet process, then a specific set of parameters $\theta^o$ is sampled from this $G$, and finally the data is modelled as a density with this set of parameters from $G$ and potentially other parameters $\nu$:

\[
[W = w \mid (\theta^o, \nu)] \sim f(W \mid \theta^o, \nu) \tag{2.1}
\]

\[
[\theta^o \mid G(\Theta_{1:\infty}, P_{1:\infty})] \sim G
\]

\[
[G \mid (G_0, \alpha)] \sim DP(G_0, \alpha)
\]

then the posterior will be a convolution of the $G$ from the Dirichlet process with the distribution $f$. One can thus express an observation from the posterior as it relates to a single observation $w$ (either in the set of observed data $w_i$ or a future $w_{n+1}$ to be predicted) as:

\[
\int f\{w \mid (\theta^o, \nu)\}dG(\theta^o) = \sum_{j=1}^{\infty} p_j f(\theta_j, \nu) \tag{2.1}
\]
Here one can conceptualize the $f(w \mid \theta_j, \nu)$ at a corresponding $j$ as a cluster, or as a local expert, of the joint distribution in the neighbourhood of this local distribution. Here the parameter within $f$ of $\nu$ is included for situations where a parameter is needed for the distribution of $W$ which does not differ across groups. For instance, one might wish to create a mixture of normals with different means but identical precision (variance) across groups. This sum in 2.1 then represents a mixture of the local densities of $f$ at specific $\theta$ values.

Alternatively, one can think of each observation $w_i$, as having latent parameters $(\theta^0_i, \nu)$ which take on a specific value which are sampled from the DPM, this invokes an additional random variable $S$ which represents the membership of this observation, to a specific group, that is $S_i = j$ when $\theta^0_i = \theta_j$. This leads to simpler notation when conditioning on group membership $S$ of the density and group membership probabilities:

$$f(W \mid S = j, G) = f(W \mid S = j, \Theta_{1: \infty}, P_{1: \infty})$$

$$= f(W \mid S = j, \theta_j) = f(W \mid \theta_j), \text{ and}$$

$$f(S = j \mid G) = pr(S = j \mid \Theta_{1: \infty}, P_{1: \infty}) = p_j$$

This allows us to write the posterior distribution more succinctly in a representation independent of the sampling algorithm for the Dirichlet process. It also becomes clear that the distribution of the random variable $S$ is completely defined by $P_{1: \infty}$. If we represent all the necessary parameters in the joint model as $\Phi = \{\Theta_{1: \infty}, P_{1: \infty}, \nu, \alpha, G_0\}$ we can think of the joint distribution as the sum across possible choices of $S$.

$$f(W \mid G, \nu) = \sum_j pr(S = j \mid P_{1: \infty}) f(W \mid \theta_j, \nu)$$

$$f(W \mid \Phi) = \sum_j f(S = j \mid \Phi) f(W \mid \Phi, S = j)$$

### 2.3.4 Regression framework

This conceptualization can now easily be extended to a regression framework including representing non-linear relationships. If $W$ is comprised of two components $(W_1, W_2)$ with $W_1$ representing outcomes of interest, and $W_2$ representing a set of predictors, the joint distribution can be modeled by cascading conditional distributions within clusters, as has been demonstrated by other authors (52, 19, 29, 39, 35). If we write the $\Theta_{1: \infty}, \nu$, and $G_0$ as containing a subset of the parameters that is pertinent to $W_2$, we can reference this subset as $\Phi^{W_2}$. One can then model the local relationship as $f(W \mid \Phi, S = j) = f(W_1 \mid W_2, \Phi, S = j)f(W_2 \mid \Phi^{W_2}, S = j)$, dropping the parameters which do not contribute to the conditional probability term. In regression, we are interested then in the relationship $f(W_1 \mid W_2, \Phi)$ or $E(W_1 \mid W_2, \Phi)$ which can be expressed, using the same concept of summing across all possible $S$ groups,
and applying Bayes’ rule to the first term in the sum, as:

\[
f(W_1 \mid W_2, \Phi) = \sum_j f(S = j \mid W_2, \Phi)f(W_1 \mid W_2, S = j)
\]

\[
= \sum_j pr(S = j \mid W_2, \Phi)f(W_1 \mid W_2, \theta_j)
\] (2.2)

\[
pr(S = j \mid W_2, \Phi) = \frac{f(W_2 \mid S = j, \Phi)f(S = j \mid \Phi)}{\sum_k f(W_2 \mid S = k, \Phi)f(S = k \mid \Phi)}
\]

\[
= \frac{\sum_k f(W_2 \mid \theta_{W_2}^j)p_j}{p_j(W_2)}
\] (2.3)

\[
E(W_1 \mid W_2, \Phi) = \sum_j p_j(W_2)E(W_1 \mid W_2, \theta_j)
\] (2.4)

Thus the conditional distribution of \( W_1 \mid (W_2 = w_2) \) is dependent on a vector of values \( pr(S = j \mid W_2 = w_2, \Phi) = p_j(w_2) \) which does not contain \( W_1 \), and a conditional density which is represented in the joint model. This allows us to estimate the density (or its expectation) by substituting draws from posterior draws from a Monte Carlo Markov chain simulation and averaging over these iterations. For further simplicity, if one implements a truncated Dirichlet process with a maximum number of groups \( C \) this sum is finite, related to the mixture of experts model. Next, we must specify conditions on the forms of the cascading conditional distributions so that the necessary assumptions of the Rubin causal model are met.

One final note on the use of the DPMR, it is motivated by the fact that the Dirichlet process prior is sampling on a space of sample distributions which are dense in the space of all distributions. Suppose that there is a collection of distributions that reasonably represent the relationship between the variables measured. This means for an arbitrary choice of distribution from this collection we can find some distribution sampled from the Dirichlet process which is arbitrarily close to this initial choice. Previous work on the convergence properties of the Dirichlet process proffers some confidence that this approach will be adequately flexible ([16, 15]).

2.4 Integrating a semi-parametric regression modelling approach with causal principles

While Hill has proposed a semi-parametric approach through BART, Wager and Athey suggest that more theoretical work must be done to fully support nonparametric Bayesian approaches [56]. We seek to create a modelling technique that is built up from basic principles that align with the previous causal modelling assumptions and principles to arrive at a highly flexible while theoretically sound modelling approach. We believe that the theoretical basis will lay a groundwork for examination of the convergence properties that can draw on prior work regarding the convergence properties of Dirichlet process prior Bayesian methods.

While most methods reviewed appear to use a strong ignorability argument, the use of nonparametric or single-step estimation processes seems to require a rethinking of this assumption. For example, one exception to the strongly ignorable assumption comes in the Hoshino article [25] which made a global model ignorability assumption that involves only one of the counterfactuals to rest their results on. We are constructing an argument based on a local ignorability assumption and demonstrate the ability to
still estimate unbiased average treatment effects, but also conditional effects.

Several frequentist approaches exist to estimate the average treatment effect that can account for a large number of covariates; however, these frequently suffer from difficulties that arise from a two-step process that estimates the propensity score and then adjust a final outcome model. While doubly-robust methods have been created to protect against misspecifications at either stage of this modelling, an alternate approach is to have a highly flexible model, hence the idea of using a nonparametric approach. While some other nonparametric approaches have been forwarded, the approaches used have had to propose various adaptations of the strong ignorability assumption.

The rationale to integrate a Dirichlet process mixture regression approach to causal modelling is multi-factorial. DPMR models create latent cluster models that are useful for non-linear relationships, can easily incorporate many distributional forms (across many variable types: continuous, discrete, count) and their flexible structure can allow for approximations of multimodal data. Further, the use of DPMR invokes a highly flexible prior that is sampling from a family of distributions when mixed with a base function when used as a hyperparameter in models; this flexibility reflects a frequent desire by scientists to have uninformative priors while retaining the benefits of proper priors with tractable properties. The Dirichlet process creates clusters $S$ that act as latent classes. We propose that this latent structure will split the data (in soft splits, that is the sample space is not partitioned but rather covered by neighbourhoods with overlapping membership) and this can be used to achieve the balance desired in a causal model. The approach is outlined in the next chapter.

There will undoubtedly be drawbacks with such an approach, it will be difficult to add many covariates, as the complexity of the modelling will scale quickly with higher dimensions of confounders, in contrast with propensity score methods which aim to reduce all covariates to a one-dimensional quantity for adjustment. However, by thinking carefully about the particular structural and modelling assumptions needed in building a DPMR causal model, we can ensure that key estimates (like the ATE) will have desirable qualities, and additional estimands (for instance unconfounded conditional estimates) can also be prepared.
Chapter 3

Proposed Dirichlet Process Mixture Regression Approach to Causal Modelling

The most commonly used counterfactual model approaches in the medical literature typically involve the use of propensity scores in either matching, adjustment or stratification approaches\cite{18}. These methods involve a two-step process that first estimates the propensity score and then uses this estimate in a second model which estimates the treatment effect while adjusting for the propensity score. The uncertainty in the first model is most often not represented or accounted for in the second model. Some attempts to redress potential problems which may arise in the first stage of modelling have been proposed, specifically using doubly robust estimators. These estimators are used to protect against model misspecification in either component of the model by carefully choosing the form of the second model, the treatment effect model. Hade and Lu\cite{18} also demonstrate problems with standard and commonly used approaches that use propensity scores as a regression parameter for covariance adjustment, showing that it can result in significant bias with non-linear models (logistic or Cox regression) and as well as in some linear regression models.

We attempt to address the uncertainty in the propensity score and difficulties with unadjusted confounding in non-linear models through a process that produces estimates from a joint distribution model of all the observables (outcomes, treatment assignment and confounders). In this way, we include uncertainty in both the treatment assignment (propensity) and outcomes simultaneously, as well as capturing the uncertainty in the distribution of the covariate. We model the joint distribution of $f(R_0/R_1, Z, X)$ as a mixture of independent local models with a Dirichlet process prior on the parameters for the mixing\cite{53}. It is important to note here that we write $R_1/R_0$ to denote either $R_0$ or $R_1$ but never the two jointly, we cannot claim ever to have measured both a response under treatment 1 and a response under treatment 0 on the same unit. Therefore we could have no information to estimate a correlation between these terms; hence we never create a parameter that would need estimation from the model. The joint model
is thus written as an infinite mixture of clusters, which are indexed by $S$. This model is thus written as:

$$f(R_0/R_1, Z, X | \Phi) = \sum_{j=1}^{\infty} P(S_i = j | \Phi^o) f_j(R_0/R_1, Z, X | \Phi_j)$$

This process has therefore induced a latent cluster $S$ which adjusts for confounding by becoming locally balanced neighbourhoods (the actual estimation procedure will use a Gibbs sampling technique where observations will be able to move between reasonable clusters thereby accounting for the uncertainty in the balancing that is estimated from the observables. Each local model $f_j$ can then be written as a cascading conditional distribution. This construction ultimately allows one to easily determine functions for $f(R_1 | Z, X)$, $f(R_0 | Z, X)$ and $f(Z | X)$ and estimates from these. These conditional distributions, or functionals of them, for instance, means, quantiles ([51]) or the number of modes, can then be used to visualize regression curves. The specific model choice is partly guided by the need to easily generate these conditional distributions while simultaneously adhering to assumptions needed for unbiased estimates of causal effects (the latter will be outlined in the next section). First, we quickly review typical assumptions of causal models and present a modified assumption and arguments that will be required to produce estimates which adjust for confounding through this modelling strategy.

### 3.1 Assumptions for causal models

In the method proposed here, we adhere to a stochastic counterfactual model of counterfactuals rather than a fixed value counterfactual. In this way we imagine not that there exists one fixed response for each unit (in our work this is typically a patient involved in a non-randomized medical study or responding to some treatment in an observational study) under each possible treatment assignment, but rather that there are probability distributions for each treatment response for each unit. This response is therefore conditional on other covariates - some measured, and potentially others that have not been measured. For each individual, once a treatment is selected, then an observation is drawn from the appropriate distribution. The alternate distributions that were not drawn from are the counterfactuals in this situation. Further, the assumptions made are divided into two sections, a structural assumption regarding ignorability that is required for correct estimation of the average causal effect, and a second set of modelling assumptions that are useful for simplifying the estimating procedure, but could be loosened in future implementations.

#### 3.1.1 Structural assumption

For our model, we will need to assume no local unmeasured confounders; however, it is important to note that this does not equate to “no unmeasured confounders” globally. This is clearer when we think in terms of the conditional probability interpretation of no unmeasured confounders. When we assume the usual causal model assumption of strong ignorability, we can write this as:

$$(R_0/R_1) \perp Z | X$$  \hspace{1cm} (3.1)$$

We interpret this conditional probability as stating that we are convinced that in an analysis that includes $X$ there are no unmeasured confounders. We are essentially stating that once we know the values of covariates $X$, knowing the treatment assignment $Z$ provides us with no additional information about
the distributions of our counterfactuals ($R_0/R_1$). In our proposed model we will use a local (weak) ignorability assumption only, and we can express this as:

$$(R_0/R_1) \perp Z \mid (X, S) \quad (3.2)$$

Here one can interpret this in the following way, within a local cluster of our data, we believe that we have measured all the covariates that allow us to have independence in the local estimate of the treatment assignment and the local responses of the counterfactual outcomes. It is important to note that these two assumptions are not strictly equivalent, that is we cannot prove global conditional independence as in 3.1 from an assumption of local conditional independence as in 3.2 and pictured in figure 3.1.

![Figure 3.1: A schematic diagram of the first assumption. Squares are calculated, and circles are randomly generated conditional on the values of the parent values.](image)

However, if we proceed with an assumption of no unmeasured local confounders, we may have an advantage with our modelling strategy vis-à-vis identifying the presence of missing global confounding variables. In Chapter 4 we demonstrate a situation in which we explicitly exclude an important global confounder to determine the effect of our modelling under the local ignorability assumption. This approach may allow us to create joint models where conditional independence does not exist globally and in these cases, we may find the evidence of an unmeasured confounder.

In Dawid’s article regarding causal modelling without counterfactuals, he posits that there is an underlying full bivariate model of the counterfactual responses, he denotes this as $P$ and writes the “implied marginals” for observed $Y_t$ for treatment and $Y_c$ for control as $P_t$ and $P_c$ respectively. While he is not explicitly discussing observational data in his approach, much of his argument rests on the unknown covariance in this model that is ultimately not testable. Several of his arguments rest on taking different assumptions on the relationship between these marginals. While we still must make an assumption of ignorance of correlation between $R_1,j$ and $R_0,j$ in our model (our equivalent to his marginals) within each local cluster, it would seem that by creating clusters that are based on similar characteristics in the confounders $X$, it is possible that our approach may give some insights into the relationship between $P_t$ and $P_c$ in his formulation. Perhaps this approach can resolve some of the tension between these two viewpoints.
3.2 Linking the DPMR to Causal Modelling principles

While standard approaches thus introduce a balancing score or a propensity score, in our approach, we put forward a modified assumption of locally ignorable treatment assignment and use a model for the joint distribution. In this way, we deviate from the use of an explicit propensity score and instead use a more general joint density representation which includes a model of treatment assignment. It allows us to obtain the average treatment effects if desired, while simultaneously allowing for the possibility of expressing probability models for the outcome conditioned on covariates. In our model we also seek to give conditional estimates, to guide a patient-centred medical approach. The inclusion of a Dirichlet process prior in conjunction with the parameters in the local model of \((R_0/R_1, Z, X)\) results in a local subclassification process similar to the impact of a balancing score, dividing observations into groupings with a similar propensity. This subclassification process is further strengthened by the flexibility of the MCMC algorithm implementation of the DPMR which samples a variety of suitable classes for each observation across iterations.

In this approach, it is useful to conceptualize the observational data as being generated from four curves (surfaces) that define the process to be uncovered in any analysis: 1) the distribution of the response under treatment 0, \(f(R_0 | X)\), which is possibly wholly defined by \(E(R_0 | X)\); 2) the distribution of the response under treatment 1, \(f(R_1 | X)\), again potentially entirely determined by \(E(R_1 | X)\); 3) the propensity of treatment assignment given \(X\), equivalent to \(E(Z | X)\); and 4) the distribution of the confounding covariates \(X\), \(f(X)\). We imagine that a process that gives rise to an outcome arises in reverse order, a person seeks treatment and at that time there is a realization of their covariates \(X\). These covariates influence how a clinician might recommend a suitable treatment or may psychologically influence the acceptability of the treatment options to the patient resulting in a realization of a particular choice of treatment \(Z | X\), and until the choice is made the possible outcome distributions exist for this person of \(R_0 | X\) and \(R_1 | X\). Now given a particular choice a realization from the appropriate distribution is realized at some follow-up point after the treatment has been administered.

Given that theorems in causal modelling typically include assumptions related to the distribution of \(X\), we propose that it is critical to include it in our joint distribution and thus to model it for a complete understanding of the treatment effects. An example of curves that define a process that meets the description provided above, which is used in the first simulation of this thesis, is illustrated in figure 3.2. We assert that these same curves can be posited as surfaces for problems of higher dimension, and non-linear relationships can easily be represented.

