University of Crete
Department of Mathematics and Applied Mathematics

Bayesian Causal Feature Selection from observational and limited experimental data.

Author: Konstantina Lelova

Advisor: Sofia Triantafillou

Submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Mathematics and Applied Mathematics of the
University of Crete.

July 2023

Supervisor:

Sofia Triantafillou

Examination committee:

Sofia Triantafillou

Ioannis Tsamardinos

Ioannis Kamarianakis

Heraklion, July 2023

Acknowledgements

This thesis would not have been possible without the enduring support of my supervisor, Sofia Triantafillou. Her ability to bridge mathematical research with real-world challenges has played a pivotal role in shaping my understanding of how mathematics and engineering can be combined to create meaningful social impact. I also want to thank Ioannis Tsamardinos for introducing me to the world of machine learning and mathematics and for his willingness to listen and assist. I really appreciate his enthusiastic approach to constantly pushing the boundaries. This work was supported by a master's scholarship from the Institute of Applied & Computational Mathematics (IACM) in the Foundation for Research and Technology Hellas (FORTH).

I would like to thank Alexandros Koskinas for believing in me and supporting all my decisions without considering spatial constraints. Also my friends for standing with me and making me believe that there is no boundary that we cannot overcome together. Special thanks go to Katerina, Eleni, and Anastasia for their unstoppable support during my stay in Heraklion. And last but not least thanks to my family for always being there for me.

Contents

List of Figures

List of Algorithms

Abstract

In medical research, the selection of variables that contribute to an optimal predictive model or aid in uncovering associations between treatment, outcome, and pre-treatment variables poses a paramount goal. However, one of the most crucial challenges faced by doctors is the selection of treatments that will optimize individual patient outcomes. This objective can be effectively addressed by framing it as the problem of feature selection for predicting post-intervention outcomes using pre-intervention variables.

Experimental data from randomized controlled trials allow for unbiased estimation of the probability of post-treatment outcomes. However, such data have limited sample sizes and may be underpowered to accurately estimate conditional effects. Observational data contain many more samples but in most realistic cases, the presence of confounding variables makes it difficult to establish causal relationships. Thus, identifying a set of appropriate covariates and adjusting for their influence to mitigate confounding bias is not always possible from the observational data alone.

This thesis argues that the combination of experimental and observational data may help to improve the prediction of the post-intervention outcome and lead to an unbiased conditional treatment effect estimation.

We propose a Bayesian feature selection method for finding the Markov boundary from the observational data and using the concepts of feature selection, Bayesian inference, and Bayesian regression, we extend a recently proposed method that combines large observational and limited experimental data to identify adjustment sets and improve the estimation of causal effects for a target population. [40] This method was developed for multinomial distributions with Dirichlet priors and closed-form solutions and we present its extension for data sets with both binary and continuous explanatory variables when the outcome is binary or ordinal. In healthcare settings, the ordinal data is of great importance as it allows for the nuanced measurement of patient outcomes and a significant gap exists in effective methods for predicting post-interventional outcomes in this case.

We test our method in a simulated data set under different conditions. Results indicate that our method (a) demonstrates high performance in accurately identifying the correct Markov boundary for both binary and ordinal cases, even when applied to small observational data sets, (b) exhibits strong performance in identifying the optimal set Z that when included in a model, yields the best prediction for the post-intervention outcome P(Y |do(X), Z). The experiments were conducted using limited experimental and large observational data samples, respectively. When dealing with ordinal data, it is essential to have a larger set of experimental data compared to the binary case.

Περίληψη

Στην ιατρική έρευνα, η επιλογή μεταβλητών που συνεισφέρουν σε ένα βέλτιστο προβλεπτικό μοντέλο ή βοηθούν στην αποκάλυψη συσχετίσεων μεταξύ της θεραπείας, των αποτελεσμάτων και των μεταβλητών προ-θερα αποτελεί έναν πολύ κρίσιμο στόχο. Ωστόσο, ένα από τα πιο σημαντικά εμπόδια που αντιμετωπίζουν οι γιατροί είναι η επιλογή θεραπειών που θα βελτιστοποιήσουν το αποτέλεσμα για κάθε ασθενή ατομικά. Η προσέγγιση του προβήματος αυτού μπορεί να γίνει αποτελεσματικά, ανάγοντας το, στο πρόβλημα επιλογής χαρακτηριστικών για την πρόβλεψη των αποτελεσμάτων μετά την παρέμβαση χρησιμοποιώντας τις προ-παρέμβασης μεταβλητές.