3.3 Creating causal estimates of the average treatment effect from a DPMR model

We wish to model \(f(R_0/R_1, Z, X | \Phi)\), as a mixture of cascading conditional probabilities, however, it is crucial at this stage to clarify and specify the necessary assumptions to ensure that the model designed will lead to unbiased estimates.
3.3.1 Impact of the structural assumption of weak ignorability

The usual assumption of strong ignorability is replaced by a proposed local weak ignorability assumption, that is:

\[
(R_0/R_1) \perp Z \mid (X, S), \text{ or equivalently,}
\]

\[
f(R_0/R_1 \mid Z, X, S, \Phi) = f(R_0/R_1 \mid X, S, \Phi)
\]

Furthermore, the complete parameter space \( \Phi \) can be divided into parameters related to each conditional probability term in the cascade and parameters related to group membership, for example, \( \Phi^o = \{\alpha, P_{1:\infty}\} \). This allows us to write: \( \Phi = \{\Phi^{R_0/R_1}, \Phi^Z, \Phi^X, \Phi^o\} \), and the overall model can be expressed as the sum of the probability of being a member in a particular cluster multiplied by the joint distribution of the cluster’s local model:

\[
f(R_0/R_1, Z, X \mid \Phi) = \sum_{j=1}^{\infty} P(S_i = j \mid \Phi^o)f_j(R_0/R_1, Z, X \mid \Phi^{R_0/R_1}, \Phi^Z, \Phi^X)
\]

Effectively the Dirichlet process mixture regression is a mixture model (akin to the mixture of local experts model proposed by Jacobs, Jordan, Nowlan and Hinton in 1992 \cite{28}) of local joint distributions. These local clusters are indexed by the induced latent variable \( S \). Thus parameters for each set described above (\( \Phi = \{\Phi^{R_0/R_1}, \Phi^Z, \Phi^X, \Phi^o\} \)) in the cascade can be further split into components that are common to that term and those that are indexed specifically to a single group \( j \), e.g. \( \Phi^{R_0/R_1} = \{\Phi^{R_0/R_1}_j, \Phi^{R_0/R_1}_{j=1}, \Phi^{R_0/R_1}_{j=2}, \ldots, \Phi^{R_0/R_1}_{j=R_0/R_1}, \ldots\} \). The local joint density can then be expressed as follows (using the local weak ignorability assumption above between the second and third lines):

\[
f_j(R_0/R_1, Z, X \mid \Phi) = f(R_0/R_1, Z, X \mid \Phi^{R_0/R_1}_j, \Phi^Z, \Phi^X)
\]

\[
= f(R_0/R_1 \mid Z, X, \Phi^{R_0/R_1}_j)f(Z \mid X, \Phi^Z)f(X \mid \Phi^X)
\]

\[
= f(R_0/R_1 \mid X, \Phi^{R_0/R_1}_j)f(Z \mid X, \Phi^Z)f(X \mid \Phi^X)
\]

\[
= f(R_0/R_1 \mid X, \Phi^{R_0/R_1}_j, \Phi^Z)f(Z \mid X, \Phi^Z, \Phi^X)f(X \mid \Phi^X, \Phi^X)
\]  

\[ (3.3) \]

\[
(3.4)
\]
So each term in equation 3.4 captures the cascading conditional components related to the confounders, treatment assignment and counterfactual outcomes, and depends only on parameters that correspond to that variable. In each, some of the parameters are cluster specific and are indexed by $j$ and others are common to all clusters (for example precision parameters) and are indexed by $\bullet$.

**Theorem 3.3.1 (Weak ignorability induced balance).** If $(R_1, R_0) \perp Z \mid (X, S)$ then

$$
\theta = \mathbf{E}(R_1) - \mathbf{E}(R_0) \\
= \int \sum_{i=1}^{\infty} \{E[Y \mid Z = 1, X = x, S = s_i] - E[Y \mid Z = 0, X = x, S = s_i]\} \mathbf{P}(s_i \mid X = x)f(x)dx
$$

**Proof.** First we condition on cluster $S = s_i$, then use Bayes’ law and sum over all clusters using the law of total probability. Each cluster has probability $\mathbf{P}(S = s_i)$ of membership. Next, we condition within each cluster on the covariate $X$, applying Bayes’ rule again and integrate over the support of $X$ again using the law of total probability.

$$
\theta = \mathbf{E}[R_1] - \mathbf{E}[R_0] \\
= \sum_{i=1}^{\infty} \mathbf{E}[R_1 \mid S = s_i] \mathbf{P}(S = s_i) - \sum_{i=1}^{\infty} \mathbf{E}[R_0 \mid S = s_i] \mathbf{P}(S = s_i) \\
= \int \sum_{i=1}^{\infty} \mathbf{E}[R_1 \mid X = x, S = s_i] \mathbf{P}(S = s_i \mid X = x)f(x)dx \\
- \int \sum_{j=1}^{\infty} \mathbf{E}[R_0 \mid X = x, S = s_j] \mathbf{P}(S = s_j \mid X = x)f(x)dx
$$

Now, our assumption of conditional independence of the counterfactuals, $(R_1, R_0)$, and treatment assignment, $Z$, given the covariates within cluster, $(X, S)$, allows us to add $Z = 1$ or $Z = 0$ respectively to each of the two terms. That is the conditional independence assumption means that given that we have conditioned on $(X, S)$ already, adding treatment assignment adds no further knowledge and therefore cannot change the distribution. Further, now that we are examining expectations within each cluster and with particular treatment assignments, then the expectation (derived from the posterior draws) of the cluster’s $R_0$ or $R_1$ distribution can be replaced with the expectation from the observed data $Y$, which
is most certainly also conditioned on the treatment assignment.

\[
\theta = \int_{\infty}^{1} \sum_{i=1}^{\infty} E[R_1|Z = 1, X = x, S = s_i] P(S = s_i|X = x)f(x)dx \\
- \int_{\infty}^{1} \sum_{j=1}^{\infty} E[R_0|Z = 0, X = x, S = s_j] P(S = s_j|X = x)f(x)dx \\
= \int_{\infty}^{1} \sum_{i=1}^{\infty} E[Y|Z = 1, X = x, S = s_i] P(S = s_i|X = x)f(x)dx \\
- \int_{\infty}^{1} \sum_{j=1}^{\infty} E[Y|Z = 0, X = x, S = s_j] P(S = s_j|X = x)f(x)dx
\]

\[\theta = \int_{\infty}^{1} \sum_{j=1}^{\infty} \{E[Y|Z = 1, X = x, S = s_j] - E[Y|Z = 0, X = x, S = s_j]\} P(S = s_j|X = x)f(x)dx \quad (3.5)\]

While we have proven the result desired, we must also now demonstrate how this quantity may be estimated, and one simplifying assumption will be needed to make this more easily estimated.

### 3.3.2 Modelling assumptions used to simplify estimation procedures

So now we can base a causal estimate based on what is reflected in equation 3.5 which depends on an integral with four key terms. The two expectations \(E[Y|Z = 1, X = x, S = s_j]\) and \(E[Y|Z = 0, X = x, S = s_j]\) are locally modelled in our proposed joint density approach in equation 3.4 as the term \(f(R_0/R_1|X, \Phi_{R_0/R_1})\), and so draws of the parameters involved in this term \(\Phi_{R_0/R_1}\) from a Markov chain Monte Carlo (MCMC) sample can be used to estimate the expectation needed for these terms. The second term, \(P(s_j|X = x) = p_j(X)\) represents the probability of group membership conditional on the covariate value of \(X\), and the final term \(f(x)\) represents the density of \(x\) in the entire population. This vector of \(p_j(X)\) can be thought of as representing latent propensity score classes. These last two terms could be estimated from an MCMC sample; however, the entire process is simplified further if an additional simplifying model choice is made.

Specifically, if one chooses local models where the expectation of \(E[f(R_0/R_1|X, \Phi_{R_0/R_1})]\) is invariant over values of \(X\) then the estimate will eliminate any dependence on \(X\) at all. For instance, if the expectation is fixed at a constant, as would be the case if the expectations are \(\mu_{R_0}^{R_0}\) and \(\mu_{R_1}^{R_1}\) within each group, then the expression in equation 3.5 will simplify further. Under this modelling choice, and so long as we can exchange the integral and summation (using Fubini’s theorem this implies some conditions on the moments, e.g. finite second order moments), then this results in the following simplification.
$\theta = \int \sum_{j=1}^{\infty} \{E[Y|Z=1, X=x, S=s_j] - E[Y|Z=0, X=x, S=s_j]\} P(S=s_j | X=x) f(x) dx$

$= \int \sum_{j=1}^{\infty} \{\mu^{R_1}_j - \mu^{R_0}_j\} P(S=s_j | X=x) f(x) dx$

$= \sum_{j=1}^{\infty} \{\mu^{R_1}_j - \mu^{R_0}_j\} \int P(S=s_j | X=x) f(x) dx$

$= \sum_{j=1}^{\infty} \{\mu^{R_1}_j - \mu^{R_0}_j\} P(S=s_j)$ \hspace{1cm} (3.6)

This formulation is then akin to a mixture of experts model, and the term $P(S=s_j)$ is acting as a weighting term within a propensity score class to generate the estimate under the predicted density of $X (f(x))$. We can easily achieve the condition of Theorem 3.3.1 by limiting the local model of $f(Z | X, \Phi^Z_j, \Phi^o)$ to be independent of $X$ by careful selection of its model and parameters. Again, one way to meet the conditions of our proposed theorem is to model $E(Z | X, \Phi)$ to rely only on a constant; for example, $f(Z | X, \Phi^Z_j, S = j) = f(Z | X, \theta_{zj}, S = j) \sim \text{Bernoulli}(\theta_{zj})$. That is we are modelling groups with a fixed propensity score within each cluster.

So by choosing the local conditional models of $R_0$ and $R_1$ to rely only on constants without any dependency on $X$, and also choosing the model for $Z$ to similarly have no dependence within cluster on $X$ then marginal equations result which again are dependent on $X$ only through $S$ or $p_j(X)$, implying that conditioning on $X$ or $p_j(X)$ will be equivalent. However, to estimate the average treatment effect, this then further reduces to reliance only on $P(S=s_j)$, or the probability of group membership, which is estimated at each iteration of the MCMC implementation of the Dirichlet process prior mixture model. This simplification is illustrated in figure 3.3. While at first glance it may seem oversimplified, the presence of a theoretically infinite number of available clusters in the Dirichlet process allows for flexibility to accommodate non-linear relationships between variables and greater variability in some regions of the joint distribution space, while still maintaining the need for weak ignorability to provide reliable estimates.

Figure 3.3: A schematic diagram of the consequences of both the structural and modelling assumptions.
Hence, overall the joint distribution for \((R_0/R_1, Z, X)\) within the \(j\)th cluster is modelled by a cascade of conditional distributions as follows to allow for unbiased estimates of the average treatment effect and, as will be demonstrated in the next section, for additional conditional treatment effects.

\[
\begin{align*}
[R_0 \mid X, S = j] &\sim f(R_0 \mid \Phi^R_0, \Phi^o_j) \text{ s.t. } E(R_0 \mid X, S = j, \Phi^R_0, \Phi^o_j) = h_{R_0}(\Phi^R_0, \Phi^o_j) = \mu^R_{j,0} \\
[R_1 \mid X, S = j] &\sim f(R_1 \mid \Phi^R_1, \Phi^o_j) \text{ s.t. } E(R_1 \mid X, S = j, \Phi^R_1, \Phi^o_j) = h_{R_1}(\Phi^R_1, \Phi^o_j) = \mu^R_{j,1} \\
[Z \mid X, S = j] &\sim f(Z \mid \Phi^Z_j, \Phi^o_j) \text{ s.t. } E(Z \mid X, S = j, \Phi^Z_j, \Phi^o_j) = h_Z(\Phi^Z_j, \Phi^o_j) = \mu^Z_j \\
[X \mid S = j] &\sim f(X \mid \Phi^X_j, \Phi^o_j)
\end{align*}
\]

### 3.4 Conditional estimates

While section 3.3 outlines a process by which the average treatment effect can be estimated, the model as we have proposed it is particularly useful for estimating and visualizing conditional effects. These estimates and visualizations can then be used to understand better the likely outcomes for new patients/units with measured covariate values for the confounders which have been included in the model. The visualization of these conditional outcomes can be accomplished in various ways that highlight different aspects of the joint distribution. Specifically, one can visualize the mean of the joint mixture distribution, the density, or the modes of the density. Furthermore, the technique is quite flexible given that there is a prediction of the joint density; so that other functionals of the density can easily be visualized using a similar process to what is outlined below for the modes. For instance, one may wish to generate truncated estimates, for example, modes greater than a specified cut-off. To create these visualizations, we begin with the estimates derived in section 2.3.4 and review them here for ease of demonstration.

\[
f(W_1 \mid W_2, \Phi) = \sum_j f(S = j \mid W_2, \Phi)f(W_1 \mid W_2, \Phi, S = j) = \sum_j pr(S = j \mid W_2, \Phi)f(W_1 \mid W_2, \theta_j) \tag{2.2}
\]

\[
pr(S = j \mid W_2, \Phi) = \frac{f(W_2 \mid S = j, \Phi)f(S = j \mid \Phi)}{\sum_k f(W_2 \mid S = k, \Phi)f(S = k \mid \Phi)} \tag{2.3}
\]

\[
E(W_1 \mid W_2, \Phi) = \sum_j p_j(W_2)E(W_1 \mid W_2, \theta_j) \tag{2.4}
\]

#### 3.4.1 Conditional mean regression

An advantage of modelling with a DPMR for the causal model is that one can get estimates of the conditional expectations such as \(E(R_1 \mid X)\), \(E(R_0 \mid X)\), and \(E(Z \mid X)\), by using equation 2.4. Starting with this equation, we select \(W_1\) to match the desired outcome and \(W_2 = X\). This choice \((W_2 = X)\) results in a need to calculate 2.3 to estimate any of these conditional expectations. In this section, and the two sections to follow, the set of parameters and related hyperparameters \(\Phi\) is presumed in
each density and is not explicitly written. It is presumed that only the necessary elements of $\Phi$ are included in each expression, that is, only the subset of $\Phi$ that is relevant to the particular cluster being considered. Thus $\Phi$ is constrained to parameters and hyperparameters belonging to the variables which are included in the submodel being described. By using a Markov chain Monte Carlo sampling method, we will have $m$ draws of these parameters and hyperparameters from the posterior joint distribution, and can thus calculate estimates for all of these desired quantities in the following way. With the sample of realizations of the parameters $(p_j, \mu^{R1}_j, \mu^{R0}_j, \mu^Z_j, \mu^X_j)_{1:\infty}$ from one iteration, one can then obtain a realization of predictive and conditional predictive distributions as functions of the model.

The formulae for these functions (within each iteration) are the following:

\[
\begin{align*}
E(R_0 \mid X) &= \sum_{j=1}^{\infty} p_j(X)\mu^{R0}_j \\
E(R_1 \mid X) &= \sum_{j=1}^{\infty} p_j(X)\mu^{R1}_j \\
E(Z \mid X) &= \sum_{j=1}^{\infty} p_j(X)\mu^Z_j \\
p_j(X) &= \frac{f(X \mid \theta^X_j)p_j}{\sum_k f(X \mid \theta^X_k)p_k}
\end{align*}
\]

Then a set of posterior conditional means are available, and by taking the mean of these estimates, we can plot over a grid of suitable values of $X$ a curve of expected outcomes, or treatment assignment probabilities. In addition, the credible region for these conditional estimates can be plotted using the desired quantiles from this set of posterior conditional expectation draws. Furthermore, a function of these conditional responses, such as $E(R_1 - R_0 \mid X) = E(R_1 \mid X) - E(R_0 \mid X)$ can also be determined using equation 2.4 in a similar way. By deriving the difference at each iteration and then using this to plot a density of its posterior at a range of values of $X$ we can visualize an expected difference with a credible region to determine regions of the covariate space $X$ where one of the treatments may dominate in effect. Specifically, from the expressions above then one can get a realization of the difference between treatments given $X$ as:

\[
E(R_1 - R_0 \mid X) = \sum_{j=1}^{\infty} p_j(X)(\mu^{R1}_j - \mu^{R0}_j)
\] (3.7)

These conditional response curves could be helpful in guiding treatment choices as they allow a clinician to visualize the expected response for a patient with a specific set of observed covariates. While in the simulations and example problem in this thesis a single covariate is used, one could envision more sophisticated plotting that might help practitioners understand which covariates may be most important in considering and visualizing both the expectation and the range of possible outcomes, by additionally including the credible region for these conditional estimates.

### 3.4.2 Density estimation

While understanding the conditional mean effects is useful in some sense to predict treatment outcomes, further information can be gleaned from a better understanding of the variability of response under each
treatment regimen that can be better appreciated through visualization of the densities of responses, conditioned on covariates. Further, for a clinician to decide the utility of a study for her clinical practice, assessing whether the population under study in a particular observational study has reasonable coverage of the clinical population that she is treating currently, can more easily be done by visualizing the density function of the covariates \( f(X) \). These density estimates can also be of use to policymakers who may wish to decide the utility of a new treatment option by assessing whether the population in their community reflects a similar population in the study. The density can be estimated by use of \( \ref{eq:2.2} \) again in conjunction with \( \ref{eq:2.3} \) by choosing \( W_1 \) to be the outcome of interest \((X, R_1, R_0, Z)\) and \( W_2 \) to be \( X \) with the exception of when we wish to estimate \( f(X) \):

\[
\begin{align*}
  f(X) &= \sum_{j=1}^{\infty} p_j f_X(X^j, \sigma^X) \\
  f(Z \mid X) &= \sum_j p_j f(Z \mid X) \\
  f(R_0 \mid X) &= \sum_j p_j f(R_0 \mid X) \\
  f(R_1 \mid X) &= \sum_j p_j f(R_1 \mid X)
\end{align*}
\]

This calculation is a computationally intense task as a grid of both \( X \) and, when relevant, the outcome of interest will be needed. This grid of \( f(\cdot) \) values can then be visualized in slices (at particular relevant values of the covariates), through contour plots, using 3D perspective graphing or in other manners as relevant to the particular problem to be solved. This approach can be conceptualized as a form of scatterplot smoothing based on the joint density estimate, and other others have proposed higher density considerations, c.f. Scott (2015) [50].

### 3.4.3 Modal regression

In some situations, and with some problems, the regression function of most interest may not be the conditional mean but may instead be the modes of the density function. For an example of this, see Einbeck and Tutz, 2006 [9] who present a study on traffic flow on a four-lane California highway. In this example, two patterns seem to emerge in a plot of the raw data with speed plotted against flow. The vast majority of the data shows a steady flow at high speed, with decreasing speeds with higher flow; however, a subset of the data shows a much stronger relationship in the reverse direction with lower speeds related to lower flow. It was proposed that this pattern arises in situations with traffic congestion and typical regression methods would not elucidate the critical pattern in the data needed for traffic planning. They prepare conditional density plots, and modal regression by choosing an appropriate loss function \( l(\cdot) = -\delta(\cdot) \) and using:

\[
m(x) = \arg\min_a [El(Y - a) \mid X = x]
\]

They also discuss how alternate choices of the loss function can create conditional median or quantile estimates. While they suggest that a search on a grid is computationally intensive, such a grid is useful
for plotting of the joint density as outlined in section 3.4.2 and thus a grid may already be prepared for that. In the case of a much larger covariate space $X$ this will become more challenging and will require alternate approaches. In some more complex problems where a density estimation is not required but modal regression may be beneficial, one could make use of the proposals in Einbeck and Tutz’s paper. There, in addition to using the modes of the conditional kernel density estimator, they also propose a search strategy using a mean-shift procedure which may be applicable to data with larger sets.

### 3.4.4 Pointillism plots - a computationally less expensive proposal

Another visualization technique that we propose we refer to as a “pointillism” plot. It is introduced in an attempt to address the computational complexity of generating the density or modes. If one considers the density estimation to be a type of kernel smoothing technique around the multivariate centres identified in the MCMC algorithm from the posterior, then a plot of these cluster centre points visually weighted by their probability ($p_j$) using a grey scale, gives a sense of both where the modes may accumulate. It can act then as a rudimentary version of the density estimation. This plot gives a reasonably clear indication of how the overall structure of the joint distribution is being created through cluster centres. When using a truncated Dirichlet process prior, by comparing models using different numbers of clusters, we can also get a more explicit indication of the impacts of this choice, assisting in the assessment of cluster size determination.

Given that the algorithm will be testing many ‘possible’ cluster centres which may blur the images produced by including points which may not ever be retained, the $M$ sampled $\Theta_{1:\infty,m}$ posterior parameter points can be thinned. One strategy to offset this blurring is to set a lower bound to a probability of cluster assignment $p_j$ which is close to the probability of a single observation point being found in this cluster for inclusion in the plot. This approach would eliminate points which most likely represent the model attempts to explore alternate clustering rather than clusters with probable membership. In addition, parameters can be sorted from low $p_j$ cluster probability values to higher $p_j$ values when plotted to ensure that the visualization captures a more representative image by plotting high probability centres on top of low probability ones. Finally, by scaling this on a logarithmic grey scale, we adjust to the capacities of human vision to interpret weights more appropriately. This truncation and sorting require very little in the way of calculation but through our simulations is found to provide a reasonable approximation to other visualizations which are more computationally intensive.
Chapter 4

Simulation Results

4.1 The actual model for the subjects: Omniscient view

Data are simulated under two scenarios to explore the properties of this model and how well it is able to estimate the difference between the counterfactuals in the classical sense of the marginal effect (the average treatment effect) and in conditional models. The first simulation is intended to provide a comparison of treatment effect estimation of commonly used causal approaches in the simplest construction, that is one where there is only one measured confounder. We will then proceed to explore the additional results that can be generated from this method, including densities, conditional effects and other functional estimates from the density estimation (for instance modes). The second simulation aims to understand the behaviour of the model in a very similar set up to the first simulation, however, with the additional wrinkle that two confounders exist (one continuous and one categorical), but only one is considered to be measured by the researcher. In both cases, the data are modelled as if they come from a study of a geriatric population where age is a known and measured confounder. For simplicity, in the first simulation, the only covariate/confounder considered is age.

The data are constructed as if they arise from a study which recruits individuals presenting to clinical care at a clinic with an age requirement of being 60 years of age or older, hence the distribution of ages starts at 60 years and is generated as coming from a skewed distribution with a heavier right tail. For the second simulation, we include a second covariate/confounder, gender. Gender is presumed to alter both the probability of treatment assignment and the counterfactual responses; however, we proceed as if gender is not available at the analysis stage (hence represents an unmeasured confounder - a violation of the assumptions of most models). The response variable in both simulations is assumed to be measured on a scale that ranges roughly between 0 and 1, with higher scores being advantageous, and it is further assumed that there are floor and ceiling effects in the response at extremes of age. The study is assumed to be observational; there are no “fixed” or “experimenter manipulated” effects. There are two different potential outcomes of the study considered (one for each treatment); in the case of the second simulation each counterfactual consists of a pair of treatment response functions which is determined by gender. The mean response by age (or age by gender, in the case of the second simulation) and treatment assignment probabilities follow specific curves which are illustrated in their respective section.
Simulation 1 - One measured known covariate

For the first data set of 500 observations, data are simulated corresponding to figures 4.1 and 4.2 (in this figure randomly generated values are in circles and calculated values in squares, where two values are separated by a virgule ‘/’ are independently generated or calculated with the quantity with the matching subscript). The simulation data are sampled sequentially, first a sample of age $X$ is constructed by sampling a value $t$ from the lowest 90%tile of a Gamma distribution with $\alpha = 2$ and $\beta = 0.15$ and then adding 60 to $t$. This corresponds to the blue curve on the right side of figure 3.2. Then a response for both $R_0$ and $R_1$ is sampled for each observation, by adding two sources of random normal noise both with a standard deviation of 0.125, $U$ and $\epsilon_0/\epsilon_1$. One, $U$, is common to both outcomes and represents unmeasured covariates that are not also confounders and the second, $\epsilon_0/\epsilon_1$, is independently sampled for each treatment arm and represents the stochastic nature of the response; both are added to the estimated mean response at that age $X$, each value is calculated from its respective equation found in equations 4.1 through 4.4. Treatment responses decline as age increases in both group and are visualized through the expected conditional response portrayed the left side of figure 4.1. These curves are constructed such that the response functions under both treatments converge to approximately the same level for at the extremes of age in the sample, that is, the youngest and oldest subjects have similar probability of outcomes under each treatment option. For the ‘new’ treatment, $R_1$, it is assumed that the decline in the response over age, starts having impact at a later age and is slower. After the confound, and counterfactual outcomes are generated, a treatment assignment is randomly generated using a Bernoulli distribution with probability given by $E(Z | X)$ which is displayed as the black curve on the right side of this figure. Finally, $Y$ is calculated from $R_0$ or $R_1$ according to the treatment assignment using $Y = Z \cdot (R_1) + (1 - Z) \cdot (R_0)$. The curves are defined by the following functions:

$$E(R_0 | X) = 0.2 + \frac{0.6}{1 + e^{0.4(x-70)}}$$  \hspace{1cm} (4.1)

$$E(R_1 | X) = 0.2 + \frac{0.6}{1 + e^{0.25(x-75)}}$$  \hspace{1cm} (4.2)

$$E(Z | X) = 0.2 + \frac{0.6}{1 + e^{0.35(x-75)}}$$  \hspace{1cm} (4.3)

$$f(X) \sim 60 + Gamma(t|\alpha = 2, \beta = 0.15)1_{F(t)<0.9}$$  \hspace{1cm} (4.4)

An analytic calculation of the “true” effect cannot be directly calculated, and so this effect was estimated from the data using a Monte Carlo integration estimate. Since we are interested in the estimate of the expectation of the difference between the counterfactuals, we can take a double expectation on a conditional model with respect to $X$ as demonstrated below. Since we can easily sample from $X$ according to the truncated gamma distribution outlined in 4.4, we can use the expectations as given by equations 4.1 and 4.2 to find the difference as in 4.5. The integral is then estimated by averaging the values...
Figure 4.1: Simulation 1: Left: Response curves for each treatment. Right: Propensity and distribution of X.

Figure 4.2: Simulation 1: A schematic diagram of the simulated dataset. Squares are calculated, and circles are randomly generated conditional on the values of the parent values.

calculated in the middle term from the sample of X drawn from $f(X)$.

\[
E(R_1 - R_0) = E[E(R_1 - R_0 \mid X)] \\
= \int (E(R_1 - R_0 \mid X)f(X)\,dX \\
= \int [E(R_1 \mid X) - E(R_0 \mid X)]f(X)\,dX \\
= \int [0.2 + \frac{0.6}{1 + e^{0.4(X-70)}} - 0.2 - \frac{0.6}{1 + e^{0.25(X-75)}}]f(X)\,dX \\
= \int \left[ \frac{0.6}{1 + e^{0.4(X-70)}} - \frac{0.6}{1 + e^{0.25(X-75)}} \right]f(X)\,dX
\]

(4.5)

A sample of 1,000,000 data points were used and this yielded a “true effect” of 0.1171. Restricting this estimate to the 500 observations that were generated for this simulation yields a sample based effect of 0.1147. This sample effect is estimated by taking the mean difference between the two counterfactuals which were randomly generated for the simulation.
Simulation 2 - One measured covariate and one unmeasured covariate

For the second data set, 1000 observations of data are simulated corresponding to figures 4.3 and 4.4, and approximately 500 observations per gender. In this simulation, gender is designed to have influence on all variables: the covariate $X$, treatment assignment $Z$ and treatment response $(R_0, R_1)$. A slightly different distribution of age compared with the first simulation was used: in this dataset women were sampled from a distribution with older ages, and both the response and treatment assignment are influenced by age and gender in notably different ways. The data are sampled sequentially, as in the first simulation. First a gender, $X_g$, is sampled from a Bernoulli trial with 0.5 probability. Then age, $X_a$, is constructed by adding 60 to a sampled value of $t_g$ from the lowest 90\% tile of the gender specific Gamma distribution: $\alpha = 1.8$ and $\beta = 0.15$ for men and $\alpha = 2.2$ and $\beta = 0.15$ for women, that is $X_a = 60 + t_g$. Then a response for both $R_0$ and $R_1$ is sampled for each observation, by adding random normal noise with a standard deviation of 0.25 to the estimated mean response at that age $X_a$ from each gender specific curve. All treatment response functions decline as age increases, but at different rates and starting at and descending to different plateaus depending on gender and treatment assignment. Next, a treatment assignment is randomly generated using a Bernoulli distribution with probability given by $E(Z \mid X_a, X_g)$; men are more likely to be assigned treatment 1 at older ages and women at younger ages. Finally, $Y$ is assigned from $R_0$ or $R_1$ according to the treatment assignment. The curves defining these expectations (or densities) are as follows:

\[
\begin{align*}
E(R_0 \mid X_a, X_g = \text{woman}) &= 0.2 + \frac{0.6}{1 + e^{0.4(x-68)}} \\
E(R_0 \mid X_a, X_g = \text{man}) &= 0.2 + \frac{0.6}{1 + e^{0.4(x-72)}} \\
E(R_1 \mid X_a, X_g = \text{woman}) &= 0.1 + \frac{0.8}{1 + e^{0.25(x-75)}} \\
E(R_1 \mid X_a, X_g = \text{man}) &= 0.3 + \frac{0.4}{1 + e^{0.25(x-75)}} \\
E(Z \mid X_a, X_g = \text{woman}) &= 0.2 + \frac{0.6}{1 + e^{0.35(x-75)}} \\
E(Z \mid X_a, X_g = \text{man}) &= 0.2 + \frac{0.6}{1 + e^{0.35(x-75)}} \\
\end{align*}
\]

\[
f(X_a \mid X_g = \text{woman}) \sim 60 + \text{Gamma}(t|\alpha = 2.2, \beta = 0.15)I_{F(t) < 0.9} = 60 + f(t_w) \\
f(X_a \mid X_g = \text{man}) \sim 60 + \text{Gamma}(t|\alpha = 1.8, \beta = 0.15)I_{F(t) < 0.9} = 60 + f(t_m)
\]

A similar process as described in the section on the simulation with one measured known covariate was used to calculate the “true” effect under this more complex model, however, here two samples of equal size were used, each from the respective sample spaces of $X_g$ to account for the mixture of responses from men and women. The expected mean differences were estimated by taking the differences within each gender sampled. A sample of 1,000,000 data points per gender was used and then averaged. This process yielded a “true effect” of 0.1054, a difference of 0.0079 for men and 0.2028 for women. The sample difference, again based on the 1000 sampled points for the second simulation is 0.1111, with 0.0055 for men and 0.2159 for women.
Chapter 4. Simulation Results

Figure 4.3: **Left**: Response curves under each treatment by gender. **Right**: Propensity and distribution of X by gender.

Figure 4.4: A schematic diagram of the simulated dataset. Squares are calculated, and circles are randomly generated conditional on the values of the parent values.

### 4.2 The structure for the DPMR: the statistician’s view

As proposed in section 3.3, we first model the joint distribution of \((R_0, R_1, Z, X)\) as an infinite sum of local joint distributions and from this, we get the model for the observables \((Y, Z, X)\). Here it is important to note that for both simulations, we presume to have only one confound, while this is true for the first simulation it is not true in the second, where gender represents a second categorical confound that has been ‘accidentally’ omitted. Thus the model outlined here remains identical and is used for both simulations. Each iteration is modelled in such a way that there is a constant mean within clusters for each of the counterfactual outcomes and a constant probability of treatment assignment to each group within each cluster. We propose for these simulations that there will be a normal distribution locally for each of the counterfactuals and the covariate age. The structure within each cluster of the hierarchical model is thus the following:

\[
R_0 \mid X, S = j \sim N(\mu_{R_0}^j, \sigma_R^j)
\]

\[
R_1 \mid X, S = j \sim N(\mu_{R_1}^j, \sigma_R^j)
\]

\[
Z \mid X, S = j \sim Bernoulli(\pi_j^Z)
\]

\[
X \mid S = j \sim N(\mu_x^j, \sigma_X^j)
\]
In other scenarios, we could replace the local normal distribution assumption with a more appropriate model if the outcome or covariate in question had a different local structure (for example, Poisson for count data, binomial for logistic). Further, we presume that we are observing $Y$ based on the treatment assignment $Z$ in the following way: $Y = Z \cdot R_0 + (1 - Z) \cdot R_1$, which then yields the following distribution for $Y$:

$$
[Y \mid X, Z, S = j] \sim N(\mu_j R_0^1_{(Z=0)} + \mu_j R_1^1_{(Z=1)}, \sigma^R)
$$

The hierarchical priors for the parameters of this model are chosen separately for each of the observed data elements $(Y, X)$ to reflect two considerations: 1) the possible variation in response in the variable of question and 2) to capture two possible extreme states for the cluster representation of this data, described in the next paragraph. The range of possible values for the covariates and outcomes is typically a scientific question, one which prior knowledge of the matter under study could address. For instance, perhaps it is known that the covariate or outcome has a certain anticipated span of reasonable values. In our specific example, we take the range of ages to be from 60 to 85, imagining that the data arise from a clinic that has age restrictions on the population seen. Further, we imagine the scientist has told us that they anticipate outcomes to be from -0.5 to 1.5 on the ‘scale’ used for this study. The mean of these hierarchical values are again thought of as scientific questions, and for this particular simulation, we choose the midrange of the values, but a scientist may have prior knowledge that in their population of study they anticipate a different midrange value.