Τα πειραματικά δεδομένα από τυχαιοποιημένες ελεγχόμενες δοκιμές επιτρέπουν τον αμερόληπτο υπολογισμό της πιθανότητας του αποτελέσματος αφού έχει χωρηγηθεί θεραπεία. Ωστόσο, αυτά τα δεδομένα έχουν περιορισμένα μεγέθη δείγματος και μπορεί να υπολειτουργούν για την ακριβή εκτίμηση των υπό συνθήκη επιδράσεων. Τα παρατηρησιακά δεδομένα περιέχουν πολλά περισσότερα δείγματα, αλλά στις πιο ρεαλιστικές περιπτώσεις, η παρουσία συγχυτικών μεταβλητών καθιστά δύσκολη τη δημιουργία αιτιωδών σχέσεων. Έτσι, ο προσδιορισμός ενός συνόλου κατάλληλων συμμεταβλητών και η προσαρμογή για την επιρροή τους για να μετριάσουν την συγχητική προκατάληψη δεν είναι πάντα δυνατή μόνο από τα παρατηρησιακά δεδομένα.

Η θέση αυτή υποστηρίζει ότι ο συνδυασμός των πειραματικών και παρατηρησιακών δεδομένων μπορεί να βοηθήσει στη βελτίωση της πρόβλεψης της έκβασης μετά την παρέμβαση και να οδηγήσει σε μια αμερόληπτη εκτίμηση της επίδρασης υπό όρους θεραπείας.

Προτείνουμε μια Μπεϋζιανή μέθοδο επιλογής χαρακτηριστικών για την εύρεση του Μαρκοβιανού ορίου από τα παρατηρησιακά δεδομένα και χρησιμοποιώντας τις έννοιες της επιλογής χαρακτηριστικών, της Μπεϋζιανής συμπερασματολογίας και της Μπεϋζιανής παλινδρόμησης, προτείνουμε ένα πλαίσιο που συνδυάζει μεγάλα παρατηρησιακά και περιορισμένα πειραματικά δεδομένα για τον προσδιορισμό των συνόλων *προσαρμογής* (adjustment) και τη βελτίωση της εκτίμησης των αιτιωδών επιπτώσεων για έναν πληθυσμό-στόχο. Η μέθοδος μας είναι έγκυρη για σύνολα δεδομένων με δυαδικές και συνεχείς ανεξάρτητες μεταβλητές όταν το αποτέλεσμα είναι δυαδικό ή τακτικό. Στα περιβάλλοντα υγειονομικής περίθαλψης, τα τακτικά δεδομένα είναι μεγάλης σημασίας, καθώς επιτρέπουν τη λεπτή μέτρηση των αποτελεσμάτων των ασθενών και υπάρχει σημαντικό χάσμα στις αποτελεσματικές μεθόδους για την πρόβλεψη των μετεπεμβατικών αποτελεσμάτων σε αυτήν την περίπτωση.

Δοκιμάζουμε τη μέθοδό μας σε ένα σύνολο δεδομένων που δημιουργήθηκαν μέσω προσομοιώσεων υπό διαφορετικές συνθήκες. Τα αποτελέσματα δείχνουν ότι η μέθοδος μας (α) δείχνει υψηλή απόδοση στον ακριβή προσδιορισμό του σωστού Markov ορίου τόσο για δυαδικές όσο και για τακτικές περιπτώσεις, ακόμη και όταν εφαρμόζεται σε μικρά σύνολα δεδομένων παρατήρησης, (β) παρουσιάζει ισχυρή απόδοση στον προσδιορισμό του βέλτιστου συνόλου

Z που όταν περιλαμβάνεται σε ένα μοντέλο, αποδίδει την καλύτερη πρόβλεψη για το αποτέλεσμα μετά την παρέμβαση P(Y) |do(X), Z). Τα πειράματα πραγματοποιήθηκαν με τη χρήση περιορισμένων πειραματικών δειγμάτων και μεγάλων δειγμάτων παρατηρησιακών δεδομένων αντιστοίχως. Όταν ασχολούμαστε με τα τακτικά δεδομένα, είναι σημαντικό να έχουμε ένα μεγαλύτερο σύνολο πειραματικών δεδομένων σε σύγκριση με τη δυαδική περίπτωση.