The critical aspect to determining precision (or variance) hyperparameters is the range of the data likely to be observed (as described in the previous paragraph), coupled with a consideration of two possible clustering outcomes. At one extreme one might imagine that the data can be described as arising from one single cluster and then all the posterior “neighbourhood” draws could be imagined as overlapping spheres that could potentially contain all the data. By this, we mean that any sphere that contains some data captures the full range of the possible observations from each variable and the sphere is centred near the expected means of the corresponding variable. At another extreme clustering state, one can imagine a “string of pearls” with all clusters lined up side by side in a straight line with minimal overlap, spanning the length of the range of the entire data. It is important to note that a curve that is not a straight line may be accomplished by more balls of the same diameter, or smoothing over the same number of spheres as the straight string of pearls with larger diameters. This reasoning procedure places an upper and lower bound on the variances within each cluster and similarly but inversely proportional upper and lower range on the location of the cluster means. While the Dirichlet process has, in theory, an infinite number of clusters, one needs to think in terms of the number of occupied clusters, especially as in our case, we will be using a truncated Dirichlet process prior representation, and hence have a fixed maximum number of occupied clusters, $C$. A reasonable range of sphere sizes to allow for various configurations of the joint distribution which can cover the observed data and a reasonable range of possible cluster means for these spheres can be approximated in the following way: we imagine having to implement a “string of pearls” in a linear fashion and therefore take the expected range of the data for each variable and then divide this range by the number of clusters to get an “average sphere size”. We then select a variance for clusters that will capture a reasonable amount of data within the sphere (for instance by taking 1/4 of this local sphere size as the variance in a normal distribution so that approximately 95% of data will fall within these bounds).

Finally, to select hyperparameters on the probability of $Z$, we draw on the positivity assumption.
Given that we wish to only make comparisons where we have some comparison group to create balance with, we wish to at minimum have one data point from each treatment group in a cluster. Here we make an assumption that observations are distributed evenly amongst clusters. This would suggest creating a lower bound of probability of $C/(N/2)$ and an upper bound of $C\cdot(N-1)/(N/2)$. We take the centre of the distribution to be 0.5 or balanced treatment assignment. This reasoning gives rise to the following hyperparameters (assuming $C=10$ clusters, these would change for a larger number of clusters as per the descriptions above):

$$\mu_j^{R_0} \sim N(\mu_0^{R_0}, \sigma_0^{R_0}), \quad \mu_j^{R_1} \sim N(\mu_0^{R_1}, \sigma_0^{R_1}), \quad \mu_j^Z \sim N(\mu_0^Z, \sigma_0^Z), \quad \mu_j^X \sim N(\mu_0^X, \sigma_0^X),$$

and the standard deviations/precisions given by:

$$\sigma^{R_0} \sim Unif(0.0475, 0.475), \sigma_0^{R_0} \sim Unif(0, 0.475), \sigma_0^Z \sim Unif(0, 1.3), \sigma_0^X \sim Unif(0.5, 6.5),$$

The Dirichlet process is implemented using a truncated stick breaking algorithm [51] (see section 2.3.2 for details), sampling $V_{1:(C-1)}$ from $Beta(1, \alpha)$, and setting the final $V_C = 1$ for the last cluster. Additionally, $\alpha$ is given a prior distribution of $Unif(0.1, 20)$.

### 4.3 Simulation Results

R and OpenBUGS [5, 38] are used to implement a truncated Dirichlet process prior with a range of number of clusters (specifically, $C = 10$, $C = 20$, $C = 30$ and $C = 40$ groups) and using a stick-breaking implementation for the cluster probabilities. Using a range of clusters allows us to examine in both simulations the effects of increasing clusters and helps demonstrate how one might determine an optimal cluster size. The model included only constant means for all variables ($R_0, R_1, Z, X$) as outlined previously. Code for the OpenBUGS model, R interface to call OpenBUGS, and R functions for post-processing and plot generation are included in the Appendices. Posterior draws from 3 chains of length 4000 after a burn-in of 1000. Some results (for models with a high number of clusters $C >= 20$) were thinned by 1 in 2 or 1 in 5 when memory allocation for post-processing analysis procedure outstripped available resources to accommodate memory limits and create plots.

#### 4.3.1 Simulation 1 - One measured known confounder

**Average Treatment Effect**

The estimated values for the overall treatment effect were very similar in all the methods used, as displayed in Table 4.1. This table includes methods that are commonly used, as described in section 2.2.1 the DPMR (with the number of clusters equal to 10, 20, 30 and 40), a propensity score stratification method (with 10 and 20 strata respectively), a simple ANCOVA model, and an inverse probability weighted regression. All methods find strong evidence of an overall marginal treatment effect and confidence intervals and credible regions are comparable. The DPMR seems to have the closest overlap with the stratification (i.e. PS) and covariate adjustment results. There is little variability between
Table 4.1: Simulation 1: Average treatment effect estimates

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>SE</th>
<th>LCR/LCI</th>
<th>UCR/UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>0.1171</td>
<td></td>
<td></td>
<td>0.1147</td>
</tr>
<tr>
<td>DPMR (C=10)</td>
<td>0.1342</td>
<td>0.0226</td>
<td>0.0975</td>
<td>0.1709</td>
</tr>
<tr>
<td>DPMR (C=20)</td>
<td>0.1323</td>
<td>0.0226</td>
<td>0.0944</td>
<td>0.1685</td>
</tr>
<tr>
<td>DPMR (C=30)</td>
<td>0.1318</td>
<td>0.0209</td>
<td>0.0973</td>
<td>0.1658</td>
</tr>
<tr>
<td>DPMR (C=40)</td>
<td>0.1309</td>
<td>0.0215</td>
<td>0.0949</td>
<td>0.1657</td>
</tr>
<tr>
<td>PS (C=10)</td>
<td>0.1323</td>
<td>0.0183</td>
<td>0.0964</td>
<td>0.1682</td>
</tr>
<tr>
<td>PS (C=20)</td>
<td>0.1337</td>
<td>0.0186</td>
<td>0.0973</td>
<td>0.1701</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>0.1358</td>
<td>0.0175</td>
<td>0.1015</td>
<td>0.1700</td>
</tr>
<tr>
<td>IPTW</td>
<td>0.1353</td>
<td>0.0235</td>
<td>0.0891</td>
<td>0.1815</td>
</tr>
<tr>
<td>IPTW (svyglm)</td>
<td>0.1353</td>
<td>0.0266</td>
<td>0.0831</td>
<td>0.1875</td>
</tr>
</tbody>
</table>

SE, standard error; LCR, lower credible region; LCI, lower confidence interval; UCR, upper credible region; UCI, upper confidence interval.

results from the DPMR models with different numbers of clusters and all credible regions include the true value and the sample value.

In figure 4.5 we can see more clearly that the DPMR is close in estimation to all the methods it is compared with through this simulation. The IPTW produces estimates with larger confidence bands, but all methods capture the true average treatment effect (plotted as the solid red line) and the sample estimate (plotted as the dashed red line) within their confidence intervals or credible regions.

Cluster Occupancy

Histograms are used to illustrate the number of occupied clusters across truncation levels in figure 4.6 to assess the adequacy of the truncation choice (C=10, 20, 30 or 40). The number of clusters in each MCMC iteration which have an observed data point assigned to a cluster was recorded. The number of these occupied clusters were averaged by chain and the frequency count recorded. These figures demonstrate that almost all 10 clusters are occupied when the clusters are truncated at 10, whereas for larger truncation limits most iterations result in less than the maximal number of clusters being occupied.

Conditional Treatment Effect

The conditional means for \((Z, R_0, R_1)\) were estimated and plotted using a grid across the covariate \(X\), starting at \(X = 60\) to \(X = 86\) in increments of \(\Delta X = 0.2\). In figure 4.7 we examine the model’s prediction of treatment assignment by covariate. All fitted models from \(C = 10\) to \(C = 40\) clusters (always displayed from left to right with increasing cluster counts on the right) give very similar estimates. These conditional expectation estimates are close to the true mean curve, and the true curve is consistently contained within the credible region. As we move toward the edge on these plots, the credible region expands, and the mean tends toward the central value. These edges correspond both to potential areas with little data to support a specific prediction and likely coincide with areas where balance may be harder to achieve. In this way, the credible region appears to expand to capture the uncertainty of prediction in this region of the covariate space. This characteristic of regression using conditional models can be seen consistently through this section.
Figure 4.5: Simulation 1: Average Treatment Effect (ATE) Estimates. The solid red line represents the true treatment effect, and the dashed red line represents the treatment effect in this particular sample. All methods produce very similar effect estimates with similar credible regions and confidence intervals that include both the true and sample treatment effects.

In Figure 4.8, one can visualize the two conditional expectations of response based on the counterfactuals. The true underlying model is pictured in red, and the fitted model in turquoise. As in the previous diagrams, the credible regions have been constructed as the 2.5%ile and 97.5%ile observations from the conditional estimates based on the posterior draws of the parameters for the joint distribution; these credible regions are denoted by the black lines. The credible region of the fitted model contains the true curve most of the time, and it appears that there is little advantage gained by increasing the number of clusters, as all figures are quite similar. It would seem that the model has captured the underlying counterfactual curves reasonably accurately.
Figure 4.6: Simulation 1: Cluster occupancy - the number of clusters that are assigned at least one data point in an iteration is calculated and plotted in a histogram, the mean occupancy of each chain is calculated and noted in each plot. The histograms range from C=10 clusters (left side) to C=40 clusters (right side).
Figure 4.7: Simulation 1: Propensity score \( pr(Z \mid X) \) estimates ranging from \( C=10 \) clusters (left side) to \( C=40 \) clusters (right side). The pink line represents the true propensity score, the blue line is the fitted estimated propensity, and the black lines represent the 95% credible region of this. The raw data used for this simulation is divided into groups and plotted as those assigned to treatment 1 (green points above the graph) and treatment 0 (salmon points below the graph). The estimated propensity well follows the true propensity and is contained within the credible region throughout.
Figure 4.8: Simulation 1: Counterfactual curves for $R_0$ (top row) and $R_1$ (bottom row) ranging from 10 clusters (left side) to 40 clusters (right side). The pink line represents the true conditional mean effect, the blue line is the fitted conditional expectation value and the black lines represent the 95% credible region. The raw data used for this simulation is plotted as grey points. The $R_0$ credible region estimate nearly contains the true values at all regions in $X$ except for a small region between 70 and 74 years of age. The $R_1$ credible region contains the true curve throughout. It is interesting to note that in areas with low amounts of data, the $R_0$ estimate is further away from the true value, but has larger credible regions.
Figure 4.9: Simulation 1: **Left:** Response curves under both treatments. **Right:** Expected difference in treatments by $X$.

From these curves, one can see that the difference is being well captured, and especially using the right side plot, one can detect a region where the credible region is clearly above zero, approximately between 70 and 80 years of age.

In addition to inspecting the counterfactual response curves individually, it is also possible, and particularly useful, to compare the two treatments directly by plotting them on the same graph as in the left side of Figure 4.9 plotted here using the $C = 20$ model fit. Equally helpful is plotting $E(R_1 - R_0 | X)$, as in the right side of Figure 4.9 which examines the predicted difference between the treatments. Since this is estimated at each iteration of the MCMC implementation, it is easy to obtain the posterior mean and credible region using the same approach as with the estimation of the counterfactual response curves themselves, described explicitly in the next paragraph. This latter curve can provide guidance as to regions of the covariate space where the difference between treatments is sufficiently different to lend support for choosing one treatment over another, based on the lower credible region bound (5th %ile) of the difference being greater than zero. This is estimated in Table 4.2. This table’s results are taken from the $C = 10$ and $C = 20$ model fit.

The difference curves have been constructed by calculating the expected difference between the predicted response in $R_1$ and the predicted response in $R_0$ for each iteration (or a thinned subset) on a grid across reasonable values of $X$ using the weighted estimates as predicted by equation 3.7. This process creates a matrix of responses, indexed by location $X$ and iteration $m$; thus the credible region

<table>
<thead>
<tr>
<th>DPMR model</th>
<th>Lower value</th>
<th>Upper value</th>
</tr>
</thead>
<tbody>
<tr>
<td>($C=10$)</td>
<td>68.1</td>
<td>79.7</td>
</tr>
<tr>
<td>($C=20$)</td>
<td>66.9</td>
<td>79.9</td>
</tr>
</tbody>
</table>

DPMR, Dirichlet process mixture regression.

Table 4.2: Simulation 1: Region of the covariate $X$ where treatment 1 is preferred.
can be constructed based on choosing the 5th and 95th quantiles of response predicted at each level of the covariate $X$ observed from this posterior. The region below the lower credible region that rests above the x-axis then represents a region where we can be 95% certain that a positive difference exists between the means. By locating the intercepts, we are able to define a region that can be helpful in clinical decision making (approximately between 70 and 79.5 years of age) where treatment 1 appears to dominate.

Density Estimation

Density estimation (and visualization) can also provide insights into the variability of response, and can potentially be used in clinical situations for decision making by better appreciating the potential outcomes that may be observed. A frequent question put to physicians by their patients is “What are my chances?” and the answer to this question is almost always intended as the conditional one, that is “What are the chances of someone like me (same age, same initial predictors) having a good outcome?” The most frequent response given to a patient is “It depends,” and the content of what follows that statement is typically a clinician’s best ad hoc estimate of the conditional outcomes. However, rarely is the data from studies made available in such a way as to aid physicians in answering this question. Furthermore, assessing the relevance of a study to a new population, one would want to assess the similarity of the original and new target populations, the density estimate of the covariates serves as a useful tool in clarifying the relevance.

The simplest density estimate produced from this simulation is the density of $f(X)$ which is plotted in figure 4.10. In this figure, we can see that there is very little difference in our estimates of the population covariate with increasing numbers of clusters to fit the model. This observation is not unsurprising as each cluster contributes to this estimate in proportion to the predicted membership, and it is likely that smoothing over a small number of clusters can approximate almost any one-dimensional density estimate, so long as it does not contain more modes than the total number of clusters. This process also leads to an estimate that is not smooth as it is likely overfitting data clusters in the joint space to capture the differences in the counterfactual responses more accurately at the expense of the smoothness of the covariate density. However, the estimate has the true density (pictured in red) within its credible bounds. As described before such an estimate can be useful for visualizing whether the group under study appears to be similar or different to the general population that could be offered an alternate intervention or to a specific help-seeking population that presents to a particular clinic depending on the situation.

Next, we move on to examine the contour plots of the conditional density of the counterfactual responses by covariate predictor $X$, as well as plots of the conditional density at selected values of $X$ illustrated as the cross-sectional slice. These plots are also highly similar between analyses conducted with different cluster sizes, and so only the $C = 10$ plots, pictured as the two plots on the left side of each row, and $C = 40$ plots, pictured on the right, are included here.
Figure 4.10: Simulation 1: Distribution of $X$ (age) density estimates ranging from $C = 10$ clusters (left side) to $C = 40$ clusters (right side). The pink line represents the true density, the blue line is the estimated density, and the black lines represent the 95% credible region of the height of the density at each value of $X$. The raw data is plotted below the graph with random noise added to aid in the differentiation of regions of high and low density of observations. The estimated density is quite close to the true density and is contained within the credible region throughout. The lack of smoothness in the density estimate is likely induced by the need of the model also to capture the response variables accurately.
Figure 4.11: Simulation 1: Distribution of $R_0 \mid X$ (top row) and $R_1 \mid X$ (bottom row) from $C = 10$ clusters (leftmost two) and $C = 40$ clusters (rightmost two) plotted both as contour plots and density slices. In the plots of the slices of the density function, the black line represents the model estimate and the red line the true underlying conditional density. Increasing the number of clusters does not appear to improve the fit of the density functions significantly. However, these plots give a clue to how one might be able to provide a clinician with the potential responses for a particular age superimposed from each curve $R_1$ and $R_0$ on the same graph to aid in decision making and to provide a more realistic view of what the possible outcomes may be.
These plots are much smoother than the covariate estimates and well match the true values at most locations. As we approach the boundaries with less observed data (at the high end of the covariate estimate), we can see that the curves begin to diverge with the estimate being shrunk towards the overall mean response across all covariates. This observation makes interpreting these findings important to do in the context of the density of the covariate in Figure 4.10 or the credible region of the responses as illustrated in Figure 4.11 so that we can appreciate the level of uncertainty in our estimate at these points. It is harder to visualize what the range of possible conditional densities might be unless we plot each iterations estimate as a single line and then superimpose the mean of these (the overall estimate pictured here).

The conditional density cross-sectional plots were illustrated using an arbitrary selection of specific covariate values, \( X = \{60, 65, 70, 75, 80\} \). However, if one imagines disseminating findings from such a trial, it may be worthwhile to create curves for covariate values that are common, or to make available the data needed to construct these curves at a variety of other covariate values. Specifically thinking about a situation where a clinician is working with a patient and about to consider two treatment options, then these estimates of the counterfactual curves of \( R_1 \) and \( R_0 \) that match the covariate values of this new individual can be superimposed on the same plot to inform the clinical decision making. This plot would thus demonstrate the distribution of likely responses at this particular covariate value; which can answer the question our patients typically ask. This approach would require a different approach to result dissemination, but given the capacity to send large data files very quickly over the internet, this could be implementable.

Modal Regression

Given that the modelling results in an estimate of the joint density, several different choices are available to visualize aspects of the joint distribution which may solve important questions critical to the researcher. To demonstrate one possible alternative way of visualizing the data, we have created plots of the predicted modes in the conditional densities. The first technique to locating the modes was begun by constructing the joint density from each iteration of the MCMC sample and then identifying approximate modes on this iteration’s density at a range of specific values of \( X = \{60, 60.2, 60.4, \ldots, 86\} \). Specifically, in each density slice (at a particular value of \( X \)) we search for local elevated values along a grid of potential value of the outcomes \( Y = \{-0.5, -0.4875, -0.4750, \ldots, 1.5\} \). The number of times across all iterations that a mode was found on the \( y \)-grid for each conditional density on the \( x \)-grid was recorded and was used to weight the estimate visually plotting these points using a greyscale colour with darker points representing more commonly identified modes. A second approach to finding the modes used a single estimate of the density (the mean estimate of all the density estimates) which thus represents a global density estimate smoothed over all estimates. A similar search strategy as outlined above is then applied with this estimated density across the same grids to find the global conditional modes.

In Figures 4.12 and 4.13 we can see that the estimated conditional modes approximate the true mode (in this particular simulation these also correspond to the mean estimate). The second approach (in Figure 4.13) gives a singular estimate whereas, the first approach (in Figure 4.12) gives a better sense of the range of possible values of the mode (akin to a credible region). The true mode generally appears within the ‘cloud’ of common modes. One could superimpose these plots to have a similar singular estimate and range of values visualization to the typical mean and credible region that is typically used.
for conditional mean estimates.

Pointillism Plots

Both the density plots and modal regression are computationally expensive; however, a simple alternative for visualization of model results is to plot the cluster means, weighting them visually by the cluster probability. We refer to these as "pointillism" plots. They may not be useful for result dissemination but may help with model evaluation. Their use in assessing missing confounders will become clearer when applied to the next simulation which includes missing confounders. The plots of this simpler simulation with no unmeasured confounders are included here to serve as a point of comparison for the next simulation model’s results. We can also begin to appreciate how the previous results were constructed from these cluster means giving a clearer intuition into the manner in which the model is fit.

From figure 4.15 we can see that the propensity score estimated curve from figure 4.7 is a smoothed version of these pointillism curves with smoothing occurring along the x-axis. In figures 4.14 and 4.16 we can also see how the cluster means are distributed along the y-axis capturing some of the extra variation in the data, with darker points generally clustering near the true mean.
Figure 4.12: Simulation 1: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using each iteration’s density estimate. The true mode is plotted with the blue line, and the darkness of points represent the number of times that point was selected as the mode across all the iterations. The cloud of points generally includes the true mode, however, is less smooth and is therefore likely following the simulated more closely.
Figure 4.13: Simulation 1: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using the mean density estimate from $C=10$ clusters (left side) through $C=40$ clusters (right side). The true mode is plotted with the blue line. Due to the use of the grid to search for the mode, we see that the estimate is jagged; however, we can see that the mode is tracking along the expected path.
Figure 4.14: Simulation 1: Pointillism plots of \( R_0 \mid X \) (top) and \( R_1 \mid X \) (bottom) from \( C = 10 \) clusters (left side) through \( C = 40 \) clusters (right side). The true conditional mean is plotted with the blue line. The cluster means \((\mu^X_0,\mu^R_0)\) are plotted using a grey scale so that more highly probable clusters (ones likely to have more membership) are darker and lower probability clusters are lighter. These plots are capturing information about the distribution of \( X \) with lower ages having more membership due to the density of \( X \), balanced with the probabilities of treatment assignment. Since the estimates arise from smoothing over many clusters, we see that the estimates derived from the models with more clusters \((C = 40)\) appear to be closer to the true values.
Figure 4.15: Simulation 1: Pointillism plots of $Z \mid X$ from C=10 clusters (left side) to C=40 clusters (right side). The true propensity score is plotted in blue. Darker points represent clusters with a higher probability of group membership. Most of the dark clusters are close to the true propensity, and in high probability regions for age appear to be symmetric around the true propensity curve.

Figure 4.16: Simulation 1: Pointillism plots of treatment difference $R_1 - R_0 \mid X$ from C=10 clusters (left side) to C=40 clusters (right side). The true conditional expectation of treatment difference is plotted in blue. The cluster means appear to be symmetrically distributed around the true difference, the clustering at some values of $X$ may represent how the curves can be well estimated as a straight line between these clusters, for instance between the ages of 64 and 68 years on all the last four plots.
Table 4.3: Simulation 2: Average treatment effect estimates

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>SE</th>
<th>LCR/LCI</th>
<th>UCR/UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>0.1054</td>
<td>Sample Value</td>
<td>0.1111</td>
<td></td>
</tr>
<tr>
<td>DPMR (C=10)</td>
<td>0.1145</td>
<td>0.0067</td>
<td>0.1043</td>
<td>0.1251</td>
</tr>
<tr>
<td>DPMR (C=20)</td>
<td>0.1125</td>
<td>0.0073</td>
<td>0.0998</td>
<td>0.1229</td>
</tr>
<tr>
<td>DPMR (C=30)</td>
<td>0.1100</td>
<td>0.0070</td>
<td>0.0977</td>
<td>0.1203</td>
</tr>
<tr>
<td>DPMR (C=40)</td>
<td>0.1078</td>
<td>0.0076</td>
<td>0.0947</td>
<td>0.1194</td>
</tr>
<tr>
<td>PS (C=10)</td>
<td>0.1128</td>
<td>0.0044</td>
<td>0.1041</td>
<td>0.1214</td>
</tr>
<tr>
<td>PS (C=20)</td>
<td>0.1122</td>
<td>0.0043</td>
<td>0.1038</td>
<td>0.1205</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>0.1108</td>
<td>0.0049</td>
<td>0.1011</td>
<td>0.1205</td>
</tr>
<tr>
<td>IPTW</td>
<td>0.1101</td>
<td>0.0130</td>
<td>0.0845</td>
<td>0.1357</td>
</tr>
<tr>
<td>IPTW (svyglm)</td>
<td>0.1101</td>
<td>0.0138</td>
<td>0.0832</td>
<td>0.1371</td>
</tr>
</tbody>
</table>

SE, standard error; LCR, lower credible region; LCI, lower confidence interval; UCR, upper credible region; UCI, upper confidence interval.

These various methods of displaying the data, allow us to visualize different aspects of how the model fitting is occurring, and provide insights into the conditional outcomes. Since the cluster means are found at key positions along the x-axis, the smoothed versions, while being computationally expensive, provide a more accurate view of the joint distribution.

4.3.2 Simulation 2 - Two confounders, only one measured

Average Treatment Effect

As in the first simulation, the estimated values for the overall average treatment effect were similar for many of the methods used, as displayed in Table 4.3. However, comparing the differing DPMR methods that use different cluster sizes we find variability in the estimates, and differences compared against the more traditional models. It is also interesting that all methods are overestimating the true ATE, and have estimates closer to the sample mean value. In this example, there is significantly more variability in the credible regions of the various methods. The DPMR seems to have an intermediate credible region size that is between the smaller confidence interval estimate from the propensity score stratification method and the largest confidence interval estimate from the IPTW survey method. In figure 4.17 we can see these distinctions more clearly with the ATE estimates very close to the sample average treatment effect (plotted as the dashed red line) and is further away from the ’true’ mean (plotted as the solid red line).

As before, in addition to determining the overall ATE estimates (which in this case we know will be inaccurate since there is an unmeasured confounder of gender), we can also explore the density estimates in various ways to understand the way the model has fit the data. We will picture the modes of the distributions, surface estimates of the density functions, and plot the cluster centres conditioned by measured covariate so that we can develop a deeper understanding of the structure of the fitted data. This process will reveal clues to the existence of an unmeasured confounder.

Cluster Occupancy

Histograms are again used to assess the adequacy of the truncation level (C=10, 20, 30 or 40) through the illustration of the number of occupied clusters in figure 4.18 for each cluster truncation choice. These
Figure 4.17: Simulation 2 - Comparing Average Treatment Effect (ATE) estimates from various methods. The solid red line represents the ‘true’ treatment effect (ignoring gender), and the dashed red line represents the treatment effect in this particular sample. All the estimates from each method have confidence intervals and credible regions that include the true and sample treatment effects. The DPMR estimates produce effect estimates that improve (are more closely centred around the true effect) with larger cluster size. The propensity score and ANCOVA methods produce very similar effect estimates with similar confidence intervals that include both the true and sample treatment effects. These standard frequentist estimates are more similar to the small cluster results from the DPMR method. Finally, the IPTW methods produce estimates with substantial confidence intervals.
figures demonstrate that almost all clusters are occupied when the clusters are truncated at 10, 20 or 30 whereas for a truncation limit of 40 most iterations result in less than the maximal number of clusters being occupied, with an overall mean of 37.2 clusters occupied.

Conditional Treatment Effect

We begin by plotting the estimate of the propensity score in figure 4.19, here we note that with increasing clusters we end up with slightly different estimates as we increase the cluster size. In all the plots in this section, in addition to the estimated conditional effect plotted in a solid green line, the credible region is plotted with black lines and the true curves for men and women are separately plotted in blue and pink respectively. Finally, the actual data is plotted on the graph with those assigned to treatment 1 plotted above the graph, and those assigned to treatment 0 below. It is apparent from this that the propensity score, while estimating somewhere in the middle of the two curves, is trying to accommodate some aspect of both curves and with more clusters is alternating between them. However, as we would expect, neither is well captured, and the credible region does not capture either curve well. This alternating graph provides our first clue to a significant unmeasured confounder.

Next, we examine the conditional estimates of the mean response under treatment 0 and treatment 1 in figures 4.20. A similar feature is noted with increasing number of clusters the estimate appears to become more unstable and while it still predicts an ATE somewhere in the middle of the two true conditional curves, it begins to oscillate at times between the two. Furthermore, the credible region grows with increasing number of clusters. This instability provides the second clue to an unmeasured confounder.
Chapter 4. Simulation Results

Figure 4.18: Simulation 2: Cluster occupancy - the number of clusters that are assigned at least one data point in an iteration is calculated and plotted in a histogram, the mean occupancy of each chain is calculated and noted in each plot. The histograms range from C=10 clusters (left side) to C=40 clusters (right side).
Figure 4.19: Simulation 2: Propensity score - $pr(Z \mid X)$. The pink line represents the true propensity score for women and the blue line the true propensity score for men, the green line is the fitted estimated propensity from the model (which cannot distinguish these two groups), and the black lines represent the 95% credible region of this estimate. The raw data used for this simulation is divided into groups and plotted as those assigned to treatment 1 (green points above the graph) and treatment 0 (salmon points below the graph). Since the true propensity cannot be determined, it would seem that with increasing clusters we find more fluctuations in the estimate with a broadening credible region.
Figure 4.20: Simulation 2: Counterfactual curves for $R_0$ (top) and $R_1$ (bottom) beginning with $C = 10$ clusters on the left side through $C = 40$ clusters on the right. The green line represents the true conditional mean response for women and the red line the true response for men, the blue line is the fitted conditional mean response from the model (which cannot distinguish these two groups) and the black lines represent the 95% credible region of this estimate. The raw data is plotted in grey. Similar to the propensity score estimate, the estimate begins fluctuating more (vacillating between the two true curves) and the credible region widens with a larger number of clusters.
Figure 4.21: Expected difference between treatments by the covariate X. The true difference for men is plotted in red, the true difference for women in green and the fitted model in blue. The counterfactual differences (given that both were simulated initially, but only one was pretended to be known) are plotted in grey.

We proceed again to compare the two treatments in figure 4.21 by plotting $E(R_1 - R_0 | X)$. Here we use the results from $C = 30$ as the results from the previous plots of $R_0$ and $R_1$ suggest that the estimates have stabilized by this number of clusters. Added to the figure is the differences from the raw data calculated by including the initial counterfactual estimate for each individual and differencing this (it is important to keep in mind this is never truly available but gives us a sense of how the data is actually clustered). Since we are creating an overall treatment effect ignoring the impact of gender, we would most likely predict a broader credible region than is reasonable for men and a narrower one for women, plus the bounds are clearly inaccurate. This finding suggests that identifying missing confounders is critical here to be able to assure oneself that the results can be trusted.

Density Estimation

Again, we proceed with density estimation. Here the utility of the modelling becomes clearer. We begin by examining the density estimate of $X$ as portrayed in figure 4.22. Similar to the conditional estimates, we see that as the number of clusters increases the estimate becomes less smooth and seems to be alternating between the two ‘true’ densities for men and women. The raw data is plotted below the graph for comparison.

The density estimates for the counterfactual responses are plotted in figure 4.22 for $C = 10$ and $C = 40$ to better demonstrate how the model fits the data in such a way as to capture the two true responses in this space. It is clear from these plots that adding additional cluster is what allows the estimate to capture both of these responses better. While there is no covariate available to distinguish the two curves and their respective propensity, there is still a reasonably accurate portrayal of what the outcomes would be at this value of the covariate.
The green line represents the true density for women and the red line the true density for men, the blue line is the estimated density from the model (which cannot distinguish these two groups) and the black lines represent the 95% credible region of this estimate. The raw data used for this simulation is plotted below the graph with random noise added to help in the visualization of the density of points. Intriguingly, as the cluster size increases, it seems as though the predicted curve is vacillating between the two underlying density estimates, becoming increasingly periodic with greater cluster numbers.
Figure 4.23: Simulation 2: Distribution of $R_0 \mid X$ (top row) and $R_1 \mid X$ (bottom row) from $C = 10$ clusters (leftmost two) and $C = 40$ clusters (rightmost two) plotted both as contour plots and density slices (conditional distributions of the outcome at a specific covariate value plotted vertically to the right of the covariate value). In the plots of the slices of the density function, the black line represents the model estimate and the red line the true underlying conditional density. In contrast with simulation 1, here increasing the number of clusters appears to improve the fit of the density functions significantly. Again, these plots could be very useful for a clinician to demonstrate the potential responses at a particular age from each treatment if we superimposed the $R_1$ curve and $R_0$ curve on the same graph. To a researcher, it would also clearly signal that there is a bimodal response for which it would be important to identify a further predictor.
So to return to the question posed with the first simulation that a patient might ask their doctor, “How might I respond to this treatment or that treatment?” This result would still accurately catch that for some individuals at some ages there is a bimodal distribution of response. It can be seen from the cross-sectional plots, that the convergence to the true density is improving with a larger set of clusters. Again here, one could plot the estimated cross-sectional density estimates at a particular covariate value on the same plot to appreciate better the likely outcomes under the two treatment regimes being offered to a specific patient. Here what is most interesting is that despite the inaccuracy in the conditional mean and difference estimates, we still can be informative (to a degree) with the information at hand.

**Modal Regression**

As described in the first simulation another possible approach to visualizing the data is to examine the density estimates for modes. In this simulation, we expect there to be differing number of modes conditional on the covariate $X$, and so here we plot the same curves as previously examined created using a global search, in figure 4.24, and a search on each iteration, in 4.25.

The global mode search was found to be prone to locating small local modes that are of questionable utility which worsen as the number of clusters is increased, whereas by weighting the frequency with which common modes were found from individual MCMC draws leads to a clearer image of where one expects significant modes, and this improves with increasing number of clusters. One could likely improve the global mode search by either suppressing all modes detected with a density estimate value below a certain lower threshold or by visually ‘weighting’ the significance of a mode by using the value (height) of the density. Specifically, one could add a colour using a grey scale, referencing against the mode with the highest density $f(Y_{mode1})$ within each $X$ slice to a black point, and then adjusting the scale towards white relative to the density value at the comparator mode, for instance by using a weight calculated by $w_m = \frac{f(Y_{mode_m})}{f(Y_{mode1})}$. 
Figure 4.24: Simulation 2: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using the mean density estimate from $C=10$ clusters (left side) through $C=40$ clusters (right side). The pink line represents the true counterfactual response for women $E(R_1 \mid X, X_g = w)$ and the blue line the true counterfactual response for men $E(R_1 \mid X, X_g = m)$. Here many small ridges appear as artifacts of attempts by the model to find clusters of data at various values of $X$ that may still create ‘ripples’ in the density. The number of these artifactual modes increase with increased cluster size and suggest that visually weighting the modes by the overall height might be useful.
Figure 4.25: Simulation 2: Modal Regression of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) using each iteration’s density estimate from C=10 clusters (left side) through C=40 clusters (right side). The pink line represents the true counterfactual response for women $E(R_1 \mid X, X_g = w)$ and the blue line the true counterfactual response for men $E(R_1 \mid X, X_g = m)$. In this estimate, we can more clearly see that the modes at each iteration are most often finding the true underlying modes to this joint distribution. With greater cluster size, these modes are clustering much more closely to the true values.
Pointillism Plots

Finally, the pointillism plots are helpful in this simulation in visualizing how the model fit is working in predicting the outcomes and propensity. In figure 4.26 of $R_0$ and $R_1$ by covariate, we can see that the cluster means as cluster size increases are more and more closely attempting to capture the two subgroups more and more accurately. The joint model is finding clusters of responses around not only the average age effect but instead around the age by gender subgroup effects. Here it is important to consider (and remind oneself) that there is no data within the observed data to distinguish which $R_1$ cluster mean is matched to which $R_0$ cluster mean which is what leads to the dispersed propensity and makes estimating the overall treatment mean fraught.