Chapter 1

Introduction

We present a causal feature selection method that combines observational and experimental data to identify causal factors influencing a target variable. The need for Bayesian causal feature selection arises from the challenges posed by both observational and experimental data when seeking to identify the proper causal factors influencing a target variable. Our motivation stems from healthcare settings, where the primary objective is often to choose treatments that result in the best possible outcomes for each patient. However, when the outcome variable is ordinal, a significant gap exists in effective methods for predicting post-interventional outcomes. This introduction serves to formalize the problem of Bayesian Causal Feature Selection, present a motivating scenario and outline the primary questions that will be addressed in this thesis.

Our contributions are the following:

- In **Feature Selection Methods** is that we can find Markov boundary from observational data with Bayesian regression methods.
- In **Causal Feature Selection** is that we present a Bayesian method for identifying the optimal set Z that when included in a model, yields the best prediction for the post-intervention outcome P(Y |do(X), Z), using observational and experimental data when it is possible, in two cases:
  - For both continuous and binary explanatory variables when the outcome is a **binary** variable,
  - and for both continuous and binary explanatory variables when the outcome is an **ordinal** variable.

In recent years, the fields of feature selection, Bayesian inference, and Bayesian regression methods have garnered significant attention and have proven to be powerful tools in various domains of data analysis and decision-making. Feature selection plays a crucial role in identifying relevant features from high-dimensional datasets, improving model performance, and facilitating interpretability. On the other hand, Bayesian inference provides a robust framework for incorporating prior knowledge and updating beliefs based on observed data, leading to more accurate and reliable statistical inference. Bayesian regression methods, within this framework, en-

able flexible modeling of complex relationships between variables while accounting for uncertainty.

The need for Bayesian causal feature selection arises from the challenges posed by both observational and experimental data when seeking to identify the true causal factors influencing a target variable. Traditional feature selection methods often fail to account for the complex interplay between variables, leading to biased or unreliable results. Bayesian causal feature selection provides a principled approach to address these challenges and extract meaningful causal relationships from the data.

In observational data, the presence of confounding variables makes it difficult to establish causal relationships. Observational studies are inherently susceptible to selection biases and unmeasured confounders, which can result in spurious correlations and misleading feature selection outcomes. Identifying a set of appropriate covariates (adjustment set) and adjusting for their influence can remove confounding bias; however, such a set is often not identifiable from observational data alone.

On the other hand, experimental data provides an opportunity for unbiased estimation of causal effects. However, due to their inherent limitation in sample size, present a challenge for traditional feature selection methods that commonly rely on statistical significance or predictive performance as selection criteria. As a result, the conditional effects that can be derived from the experimental data have high variance and may not be reliable.

These constraints necessitate the requirement for a framework that combines the large observational and the limited experimental data to identify adjustment sets and improve the estimation of causal effects for a target population.

In medical data research, accurately predicting the post-interventional outcome is crucial for assessing the effectiveness of interventions, guiding treatment decisions, and improving patient care. However, when the outcome variable is **ordinal**, a significant gap exists in effective methods for predicting post-interventional outcomes. This thesis aims to address this critical gap by developing techniques that specifically address the challenges associated with predicting post-interventional outcomes when the outcome variable is ordinal.

Our objective is to investigate and examine the ways in which we can explore how the integration of the three concepts of feature selection, Bayesian inference, and Bayesian regression methods, can establish a convincing way for estimating the post-interventional distribution of an outcome Y when intervening on a treatment X by finding the set of pre-treatment covariates which are maximal informative for the outcome using both observational and experimental data.

To conclude, this thesis is concerned with (a) finding the minimal set of features that lead to the optimal prediction of a target variable Y from the observational data, which is the Markov Boundary for the observational distribution, (b) selecting the minimal sets (if exists) of pre-treatment covariates that are maximally informative for the post-interventional distribution, P(Y| do(X)), from experimental and observational data, respectively, (c) understanding how combining observational and experimental data can improve feature selection and effect estimation and (d) extending the results of the work of *Triantafillou, Jabbari, Cooper in "Causal and Interventional Markov Boundaries"*[2021] [40] who present their results for multinomial distributions with closed-form marginals, to distributions with both continuous and binary explanatory variables in two cases: when the outcome variable is **binary** and **ordinal**.