This observation also explains how the larger number of clusters was providing more information that made these distinctions in the density estimates and modal estimates, but that also led to increasingly wide credible regions with fluctuating estimates in the conditional means and overall ATE estimate. However, since there is no way to accurately match subgroups from treatment 1 to subgroups from treatment 2, a crossing of 2 by 2 (within and between genders) potential treatment matches are being made to estimate the treatment differences that are plotted in figure 4.28.

The propensity score model is displayed in figure 4.27. It is clear from these plots that the clusters cannot distinguish the propensity, as any value between the lower threshold and upper threshold are included. Interestingly, as the density of females is lower at both age extremes, the probability of treatment assignment is less balanced by gender at these extremes, and there is little distinction in outcomes in these areas, the model does not span to this lower bound at the higher age levels. It would seem that the model is mixing observations which are suitable comparators to others while trying to continue to capture the counterfactual outcomes (as can be seen in the following plots). This is our third clue that the data is missing a key confound, as a consistent propensity score cannot be estimated at specific values of the confound.
Figure 4.26: Simulation 2: Pointillism plots of $R_0 \mid X$ (top) and $R_1 \mid X$ (bottom) from $C = 10$ clusters (left side) through $C = 40$ clusters (right side). The true conditional mean for men is plotted with the blue line and for women with pink. The cluster means ($\mu^X_j, \mu^R_j$) are plotted using a grey scale so that more highly probable clusters (ones likely to have more membership) are darker and lower probability clusters are lighter. These plots are demonstrating that the model is capturing information about the joint distribution and its bimodal nature. At small cluster sizes the distinction between the modes is much less clear; however, we see that the estimates derived from the models with the most clusters ($C = 40$) that the model is identifying clusters close to the true curves.
Figure 4.27: Simulation 2: Pointillism plots of $Z \mid X$ from C=10 clusters (left side) to C=40 clusters (right side). The true propensity score for men is plotted in blue and for women in pink. Darker points represent clusters with higher probability of group membership. In this simulation, the cluster centres are very diffuse, and high probability clusters can be found throughout. This diffusion can help to identify the possibility of a missing confounder.

Figure 4.28: Simulation 2: Pointillism plots of treatment difference $R_1 - R_0 \mid X$ from C=10 clusters (left side) to C=40 clusters (right side). The true difference for men is plotted in blue and for women in pink. The model cannot distinguish which subgroups match and as the cluster size increases, we can see four distinct lines beginning to form as male and female subgroups for treatment 0 and 1 match and cross with each other.
4.4 Summary

A model for estimating average and conditional treatment effects was proposed and fit to two simple models. The first simulation included only one known confounder, and the outputs were consistent with other approaches. The model was also able to create conditional estimates that were demonstrated through a variety of visualization strategies. These alternatives could be decided between to serve various purposes and to demonstrate specific nuances of a particular question under study.

The second simulation is based on a model with two confounders but where only one continuous confounder was measured, and a key categorical confounder was omitted to understand better how this would affect the modelling. Intriguingly this approach continues to estimate the conditional effects well and provides clues that a confound was missing. As with all other methods, the overall average treatment effect cannot be estimated accurately. Here the different visualization strategies can be seen to be useful in some scenarios suggesting ways we might use the different approaches to capture different aspects of the model results. Therefore this approach seems to include the possibility of examining our model assumptions without needing additional information outside the data available.
Chapter 5

Clinical Application

5.1 The Problem and Data: Homelessness and health care delivery

5.1.1 Motivating Problem

This problem concerns an observational trial of two possible interventions for homeless men to aid them with mental health concerns. While the literature well supports collaborative interventions (treatments that are offered jointly by primary care, mental health specialists, social services, and family members) for many populations, it is not clear whether some models of integrated care are more effective for homeless individuals. This trial sought to compare two possible approaches to collaborative care delivery. One intervention offered integrated multidisciplinary collaborative care (IMCC), where men were seen by a psychiatrist working collaboratively within a shelter-based health care clinic (with a variety of health care professionals) that participants were enrolled as patients. The alternative intervention offered shifted outpatient collaborative care (SOCC) where a psychiatrist saw men within the clinic working alongside shelter employees and coordinating with primary care services that were offered off-site. This latter intervention required fewer resources than IMCC.

The study used a quasi-experimental design with two comparison interventions with non-random treatment assignment. Homeless men, 18 years of age and older, who had one of five different categories of psychiatric diagnoses were assigned to treatment based on the shelter where they were staying at the time of enrolment into the study. The study investigators wish to determine the 'causal effect' of the care model used on outcomes. The outcomes spanned various domains and were measured pre-treatment, at six months, and at 12 months. They included: overall functioning (MCAS - Multnomah Community Ability Scale), housing status and stability, severity of mental health symptoms (BPRS - Brief Psychiatric Rating Scale), presence and severity of alcohol and substance use (ASI - Addiction Severity Index), and health care utilization in the past 6 months (ER use, and hospitalization), or 30 days for physician visits. The particular question about the housing status during the last six months of the study was of particular interest to the research team, as there was a hypothesis that individuals who receive all their primary medical and psychiatric care within the shelter might have a perverse incentive to remain homeless so as not to disrupt needed care. It was also plausible that the lifetime homelessness (which was measured at baseline) might represent the only confounder for this outcome.
This study provided an ideal opportunity to test the application of our model for estimating causal effects. It represents the types of limitations of many clinical trials, including small sample size, a limited set of measured outcomes, many pre-treatment outcomes available but mostly as categorical variables, non-random assignment based on convenience of sample recruitment, and the inclusion of non-normally distributed outcomes. The combination of small sample size and a large number of categorical pre-treatment variables may make causal approaches relying on stratification or matching methods more unwieldy, as it may be difficult to find matching cases, and the elimination of unmatched cases may reduce power to an unacceptably low level. If several of the categorical variables are identified as potential confounders, the cross tabulation of these may create cells where no observations are available, and hence comparisons may not be possible. The variables available may not include all essential confounders if there is only a small number included, if a causal modelling approach was not considered at the outset, or if critical confounders are only discovered during the period a trial is conducted and hence the key assumption of most causal approaches may be violated. Furthermore, non-normally distributed outcomes can easily be accommodated in the joint model approach we take and several of the outcomes that may be considered here (for example, physician access, days homeless, number of emergency room visits) are likely best represented by count or binomial distributions. While our model cannot correct for all of these difficulties, this particular study provides an opportunity to examine the model’s behaviour in a simple situation where we select one clear confounder for inclusion to determine what might be gleaned from this approach. In this research, several of these real-life complications are likely to be present.

5.2 The Data

From table 5.1 it can be seen that there are differences between the groups who access the two treatment models (IMCC versus SOCC). The population who accessed IMCC had been homeless for a considerably longer time in their lifetime (4.0 years compared to 1.8 years), and considerably more time unhoused in the past year 67.1% reported > 90 days in shelters on the living on the street compared with 17.1% in the SOCC group. While the psychiatric severity seemed to be similar, the classes of psychiatric disorders seem to suggest a different type of population with more psychotic illness in the IMCC group and more mood and anxiety disorders amongst the SOCC group. It could be argued that lifetime homelessness may mediate the impact of these chronic illnesses on one’s current capacity to be housed, and the impact of these illnesses may be wholly captured through its impact on lifetime homelessness within clusters of our data. Preliminary examination of the data revealed a high correlation between the various methods in which pre-study homelessness was measured, and ultimately a log-transformed measure of total years homeless was selected as a single confounder to be included in our model. A kernel density estimate of the density of log-transformed total years of homelessness suggested a bimodal distribution within both groups, with slightly displaced means as demonstrated in figure 5.1.

Figure 5.2 presents a plot of these two key variables (recent homelessness at trial completion and lifetime homelessness at baseline); individuals with missing data are represented by points in the margin of the graph. From this plot, it is clear that individuals recall their homelessness in blocks, such as a specific number of weeks or months, and hence the recorded data is clustered at common response values. As is common in many clinical trials, there was also a problem with non-response at the 12 month follow-up period. We do not specifically address this particular problem is not specifically addressed in this chapter; a final model may well need to incorporate a separate model for the missing value mechanism.
### Table 5.1: Baseline demographic and clinical characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>IMCC&lt;sup&gt;a&lt;/sup&gt; (N=70)</th>
<th>SOCC&lt;sup&gt;b&lt;/sup&gt; (N=70)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean age (in years)</td>
<td>42.3(10.8)</td>
<td>42.0(10.6)</td>
<td>0.912</td>
</tr>
<tr>
<td>Single or never married</td>
<td>53(75.7%)</td>
<td>41(58.6%)</td>
<td>0.031</td>
</tr>
<tr>
<td>Less than high school education</td>
<td>32(46.4%)</td>
<td>23(32.9%)</td>
<td>0.103</td>
</tr>
<tr>
<td>English first language learned</td>
<td>54(80.6%)</td>
<td>53(77.8%)</td>
<td>0.704</td>
</tr>
<tr>
<td>English main language</td>
<td>65(94.2%)</td>
<td>64(91.4%)</td>
<td>0.745</td>
</tr>
<tr>
<td>Foreign born</td>
<td>31(44.3%)</td>
<td>21(30.0%)</td>
<td>0.080</td>
</tr>
<tr>
<td>% Caucasian</td>
<td>34(49.3%)</td>
<td>44(64.7%)</td>
<td>0.068</td>
</tr>
<tr>
<td><strong>Homelessness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (IQR)% of lifetime homelessness</td>
<td>4.0(15)</td>
<td>1.8(5)</td>
<td>0.001</td>
</tr>
<tr>
<td>Median (IQR)% of moves in past 12 months</td>
<td>1.0(1)</td>
<td>2.0(2)</td>
<td>0.097</td>
</tr>
<tr>
<td>No. nights spent on streets or in shelters in past 12 months</td>
<td>&lt; 0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 30 days</td>
<td>12(17.1%)</td>
<td>41(58.6%)</td>
<td></td>
</tr>
<tr>
<td>31-90 days</td>
<td>11(15.7%)</td>
<td>17(24.3%)</td>
<td></td>
</tr>
<tr>
<td>&gt; 90 days</td>
<td>47(67.1%)</td>
<td>12(17.1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Self-reported Diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood Disorder</td>
<td>31(44.3%)</td>
<td>51(72.9%)</td>
<td>0.001</td>
</tr>
<tr>
<td>Schizophrenia and Related Psychotic Disorders</td>
<td>41(58.6%)</td>
<td>27(38.6%)</td>
<td>0.018</td>
</tr>
<tr>
<td>Anxiety Disorders</td>
<td>15(21.4%)</td>
<td>31(44.3%)</td>
<td>0.004</td>
</tr>
<tr>
<td>Any Substance Use Problem</td>
<td>8(11.4%)</td>
<td>13(18.6%)</td>
<td>0.237</td>
</tr>
<tr>
<td><strong>Addiction Severity Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug use scale</td>
<td>0.1(0.2)</td>
<td>0.1(0.2)</td>
<td>0.975</td>
</tr>
<tr>
<td>Alcohol use scale</td>
<td>0.1(0.2)</td>
<td>0.1(0.2)</td>
<td>0.445</td>
</tr>
<tr>
<td>Brief Psychiatric Rating Scale</td>
<td>46.2(9.9)</td>
<td>48.8(10.0)</td>
<td>0.133</td>
</tr>
<tr>
<td>Multanomah Community Ability Scale</td>
<td>60.3(9.0)</td>
<td>57.3(10.1)</td>
<td>0.066</td>
</tr>
</tbody>
</table>

<sup>a</sup> Integrated Multidisciplinary Collaborative Care, <sup>b</sup> Shifted Outpatient Collaborative Care.

Copied from Stergiopoulos et al 2015 under the Public Domain Dedication.
Figure 5.1: Estimated density of the logarithm of years homeless by treatment. The red line is the estimate for men in the IMCC arm and the blue line for men receiving SOCC.

if it is believed to be non-ignorable. Such an inclusion would be a natural extension of the principles outlined in this method along with Donald Rubin’s early work on Bayesian approaches to missing data and causal models which he outlined in his 1978 paper [48]. However, for the remainder of this chapter we will ignore a model for the missingness and proceed with an analysis based only on the observed data; we included partial data in the analysis.

5.3 The Model

In this section, we outline the parameters and hyperparameters that we used in the model implemented to address this particular problem. Again we think in terms of the local models (local neighbourhoods defined by membership in $S_j$ a partition of our $\sigma$-field of events) containing three key variables, the covariate $X$, the treatment assignment $Z$ and the outcome $Y$ (which will eventually be represented by two counterfactual responses $R_0$ and $R_1$). We consider each variable in turn and determine the distributional form within each cluster. We represent membership in the $j^{th}$ cluster by the $i^{th}$ observation with $S_i = j$.

- $X$ is the logarithm of the total years homeless, $X|(S = j) \sim N(\mu_j^X, \sigma_j^X)$
- $Z$ is the observed assignment to one treatment or another, $Z|(X, S = j) \sim Bern(\pi_j^Z)$
- $R_0$ is the number of days homeless out of 180 immediately after receiving the control intervention(SOCC), $R_0|(X, Z, S = j) \sim Binomial(\pi_j^{R_0}, 90)$
- $R_1$ is the number of days homeless out of 180 immediately after receiving the comparison intervention(IMCC), $R_1|(X, Z, S = j) \sim Binomial(\pi_j^{R_1}, 90)$
In this way the joint distribution can be written (invoking Bayes’ rule within each local model) as the product of these three distributions. Note that in order to result in a causal model, we have adhered to our assumption that \((R_0, R_1) \perp Z | (X, S)\) by not including any dependence on \(X\) in either \(R_0\), \(R_1\) or \(Z\), each is solely defined by a function that depends only on hyperparameters which partition into respective \(X\), \(Z\) or \(Y\) \((R_0, R_1)\) spaces.

Hyperparameters were decided upon with discussion with the primary research team about the range of values expected and the expected variability in these responses. Similar to the reasoning outlined in the simulation studies, two extreme states were imagined as placing constraints on these values. At one extreme one might imagine that one single cluster describes the data and then all the “neighbourhoods” could be imagined as overlapping spheres that contain the full range of each variable centred near the expected means of each variable. At the other extreme, one can imagine a “string of pearls” with all clusters lined up side by side in a straight line with minimal overlap, spanning the length of the range of the entire data. This procedure places an upper and lower bound on the variances within each cluster and similarly but inversely proportional between the cluster means. These extremes are thus approximated using the range of data at the high end and the range divided by the number of clusters at the low end.

For lifetime homelessness, one can imagine that most people would not consider themselves homeless until they were without housing for 1-2 weeks and for the chronically unhoused this could span the majority of one’s adult life (i.e. up to 50 years). On the log scale this represents \(\log(1/52) = -3.95\) and \(\log(50) = 3.912\), which was approximated as \((-4, 4)\).
Table 5.2: Propensity Score Models

<table>
<thead>
<tr>
<th>Model of $\text{logit}(p)$</th>
<th>Intercept</th>
<th>$X$</th>
<th>p-value</th>
<th>$X^2$</th>
<th>p-value</th>
<th>$X^3$</th>
<th>p-value</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear $\sim x$</td>
<td>0.08641</td>
<td>0.28770</td>
<td>0.00251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quadratic $\sim x + x^2$</td>
<td>-0.29825</td>
<td>0.42670</td>
<td>0.00056</td>
<td>0.12536</td>
<td>0.03162</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cubic $\sim x + x^2 + x^3$</td>
<td>-0.35077</td>
<td>0.19086</td>
<td>0.4366</td>
<td>0.18141</td>
<td>0.0332</td>
<td>0.04260</td>
<td>0.2825</td>
<td>186.03</td>
</tr>
</tbody>
</table>

Each logistic regression model involves progressively higher terms of $\text{logit}(p) = x + x^2 + x^3$

Figure 5.3: These plots contain three elements: propensity score estimates in strata by quintiles of lifetime homelessness (represented by points with error bars centred within each stratum), observed treatment assignment (represented by tick marks either at the top (IMCC) or bottom (SOCC) of the plot, and predicted propensity score models are plotted as curves. Each plot uses progressively higher order covariates ($x, x^2, x^3$) for these propensity models, from a linear model on the left to a cubic model on the right.

For comparison purposes, a propensity score was fit to the data and attempts were made to fit stratified, covariate-adjusted, and inverse propensity score weighted models.

5.4 Results

A sequence of propensity score models was fit using logistic regression and including increasingly higher powers of the one selected confounding covariate (logarithm of total years spent homeless). It can be seen in table 5.2 that by including the square of this variable a better model fit, as measured by AIC, is realized, and both variables are seen to be statistically significant from a frequentist perspective. When we add a cubic term, we see a relative increase in the AIC again, suggesting poorer overall model fit and simultaneously we find that only one of the three variables $x^2$ remains statistically significant. However, the images plotted in figure 5.3 provide contrasting information. In these plots, the estimate of the probability of treatment assignment, within quintiles of the covariate, is represented by the circles and plotted with estimated errors (under a normal approximation). Tick marks represent the observed data at the top and bottom of the graph. Finally, the propensity estimates from the sequence of logistic models with progressively higher order terms are plotted in sequence across the page. One can see that there is a poor fit. The propensity score was then used to create a predicted propensity corresponding to each observation.
Table 5.3: Overall treatment effect

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>SE</th>
<th>LCR/LCI</th>
<th>UCR/UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPMR (C=20) ((p_1 - p_0))</td>
<td>0.2109</td>
<td>0.0677</td>
<td>0.0996</td>
<td>0.3237</td>
</tr>
<tr>
<td>DPMR (C=20), log odds ratio</td>
<td>1.8199</td>
<td>0.5596</td>
<td>0.9379</td>
<td>2.7798</td>
</tr>
<tr>
<td>PS (C=6)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>1.2296</td>
<td>0.0416</td>
<td>1.1421</td>
<td>1.3114</td>
</tr>
<tr>
<td>IPTW</td>
<td>1.2419</td>
<td>0.3906</td>
<td>0.4763</td>
<td>2.0075</td>
</tr>
</tbody>
</table>

SE, standard error; LCR, lower credible region; LCI, lower confidence interval; UCR, upper credible region; UCI, upper confidence interval.

Due to the presence of data clusters (many individuals reported similar lengths of lifetime homelessness), it was not possible to create a stratification of the covariate in any more than six strata, and so this was used to attempt a stratified analysis. However, within the six strata, this yielded a subgroup with perfect separation of the probability of homelessness during the follow-up period. Hence, only the IPTW and covariate-adjusted models are reported and compared with an average treatment effect estimate from the DPMR causal model in table 5.3. A plot of the estimated propensity score and probability of treatment assignment with quintiles of the covariate are plotted for comparison in Figure 5.3.

Examining the outcome of interest (number of days homeless in the past 180 days) at the one-year post-intervention initiation reveals the data as plotted in figure 5.2. Since this is a self-report measure, it is important to note that individuals seem to be responding in quantities that cluster and approximate the actual length of time homeless. For example, an actual period of homelessness of 9 days may elicit a response of ‘one week’ from a participant and then be recorded as seven days; 97 days, reported as three months, is likely recorded as 90 days. While it is most likely the case that there is inaccuracy with intermediate responses in the range of 1-179 days, it is likely true that reports of 0 or 180 days are accurate and potentially abundant in a follow-up period. In the case that a person becomes entirely housed, they will report 0 days and those who remain unhoused will report 180 days. This difficulty with self-report is likely unavoidable unless we can locate some other confirmatory source for residency data.

The Dirichlet process mixture regression model was implemented with a truncated Dirichlet process using 20 clusters. As can be seen in table 5.2 there is comparable results between methods in terms of the mean estimate of the average treatment effect, but different methods result in quite different variance estimates, hence different credible/confidence regions. Cluster occupancy was assessed graphically and is demonstrated in figure 5.4. The average occupancy is 17.18, below the truncation value of 20, and with relatively low percentage of iterations with all clusters occupied.

The density of the covariate (log years homeless) is plotted in figure 5.5 (left side). This estimate is consistent with the bi-modal distribution that was identified in the kernel density estimate displayed in figure 5.1. Furthermore, plotting of the 5th and 95th percentile of the density estimate at each value of X creates a credible region for this density, appreciating that the area must remain constant, it seems quite likely that any estimate generated from this would most likely be bi-modal. This feature is interesting in and of itself and may represent two distinct types of types of homelessness amongst men with mental health struggles.

The DPMR allows us to plot our estimates of the propensity score and in figure 5.5 (right side) we
Figure 5.4: Cluster occupancy - the number of clusters that are assigned at least one data point in an iteration is calculated and plotted in a histogram, the mean occupancy of each chain is calculated and noted in the plot.

see that the DPMR estimate is rather flat at approximately 40% for those with a very brief history of homelessness. It then has an increase between one and three years and a levelling off around five years of lifetime homelessness at 60%. This relationship suggests that one shelter is more likely to attract/accept men who have a longer history of homelessness, after a certain period. This observation is in keeping with the general perception of these two shelters amongst clinicians in the area, where one shelter is recognized as providing services to a more entrenched group of homeless men. Given the findings in the density of $X$ being potentially bimodal, this might be easily explained as a mixture of two distinct propensities for care in each of the two shelters based on these subgroups of the length of homelessness.

Examining the treatment responses conditional on the confounder in figure 5.6 is rather informative, as there is a particularly distinct non-linear relationship for treatment under the IMCC condition, and a different relationship in the SOCC treatment arm. By superimposing these curves in figure 5.7 one can compare these two more directly, and it appears that both groups have somewhat similar outcomes at lower values lifetime homelessness, for those with less than five years homeless. It then appears that the treatments diverge from about three months of lifetime homelessness. This effect seems to then stabilize after five years. This model seems to suggest that while the SOCC treatment has minimal impact on one’s ongoing homelessness, IMCC has the possibility of worsening homelessness. This is more clearly examined by plotting $E[R_1 - R_0 | X]$, as in the right side of figure 5.7 and the point at which the credible region intersects with the x-axis is plotted in green, this defines a region where one treatment clearly outperforms the other, with regards to its impact on further homelessness.
Figure 5.5: **Left**: Distribution of X. **Right**: Propensity score - $Pr(Z \mid X)$.

Figure 5.6: **Left**: Response curve under SOCC, $R_0$. **Right**: Response curve under IMCC, $R_1$. 
Returning to our findings with the propensity score, IMCC was more likely to be offered to the individuals with longer histories of homelessness and was also more likely to extend homelessness for men with this more extensive homelessness. While subsequent homelessness was not the primary outcome for this trial (community functioning was) nor was it a specific target for either intervention, knowing about this consequence might inform decisions about which treatment may be best for men depending on their particular situation. This finding of differential response across the covariate was further explored using additional visualization methods.

In the pairs of plots portrayed in figure 5.8, which represent the findings for SOCC (top) and IMCC (bottom) respectively, the clustering effects are demonstrated in two ways. The left-hand side illustrates where the probability is amassing in the response variable by plotting the centres (probability of a day homelessness*180) of each cluster. They are plotted in greyscale with full black representing the cluster with the highest probability through to white representing a lower bound of probability equalling 0.014. This lower bound is derived by imagining a hypothetical situation where the probability of belonging to a cluster is equivalent to 1 observation of the 70 participants in each group belonging to this cluster 1/70 = 0.014. The plot on the right of each pair is a plot of the predicted density of response under that treatment. Note that the base of the density surfaces have the axes exchanged when compared with the cluster plots, to better visualize the peaks of the surface.

From these plots, there are two features of the clustering that are notable. The first is that of the ridges that appear horizontally across both cluster plots and density surfaces in the middle of the probability. As was noted in the raw data, individuals are likely responding in estimates that cluster around easy to recall date ranges (weeks, months) and these give rise to a density that can be difficult to detect subtle changes between the treatments. The next notable clusters are at the edges of the surface - this is not unexpected as one might expect that some men may become housed or remain entirely
Figure 5.8: Left: Local neighbourhood centres. Right: Density estimate. Top: SOCC. Bottom: IMCC.

homeless during the entirety of the study’s follow-up period. By comparing between treatments, we can see that the model predicts that SOCC has a proportion of individuals who become housed (i.e. there is a concentration of probability at 0) amongst those with a longer than one-year history of homelessness, whereas there is not such a group in the IMCC model.

Finally, visualizing the clusters of treatment assignment probabilities in figure 5.9 demonstrates that there are broad ranges of clusters at various values of the confounding covariate. As in our second simulation, this suggests that there may be additional unmeasured confounders missing from this analysis. However, in this particular analysis, there is also clustering of outcome data likely due to the typical time responses described previously, and this may also be leading to difficulties. The modelling is attempting to match outcome clusters in neighbourhoods of the covariate, but given that several clusters are likely needed to identify the outcome, without an additional covariate there may be no way to narrow in on a
single stable comparator outcome cluster in order to quantify the propensity score within that cluster.

5.5 Comments

Applying the proposed DPMR method of causal modelling to this particular problem has certain benefits and limitations. It demonstrates a similar overall estimate as the propensity score methods for the average treatment effect when compared to the IPTW method, which is more likely capturing the uncertainty in the estimate more accurately. However, it is in the conditional models that we can glean considerably more information about the relationship between the variables and are thus able to identify both clusters in the covariate and the potential impacts this is having on the outcome. There are clues that additional covariates are needed to predict clusters better (and this may correspond to subgroups of patients that scientists in this field may be able to identify). For instance, specific mental health diagnoses may help differentiate groups - psychotic disorders are often quite debilitating concerning patients’ capacity for insight, planning and reasoning and may present a barrier to becoming housed. Whereas episodic mental illnesses like bipolar disorder and depression may cause temporary difficulties with housing that resolve when illnesses remit spontaneously or through treatment. Finally, these diagnostic categories may also inform the settings in which people feel most comfortable and may capture some of the prediction of treatment choice.

The method is limited by the inclusion of only one covariate currently, and this is an area that will require further development for the method to be more applicable to real-life situations, however, with the strong theoretical underpinnings and advances in the use of DPMR with large datasets, we do not envision this as insurmountable. An additional challenge will be the visualization of data in higher dimensions, much of what makes this method powerful is the capacity to gain insights through the flexible options for data visualization. To retain this advantage, visualization methods that allow for the identification of regions of the covariate space where significant differences in treatment effects exist are essential.
Further, in this application, the self-report data clustering presents a challenge. It may be advantageous to consider alternate ways of identifying periods of homelessness - either with a timeline follow-back method or use of collateral sources. It may be possible to use some statistical techniques to smooth over responses, however, the difficulty of there being some likely highly accurate responses at the margins with much less accurate ones in the midrange makes this a complex problem to solve.

Despite these limitations and challenges, this approach has demonstrated a novel finding in this data of a type of homeless individual who may have an inadvertent entrenchment of homelessness with one of the treatments, and this can serve as information that clinicians can use to be judicious in their treatment decision making.
Chapter 6

Discussion

6.1 Introduction

This thesis aimed to develop a nonparametric, fully Bayesian approach to causal modelling. The method proposed was generated by starting with the basic principles of causal modelling, the principles of modelling using the Dirichlet process prior to model a full joint distribution, and then aligning the properties of both approaches to maintain the capacity to calculate an unbiased estimate of the average treatment effect. While others have started with a plan to implement a propensity score model, in this approach we chose to remain open to alternative ways to proceed while still adhering to the spirit of the counterfactual approach. However, it was also intended from the outset that rather than only reporting an average treatment effect, we also wish to develop a method that can also accurately predict conditional effects. The average treatment effect is most relevant to only some end users (for instance policymakers or program developers) where understanding and predicting the population level effect is the critical aspect to making a decision for funding or new program implementation. We believe that conditional estimates may be of more use in clinical care and therefore have particular relevance in biostatistics when working with clinicians who ultimately want information to guide treatment decisions for an individual level.

6.2 Summary of results

The proposed method produces average treatment effect results that are similar to the propensity score methods of stratification, covariance adjustment and inverse probability weighting. In addition, through this approach, we are able to recover the density estimate of the covariate, and three conditional curves that underlie the original generative causal model. Even in the situation of a known missing confounder, we were able to replicate the conditional densities with reasonable accuracy when a sufficient number of clusters were included in the model. While the method cannot correct for missing confounders in the estimation of the overall average treatment effect, the density visualizations and pointillism plots that are produced can provide clues of missed confounders. When this method was applied to a real data situation using only one confounder this type of analysis did point to this possibility. Also, in this real-life scenario, where a propensity score could not be implemented due to convergence issues, the DPMR method still provided estimates similar to the inverse probability weight, and covariance adjustment.
The assumption of no unmeasured confounders raises issues that contain philosophical questions regarding causation that lead to considerable debate. Dawid [8] introduces a metaphysical model which contains how $R_1$ and $R_0$ covary, which ultimately he argues is unanswerable. He goes on to suggest that a physical model is the only one that can provide direction uncontaminated by this uncertainty. He also suggests that every model that uses a counterfactual framework induces a covariance relationship between the counterfactuals by virtue of the set of assumptions that are additionally made in the fitting of the model. In our model we assume local constant means matched within clusters; however, we suspect that this creates a minimal structure on this covariance relationship. For instance, in our second simulation, it is certainly the case that our local counterfactual responses are treated as independent, but there is an underlying covariation in outcome responses by gender and by age. This method provides some insights, by capturing aspects of the difference by age, including modes for the gender response despite not being able to identify the source of this covariation. To Dawid’s point, how $R_1$ and $R_0$ covary within an individual may not be known, however, it would seem that once we condition on covariates (such as age), we begin to learn something about the joint responses conditionally on the covariates we do know. This approach is inducing a covariance structure in a highly flexible way. Exploring this covariance structure more closely in this context may be informative. As the frequently stated aphorism credited to George Box says, “All models are wrong, some are useful,” it may be that some models work better for different types of applications in this setting. It may be that by making a local assumption, we may be sidestepping some of the problems outlined by Dawid that seem to have created much controversy.

The economics literature addresses situations where both direct and indirect effects need to be accounted for; further, the concerns regarding the intrinsic error in estimating the propensity score led to the inclusion of frequentist clustering-based models to address this problem. This fully Bayesian approach may provide a novel way to address this same problem and provide some additional information regarding the joint distribution.

### 6.3 Limitations

This technique has only been developed on a single covariate confounder. However, in most datasets, there are several variables available, and work must be done to extend this work to multivariate confounders. However, the underlying framework is promising for identifying missing covariates and thus may still be suitable when minimal data is available for correction. The curse of dimensionality may present difficulties in situations where a researcher wishes to correct for an exhaustive list of confounders. However, current methods may already excel in these types of problems.

A potentially critically important aspect of this approach is determining the sample sizes needed for this method to be effective. The properties of its convergence still need to be researched and established across a variety of potential situations (for example with a higher number of modes of treatment and with small subsets of either low or high propensity for one treatment over another) and types of outcomes (continuous, count or categorical). Similarly, a larger number of applications and simulations are needed to determine which complex situations arise where the method performs poorly.
6.4 Strengths/Importance

The method is flexible with regards to types of outcomes and in the conditional space, with enough clusters, appears to capture features of the joint density that are useful for prediction, while maintaining an unbiased estimate of the ATE. The ATE may be accurately estimated with fewer clusters than the number needed for robust conditional estimates; hence the number of clusters can be calibrated to the primary task of the analyst.

The identification of conditional modes and densities allow for a visualization that can help identify the presence of key missing confounders. The conditional density allows for a prediction that includes the uncertainty in response, which may be more useful for patients and clinicians in actual decision making situations where understanding the possible and probable outcomes may be useful for treatment decisions. Since the joint density sits as the central result from this method, it is possible to derive and report on any functional of this joint density. While the mean and mode are two common results of interest, in different problems other functionals may be the most relevant.

Furthermore, the use of a Bayesian method allows for easy integration with a loss function for policy makers and clinicians to weight possible responses by the costs (financial, quality of life or otherwise). Also, given that many treatments come with side effects and these effects may be persuasive factors in treatment decisions, one can create models with multiple outcomes all plotted by confounders. This method may allow individuals to have a more explicit prediction of how they may fare both in benefits and adverse consequences as was demonstrated in the clinical application. When these consequences have clear financial impacts, the current method produces posterior draw information that could be used to create appropriate cost calculations based on appropriate loss functions while preserving the capacity to estimate the credible regions of these additional estimates.

6.5 Practical applications of work

Small, non-randomized studies often form the basis of early research into promising or new treatments. This method could both help balance the effects by adjusting for observed confounders but may also provide clues as to the need to identify additional biases in the treatment assignment. Small studies are essential to the development of support for innovative methods; especially since the costs associated with the more rigorous randomized controlled trials are often cost-prohibitive. Thus researchers are often unable to secure funding for randomized studies when the treatment under study is very new. This financial barrier may also be a problem for studies that repurpose previously approved medications or interventions as there may be little financial incentive for business investment in this research. These financial disincentives are also frequently the case with orphaned diseases, i.e. diseases that affect a very small minority of patients. Often these small trials are not only non-randomized but are also conducted with little to no funding and hence often can only collect a small number of covariates, making it harder to justify the no-unmeasured confounders assumption.

It is certainly still the case, with our estimate as with others, that no one can estimate the average treatment effect in an unbiased way when the assumption of no unmeasured confounders is violated. However, since the density estimate seems unperturbed under our new model (creating bimodal or potentially dispersed density estimate when it needs to be), it would seem that this proposed approach could provide a reasonable estimation method under the current known scientific understanding of the
counterfactuals with the already identified confounders.