1.1 Motivation

The motivation for this work comes primarily from clinical settings where you may have patients and you want to decide what's the best treatment for every patient.

Assume that you are working with doctors and clinicians who work in critical care and they have a lot of sepsis patients. One of their goals is to discover the association of some treatments with the outcome of interest and some other covariates that they have from the patient's history. But their optimal goal is to predict what the chance of in-hospital mortality is going to be for a new patient who comes in if they give them a specific treatment versus if they do not give them this treatment.
In the specific example of Figure 1.1, we want to answer the question: *Which variables should we include in a model to get the best prediction for in-hospital mortality give them steroids? (notation:**in-hospital mortality| do(Steroids))**.*

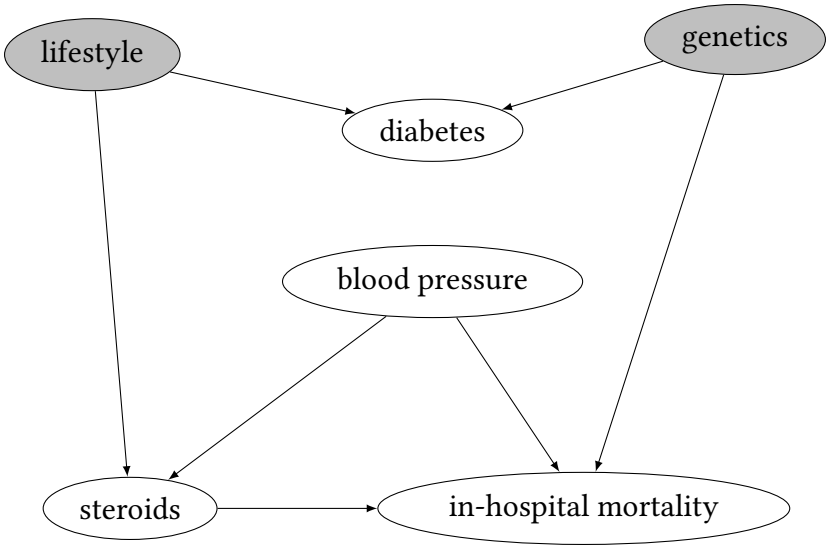


Figure 1.1: Hypothetical causal structure
Genetics and lifestyle are unobserved

Assume that Figure 1.1 describes the causal relations of some variables related to in-hospital mortality, where an arrow$X \rightarrow Y$ denotes direct probabilistic causality: direct because nothing mediates this causal relation in the context of the variables present in the graph, and probabilistic because if you intervene and change the distribution of X, then the distribution of Y will also change. Variables that are in ellipsis with gray color are unmeasured. Several graphical criteria exist, some of them will be shown next, and algorithms that learn the best adjustment set when the causal graph is known. [39]
However, in general settings and in most real-world problems, we are not gonna know the graph, and despite that fact, we want to predict the post-intervention outcome.

Imagine, for example, the following scenario (**motivating scenario**):
Assume that the **causal graph is unknown** and the doctors of the hospital where you work have various data sources. The data that they have can be divided into two categories:

- In **observational data**, that is usually historical data of large cohorts of patients that

they have already treated, they have been given steroids or not given them steroids. This data usually tend to have large sample sizes and they measure a lot of covariates because it is the entire medical history of the patients. However, you can not always use them to estimate post-interventional distributions because of possible confounders.

- and in **experimental data**. Data from Randomized Control Trials(RCTs) that they have performed in the hospital. This data is produced by dividing a group of people into two groups and forcing the first group to take a treatment and the second group not to take this treatment. They are unbiased for causal effect estimation, so you can use them to predict the post-interventional outcome. Unfortunately, they have limited sample sizes with limited covariates and they may not have the entire history of that patient.
  Doctors in the hospital want to answer two types of questions. We create scenarios for these cases:

  - **First Scenario**: Assume that you are working with doctors and clinicians who work in critical care and they have a lot of sepsis patients. They have every patient's history and they want to know the minimal set of covariates that contain all the information needed to determine the probability distribution of in-hospital mortality. They just want for example to predict: whether a patient will die in the hospital given the information that you have from them including whether you have given them steroids or not. This is just an observational prediction of the outcome of Y, so we need to **find from the observational distribution, the minimal set of pre-treatment covariates that makes all other variables independent of in-hospital mortality** (this set is also known as Markov Boundary of the variable "in-hospital mortality").
  - **Second Scenario**: Assume that you are working with doctors and clinicians who work in critical care and they have a lot of sepsis patients. A **new sepsis patient** comes in and they want to predict what the chance of in-hospital mortality is going to be for that patient if they give her steroids versus if they do not give her steroids. So they want to predict the post-interventional distribution of the outcome for their patient given different treatments and given some other covariates. We assumed that we have observational data measuring the treatment, the outcome, and some covariates and experimental data measuring the treatment, the outcome, and some covariates but they have a much smaller number of samples. Those distributions describe the relationship between your variables. The doctors want to know what is the best way to do the prediction for their post-intervention outcome and we **view this as a feature selection problem**. So, what are the best covariates (**optimal adjustment set**) to include in this model for predicting their post-intervention outcome?

The first scenario describes the feature selection problem and the second scenario the causal feature selection problem that will be answered next, using observational and limited experimental data. In this thesis, these two scenarios will be investigated, and the results of Triantafillou's, Jabbari's, and Cooper's work in "Causal and Interventional Markov Boundaries"[2021][40] will be extended to distributions with both continuous and binary explanatory variables, when the outcome variable is binary and ordinal.

1.2 Related work

Feature Selection - Markov Boundary
Covariate adjustment plays a crucial role in estimating causal effects from observational data. Extensive research has been conducted in the fields of potential outcomes and causal graph analysis to determine the appropriate sets of covariates for effective adjustment. There are several sound and complete graphical criteria when the causal graph is known.[48] [33]. As we deal with finding the Markov boundary (MB) of a target variable as a feature selection problem when the causal graph is unknown, we refer only to related work that is associated with this concept.

At first, almost all existing MB discovery algorithms are designed under the assumption of causal sufficiency, which states that there are no latent common causes for two or more of the observed variables in data.

Frequentist approaches offer different methodologies to address causal insufficiency. Researchers have developed extensions to the popular PC (Peter-Clark) algorithm which is one of the Constraint-Based Methods, such as the FCI (Fast Causal Inference) algorithm and RFCI (Restructural Causal Model FCI) [49] algorithm. These extensions incorporate additional tests and techniques to handle latent variables and unobserved confounders, allowing for more accurate identification of the Markov boundary in the presence of causal insufficiency. Also, the GES (Greedy Equivalence Search) algorithm [4] is a well-known score-based method that explores the space of causal structures by iteratively adding, deleting, or reversing edges to maximize a scoring criterion such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC).

Bayesian approaches offer different methodologies to address causal insufficiency, too. Bayesian Model Averaging (BMA) [7] and Bayesian Information Criterion (BIC) [18] can be used in the presence of causal insufficiency, but they do have limitations. But most recent works appear, that provide several algorithms (from sound and complete methods) which learn the Markov Boundary from data under causal insufficiency such algorithms from the work of K. Yu, L. Liu, J. Li, and H. Chen (M3B) [47] and the one proposed in Triantafillou's, Jabbari's, and Cooper's work in "Causal and Interventional Markov Boundaries" paper(FGESMB)[22] which has been the starting point for this project.

Our contribution to feature selection methods is that we can find the MB of a target variable using Bayesian regression methods and Markov Chain Monte Carlo(MCMC) sampling for inference. This method is an extension of the method presented for multinomial distributions which have closed-forms marginals in "Causal and Interventional Markov Boundaries" [40] and is valid for mixed datasets, with both continuous and binary variables. There are already methods that use the hierarchical Bayesian Model for Bayesian feature selection and make inferences with Gibbs sampling, but they are appropriate only for multivariate normal distributions with conjugate priors that their posterior can be calculated with closed-form marginals.[46]

Causal Feature Selection - Optimal Adjustment Set
Identifying causal effects is a critical task that has garnered significant attention in the research community. A vast body of literature has been dedicated to this endeavor, exploring various methods to determine optimal adjustment sets for accurate estimation of average treatment effects.