Given that the identification of confounders is also a research output in its own right, it is somewhat difficult to justify an assumption of no unmeasured confounders, especially in the early phases of research into an area. As research into important exposures advances, we are better able to recognize which factors may be influencing both treatment assignment and outcomes. Having to wait until a perfect understanding of the categorization of potential confounders into relevant and irrelevant before a model can be trusted seems to be a high bar to set. The current alternative strategy to this method has been to include all available covariate ("everything and the kitchen sink"), even when researchers doubt (or cannot justify) the actual influence of a particular included covariate. As a point of irony, we chose to use age and gender, covariates which are frequently included in all models even when there are no theoretical or previous results suggesting the relevance of their inclusion.

In the case of providing care and choosing between options, in the standard propensity score models we find results that can help a policy maker choose between making some treatments more readily available. However, this is not the most relevant question for clinicians. With clinicians, the variability in response represents the uncertainty in outcomes that a patient will face following the treatment decision and needs to be directly addressed in a model for it to be useful. When a patient selects their treatment we (and they) would like the probability of a good outcome to be maximized and it may be that in some situations, while the average treatment effect may seem to demonstrate ‘superiority’ of one treatment, this does not guarantee that in the conditional space it is always the superior treatment. Patient’s questions about probable outcomes under different treatment options are always intended in the conditional meaning; that is, “How will this treatment work for me?” (or as we would read it, people like me). This approach to modelling and reporting conditional effects thus provides a way to personalize treatment choices. This principle has found much support recently both in genetic studies (picking psychiatric treatments based on the genetic prediction of side effects, metabolism and thus likely treatment effects) and in treatment algorithms based on observed disease factors. The more widespread use of calculators and the possibility of smartphone and EMR (Electronic Medical Records) integration to assist clinicians in treatment decisions make this approach timely. Whereas in the past KTE involved reporting the average response by treatment group, the dissemination of an algorithm, formula or conditional plots may be a more powerful way to influence uptake of treatment advances.

The implications of using this new approach are multi-faceted. It would suggest that including a smaller set of known confirmed confounders may provide us with conditional models that are still useful, and easier to implement for a clinician with limited time to ask and record every possible confounder, despite its minimal contribution to overall treatment response. If this minimal set remained useful, it would imply that the ‘curse of dimensionality’ is less important than previously considered. This curse has been driven by the “everything and the kitchen sink” approach to variable selection. Even in the absence of all a model with all confounders, it may be that by having clinicians inspect and compare predicted outcome curves from these conditional models in clinical situations as described above, clinicians may begin to recognize the factors that distinguish which patients have responses from various parts of the predicted response curves. For instance, imagining for the moment that the second simulation represents a true causal model, it might be that once disseminated widely, clinicians begin to recognize the modes in response correspond to men or women. If this had a been an extra dispersed density, it may be that clinicians while using a tool informed by this method, might recognize a continuously-valued covariate (severity of illness) or count-valued covariate (number of relatives) that
helps distinguish which part of a curve responses cluster at, thereby informing the next cycle of clinical research.

6.6 Future work

More work is needed to extend the current work from a univariate confounder to multi-dimensional covariate spaces. Also, mixtures of confounder types (classes, continuous measures and frequency counts) are extremely common, and methods that can simplify the covariate space are needed so that the method can still work in small data settings. Current covariate selection tools and dimensionality reduction strategies could be accomplished by applying tools already developed and in use with Dirichlet process mixtures. Modelling outcomes and covariates from other (non-normal) distributions in simulation studies to establish the performance of this method would also be a beneficial extension, as other methods are noted for their bias in the generalized setting [2]. Given the Dirichlet process samples from a dense set of densities, and this method does not require the separate calculation and inclusion of a propensity score, it is entirely possible that this method, when used on generalized outcomes may prove to have desirable convergence properties. Additionally, extending the method to include a mixture of covariates that are not confounders (i.e. do not contribute to the propensity score) would be helpful. Simulation studies with such a mixture of covariates and confounders would allow for determining the impact of including these variables in the analysis.

Dirichlet process priors and Dirichlet process mixture regression are currently being applied to ‘big data’ in many settings. The findings and developments in those areas from this recent expansion of techniques are therefore available to this causal modelling approach. This flourishing area can provide a rich set of possible analytic approaches to, and integrations with, the high-dimensional data commonly available from observational studies. To capitalize on these extensions a more fulsome review of DPMR models and careful consideration of if, and how, they could be used would need to be further explored.

While we have a conceptual system to justify a weak ignorability assumption, this assumption has not been the norm in the Bayesian or frequentist models advanced to date. Work is needed to more deeply understand the consequences of this assumption and the relationship with the more standard assumption of strong ignorability. The one other exception to the strongly ignorability assumption was found in the work of Hoshino [25]. They used an alternate assumption that included the value of one counterfactual to inform the other; perhaps connections can be drawn by determining what differing impacts and consequences of these assumptions. While we believe that the assumption we begin with is weaker than the general assumption (and hence broader and more likely to be true), others will need to be convinced of this.

The simplified pointillism plotting approach we proffer in this thesis as capturing aspects of the density and its modes can allow for faster production of visualizations of data, however further developments are needed to understand which situations may lead to erroneous interpretation. Further, additional adjustments to the approach can be developed. For instance, creating elliptically shaped “points” that capture the variation in each dimension (or inclusion of colour coding to denote the probability of class membership) could provide an opportunity to improve the effectiveness and perceived accuracy of these visualizations. Our pointillism plots could also be adjusted by the observed class membership probabilities (the proportion assigned to the class) rather than the predicted class membership probability \( p_j \) if a prediction closer to the observed data is desired. The properties of these alternate visualizations need
further exploration.

Finally, further testing of the approach on real datasets and speaking to clinicians about the utility of the outputs is critical. While we propose that this approach could be of use in clinical situations, such use would require a different method of dissemination, and this would need to be practical, available and acceptable to clinicians. Working closely with the end-users in such a development would be essential to ensure wider dissemination.
Appendices

Appendix 1

Example BUGS code

The core BUGS code for the observed data (and counterfactuals) is included here plus an example of the hierarchical code used for the treatment assignment. The other hierarchical hyperparameters follow a similar structure as the treatment assignment.

# The joint model is built from the following local (cluster) conditional models
#
# X ~ Normal (MU_X,TAU_X)
# Z (TREATMENT) | X ~ bernoulli (pi_TREATMENT)
# R1,R0 | X, Z ~ Normal (mu_R0/mu_R1, tau_R0/tau_R1)
# Y <- R1, R0 & TREATMENT

model {
  for (i in 1:N) {
    muX[i]<-muX.group[H[i]]
    x[i]~dnorm(muX[i],tauX)
    logit(piT[i])<-muT.group[H[i]]
    z[i]~dbern(piT[i])
    muR0[i]<-muR0.group[H[i]]
    muR1[i]<-muR1.group[H[i]]
    muY[i]<-muR0[i]*(1-z[i])+muR1[i]*z[i]
    y[i]~dnorm(muY[i],tauY)
    H[i]~dcat(p[i])
    for (j in 1:C) {
      SC[i,j]<- equals(j,H[i])
    }
  }
}

# Alpha prior for the Dirichlet process prior
alpha~dunif(0.1,20)

# Constructive truncated prior for the p[] of the Dirichlet process prior
V[C]<-1
for (j in 2:C) {
  p[j]<-V[j]*(1-V[j-1])*p[j-1]/V[j-1]
}
for (k in 1:C-1) {
  V[k]~dbeta(1,1)
}

# Specifying the hyperparameters for the covariate
tauX<-1/(sdX*sdX)
sdX~dunif(sdX_L,sdX_U)
tauX0<-1/(sdX0*sdX0)
sdX0~dunif(sdX0_L,sdX0_U)
muX0~dnorm(muX00,tauX00)
for (c in 1:C) {
  muX.group[c] ~ dnorm(muX0,tauX0)
}

# Specifying the hyperparameters for the probability of treatment assignment
muT0~dnorm(muT00,tauT00)
taupiT <- 1/(sdpiT*sdpiT)
sdpiT~dunif(sdpiT_L,sdpiT_U)
for (c in 1:C) {
  muT.group[c] ~ dnorm(muT0,taupiT)
}

# Specifying the hyperparameters for the outcome and counterfactuals
muR00~dnorm(muR000,tauR0)
muR10~dnorm(muR100,tauR0)
tauR <- 1/(sdR*sdR)
sdR~dunif(sdR_L,sdR_U)
tauY <- 1/(sdY*sdY)
sdY~dunif(sdY_L,sdY_U)
for (c in 1:C) {
  muR0.group[c] ~ dnorm(muR00,tauR)
  muR1.group[c] ~ dnorm(muR10,tauR)
}

# Calculating total clusters used (with non-zero membership)
Cstar<- sum(CL[])
for (j in 1:C) {
    CL[j]<-step(sum(SC[,j])-1)
}
}
Appendix 2

Example R code to call OpenBUGS

Data, parameters, hyperparameters and MCMC implementation factors are all coded in R, through the use of the "R2OpenBUGS" package. This package allows for easy integration with projects that are already in R, and the code below can be adapted for different problems and changed to model a variety of cluster sizes. The following is an example of code that was used for the first simulation study with C=10 clusters.

```
library(boot)
library(R2OpenBUGS)
setwd('C:/Users/Tim/OneDrive/Data/Thesis Simulations/Geri10')
GeriatricData<-read.csv("../CommonSimulationFiles/GenGerPropenBaseDataFeb2014.csv")
attach(GeriatricData)

# First I determine how many clusters I wish to have in the model, and the number of # datapoints
C<-10
N<-length(x)

# Setting bounds on the ball size for X:
# The principle here is always the same, if one cluster exists (that is there is no
# distinction over x in response and/or treatment assignment) then the data will be
# within 2 standard deviations of the mean. Here I assume that will be midrange,
# and so the range / 4 will give an upper bound for the size of the ball.
# Age generally spans 60 to 86 = 26 units/4=6.5, so the lower bound needed (imagining
# a string of C pearls each with 4*sdx_lower to span the range of X

Xrange<-26  # could also make this data driven, and use an empirical Bayes appoach
sdX_L<Xrange/(4*C)
tauX_U<-1/sdX_L^2
sdX_U<Xrange/4
```
\[ \tau_{X_L} = \frac{1}{\text{sd}_{X_U}^2} \]

# Equivalently for the hyperparameter, either there is one grand mean and the sd between them is 0 (very small) or there is a chain of \( C \) pearls and the means of the balls span almost the entire \( x \) space, so then we can use the upper bound described above (range/4) to bound the hyperparameter.

# Here I assume the researcher has told me that they expect an average age of 73

\[ \mu_{X00} = 73 \]
\[ \tau_{X00} = \frac{1}{\text{sd}_{X_U}^2} \]
\[ \text{sd}_X_{0_L} = 0 \]
\[ \text{sd}_X_{0_U} = \text{sd}_{X_U} \]

# Here I assume the mean treatment probability will be 0.5 (this may not be reasonable for some problems), and I do not want any group membership to fall below one observation. So if everyone (\( N \) participants) were divided evenly into \( C \) groups, \( N/C \) per group, and looking to have at least one in treatment 1 or at least one in treatment 0, then minimally \( 1/(N/c) \) and most \( (N/C-1)/(N/C) \). So I take 3 standard deviations of the logit\((C/N)=-3.9\) or \( \text{sd}=1.3 \) is the max

\[ \mu_{T00} = 0 \]
\[ \text{sd}_{piT_U} = \text{abs}(\text{logit}(C/N))/3 \]
\[ \text{sd}_{piT_L} = 0 \]
\[ \tau_{T00} = \frac{1}{\text{sd}_{piT_U}^2} \]

# To estimate the outcomes, again I know the range is between -0.5 to 1.5 spanning about 2 units, This gives bounds as follows:

\[ \text{Yrange} = -2 \]
\[ \text{sd}_Y_{U} = \text{Yrange}/4 \]
\[ \text{sd}_Y_{L} = \text{Yrange}/(4*C) \]

# For the hyperparameters I assume a midpoint at the midrange of the scale 0.5

\[ \mu_{R000} = 0.5 \]
\[ \mu_{R100} = \mu_{R000} \]
\[ \text{sd}_R_{L} = 0 \]
\[ \text{sd}_R_{U} = \text{sd}_Y_{U} \]
\[ \tau_{R0} = \frac{1}{(\text{sd}_R_{U})^2} \]

# The three data elements needed are \( x,y, \) and \( z \), however, must also pass in sample size \( (N) \) and maximum number of clusters \( (C) \), along with fixed hyperparameters: \( \text{sd}_R_{L}, \text{sd}_R_{U}, \mu_{R000}, \mu_{R100}, \tau_{R0}, \text{sd}_{piT_L}, \)
Chapter 6. Discussion

# sdpiT_U, muT00, tauT00, sdX0_L, sdX0_U, sdX_L, sdX_U, sdY_L, sdY_U

dataIn<-list("x", "y", "z", "N", "C", "sdR_L", "sdR_U", "muR000", "muR100", "tauR0",
        "sdpiT_L", "sdpiT_U", "muTO0", "tauTO0", "sdXO_L", "sdXO_U",
        "sdX_L", "sdX_U", "sdY_L", "sdY_U", "muX00", "tauX00")

# The other random components needing initialization are:
# H[] <- length N, categories (1:C)
# alpha <- alpha is likely to be between 1/N and N^2
# V[] <- length C, last one is 1, others are dbeta (1, alpha)
# sdY~dunif(sdY_L,sdY_U)
# sdR~dunif(sdR_L,sdR_U)
# muR00~dnorm(muR000,tauR0)
# muR10~dnorm(muR100,tauR0)
# muR0.group[c] ~ dnorm(muR0,tauR)
# muR1.group[c] ~ dnorm(muR1,tauR)
# sdX~dunif(sdX_L,sdX_U)
# sdX0~dunif(sdX0_L,sdX0_U)
# muX0~dnorm(muX00,tauX00)
# muX.group[c] ~ dnorm(muX0,tauX0)
# muT0~dnorm(muT00,tauTO0)
# sdpiT~dunif(sdpiT_L,sdpiT_U)
# muT.group[c] ~ dnorm(muT0,tauT0)

inits<-function(){
    alpha<-5
    V<-c(rbeta(C-1,1,alpha),NA)
    # If I need to create multinomials - will have to change
    # from the matrix to a vector
    # p<-rep(V[1],C)
    # for (j in 2:C) {
    #     p[j]<-V[j]*(1-V[j-1])*p[j-1]/V[j-1]
    # }
    p<-c(0.35,0.25,0.15,0.13,0.12)
    Htemp<-rmultinom(1,N,p)
    H<-rep(1,N)
    H<-sample(c(rep(1,Htemp[1]),rep(2,Htemp[2]),rep(3,Htemp[3]),
        rep(4,Htemp[4]),rep(5,Htemp[5])))
    # I choose a single reasonable value between (0.05,0.5)
    # by narrowing the range
    sdY<-runif(1,0.2,0.3)
    # similarly for sdR from (0,0.5)
    sdR<-runif(1,0.1,0.3)
    muR00<-rnorm(1,muR000,sd=sqrt(1/tauR0)/2)
muR10 <- rnorm(1, muR100, sd = sqrt(1/tauR0)/2)
muR0.group <- c(by(y, list(H, z), mean, rm.na = T)[1:5], rnorm(C-5, muR00, sdR))
muR1.group <- c(by(y, list(H, z), mean, rm.na = T)[6:10], rnorm(C-5, muR10, sdR))

# now I initialize the probability parameters and logits
muT0 <- rnorm(1, muT00, sd = sqrt(1/tauT00)/2)

# I choose a single reasonable value between (0,1.3) by narrowing the range
sdpiT <- runif(1, 0.4, 0.9)
muT.group <- c(logit(by(z, H, mean)), rnorm(C-5, muT0, sdpiT))

# now the same for the housing binomial
muX0 <- rnorm(1, muX00, sd = sqrt(1/tauX00)/2)

# I choose a single reasonable value between (0.5,6.5) by narrowing the range
sdX <- runif(1, 2.5, 4.5)
# I choose a single reasonable value between (0,6.5) by narrowing the range
sdX0 <- runif(1, 2, 4.5)
muX.group <- c(by(x, H, mean, rm.na = T), rnorm(C-5, muX0, sd = sdX0))

list(H = H, alpha = alpha, V = V, sdY = sdY, sdR = sdR, muR00 = muR00, 
    muR10 = muR10, 
    muR0.group = muR0.group, 
    muR1.group = muR1.group, 
    muT0 = muT0, 
    sdpiT = sdpiT, 
    muT.group = muT.group, 
    muX0 = muX0, 
    sdX = sdX, 
    sdX0 = sdX0, 
    muX.group = muX.group)

# Now I list the parameters that I would like to follow
parameters <- c("p", "muR0.group", "muR1.group", "muR00", "muR10", "sdY", 
    "sdR", 
    "muT0", "muT.group", "sdpiT", "sdX", "sdX0", "muX.group", "muX0", 
    "Cstar", "alpha")

numChains <- 3
numIters <- 5000
numBurn <- 1000

Geri10.sim <- bugs(dataIn, inits, parameters, model.file = "modelPropensity.txt", 
    n.chains = numChains, n.iter = numIters, n.burnin = numBurn, 
    n.thin = 1, debug = FALSE)

save.image("GeriSimOutput10.Rdata")
Bibliography