One line of work tries to select an adjustment set from observational data. From 2011 VanderWeele and Shpitser [44] proposed to control on a set of covariates that satisfy the "disjunctive set criterion", and their method guaranteed to adjust for a valid adjustment set, if one exists. This

method requires that we know which variables cause X and Y, while we make no such assumption. Henckel, Perkovic, and Maathuis [9] provide methods for selecting an optimal adjustment set for linear Gaussian data with no hidden confounders when we know all valid adjustment sets. They also provide a pruning method that takes as input a valid adjustment set, and returns a smaller valid adjustment set with lower asymptotic variance, if one exists. Smucler, Sapienza, and Rotnitzky (2020) [34] show that the results hold for broader types of distributions and there are also extends to DAGs with latent variables. Thus, in contrast to our method, this line of work assumes there is no uncertainty on whether a set Z is an adjustment set.

There are also several works on identifying causal effects when the causal graph is known. Given a graph (DAG/PDAG or SMCM), these methods apply a graphical adjustment criterion to identify a set of valid adjustment sets for estimating the average treatment effect of X on Y. Then, they try to identify the set that leads to the estimator with the smallest asymptotic variance among all the valid adjustment sets [25] [27] [34]. These methods are not directly comparable to ours since they focus on identifying average treatment effects while our method focuses on conditional effects and combines observational with experimental data when the graph is unknown. While these methods have different objectives compared to ours, they have some connections with our work, since for pre-treatment variables, they found the adjustment sets.

There is also a line of work that tries to combine experimental and observational data sets but with different settings than ours. For continuous data and linear relationships, observational and limited experimental data (2010)[6]. Kallus, Puli, and Shalit (2018) [12] propose a method for improving conditional interventional estimates. However, this method requires a binary treatment and continuous covariates and outcomes. Rosenman, Baiocchi, and Owen (2018) [26] propose combining RCT and observational data to improve causal effect estimates, based on some similar assumptions to ours. Nevertheless, this method assumes no hidden confounders. In 2020, the work of [45] observational and limited experimental data combined but they focus on the identifiability of causal effects and they assume no hidden confounders. There is also a lot of work on combining observational and experimental data on the basis of independence constraints [42] [19]. Still, these methods require larger experimental data sets to make meaningful inferences.

Our method tries to predict the post-intervention outcome using observational and limited experimental data using Bayesian regression methods and marginals. The important difference compared to other methods is that we assume nothing about the nature of the predictors and the response variable may be either binary or ordinal. We are trying to address the significant gap that exists in effective methods for predicting post-interventional outcomes when the outcome variable is ordinal. Also, we make no assumption about confounding and this method can be applied even with very few experimental data versus the existing constraint-based methods.

Chapter 2

Predicting binary and ordinal response variable using Bayesian Logistic Regression

2.1 Bayesian Inference

There are two main approaches to statistical machine learning: frequentist and Bayesian methods. In frequentist inference, probabilities are interpreted as long-run frequencies and the goal is to create procedures with long-run frequency guarantees. In this work, Bayesian inference is adopted, where probabilities are interpreted as subjective degrees of belief and the goal is to state and analyze these beliefs.

Let $X_1, ...X_n$ be n observations sampled from a probability density $p(x|\theta)$. The parameter θ is viewed as a random variable and $p(x|\theta)$ represents the conditional probability density of X conditioned on θ.

The Bayesian Inference procedure follows three main steps:

1. A probability density $\pi(\theta)$ is chosen- called the prior distribution - which expresses our beliefs about a parameter θ before seeing any data.
2. A statistical model $p(x|\theta)$ is chosen that reflects our beliefs about x given θ.
3. After observing Data $D_n = \{X_1, ..., X_n\}$, we update our beliefs and calculate the posterior distribution $p(\theta|D_n)$.

Theorem 2.1.1. ***Bayes Rule for Inference***
Bayes' theorem is stated as the following equation:

$$p(\theta|X_1, ..., X_n) = \frac{p(X_1, ..., X_n|\theta)\pi(\theta)}{p(X_1, ..., X_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta), \tag{2.1}$$

where $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(X_i|\theta)$ *is the likelihood function and*

$$c_n = p(X_1, ..., X_n) = \int p(X_1, ..., X_n|\theta)\pi(\theta)d\theta = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called the evidence.

In Bayesian probability theory, if the prior is a conjugate prior for the likelihood function, the posterior distribution can be calculated easily in closed form. Our contribution is that no conjugate priors were used in this work and the posterior distributions and marginals were calculated with sampling.[14]

2.2 Bayesian Logistic Regression for binary response

Bayesian logistic regression is a statistical model that combines logistic regression with Bayesian inference. It is used to model the relationship between a binary dependent variable and one or more independent variables while incorporating uncertainty and prior knowledge.

Logistic regression could be used to model, for example, the relationship between the binary outcome coronary heart disease status (yes or no) and age, adjusted for confounding due to other factors. As with linear regression, the model can be used to test the null hypothesis of no association, estimate the magnitude of the association and its 95% confidence interval, and predict the outcome at a given age. For logistic regression, the raw prediction from the model is on the log-odds scale but can be transformed to produce a predicted probability.

Assume N samples and each sample has a binary dependent variable Y and a set of k predictors, independent variables, $\mathbf{X} = (X_1, ...X_k)$. Assume, also, that Y_i is one of N response discrete variables that can only take two values, 0 or 1, the **Bernoulli distribution** is appropriate to describe this case.
If $\pi_i = P(Y_i = 1)$ denotes the probability of an event to occur, then:

$$Y_i|\pi_i \sim Bern(\pi_i)$$

As the logistic regression model is part of a broader class of generalized linear models, an appropriate link function of π_i, $g(.)$, that is linearly related to the predictors $X_{i1}, .., X_{ik}$ must be identified:

$$g(\pi_i) = b_0 + b_1X_{i1} + ... + b_kX_{ik}$$

The probability of an event to occur, π_i depends upon predictors $X_{i1}, ...X_{ik}$ through the **logit link function** $g(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$:

$$Y_i|(b_1, ...b_k) \sim Bern(\pi_i) \text{ with } log(\frac{\pi_i}{1-\pi_i}) = b_0 + b_1X_{i1} + ... + b_kX_{ik} \tag{2.2}$$

This assumes that the **log(odds of the outcome)** is linearly related to the predictors. To work on scales that are much easier to interpret, the relationship in terms of odds and probability can be rewritten following the properties of the log function as follows:

$$\frac{\pi_i}{1-\pi_i} = e^{b_0+b_1X_{i1}+...+b_kX_{ik}} \text{ and } \pi_i = \frac{e^{b_0+b_1X_{i1}+...+b_kX_{ik}}}{1 + b_0 + b_1X_{i1} + ... + b_kX_{ik}}$$

$$\pi_i = expit(b_0 + b_1X_{i1} + ... + b_kX_{ik}) \tag{2.3}$$

Note: $logit(p) = log(\frac{p}{1-p})$ and the inverse function: $expit(p) = \frac{e^p}{1+e^p}$.

Now, the relationships on the odds and probability scales are represented by nonlinear functions. These transformations preserve the properties of odds, which must be non-negative, and probability, which must be between 0 and 1.

In Bayesian logistic regression, the model incorporates prior distributions for the coefficients of the logistic regression model. These prior distributions reflect our beliefs or knowledge about the coefficients before observing the data. To estimate the coefficients, Bayesian inference updates the prior distributions using observed data. The posterior distribution, which represents our beliefs about the parameters given the data, is obtained by combining the prior distributions with the **likelihood** of the data.

The probability mass function of a Bernoulli distribution with π the probability of an event to occur is:

$$Bernoulli_{PMF}(Y, \pi) = \pi^{Y}(1-\pi)^{1-Y},\ for\ Y \in \{0, 1\}$$

Thus, for N samples, the likelihood of Y's will be observed when the true value of the parameters are $\{\pi_1, ...\pi_N\}$:

$$Likelihood = \prod_{i=1}^{N} [\ \pi_i^{Y_i}\ (1-\pi_i)^{(1-Y_i)}\ ] = \prod_{i=1}^{N} Bernoulli_{PMF}(Y_i, \pi_i) \tag{2.4}$$

$$LogLikelihood = \sum_{i=1}^{N} [log(Bernoulli_{PMF}(Y_i, \pi_i))] \tag{2.5}$$

The computation of the posterior distribution in Bayesian logistic regression typically involves techniques such as Markov Chain Monte Carlo (MCMC) or variational inference. These methods allow for sampling from the posterior distribution or approximating it to obtain the desired inference. In this work, MCMC is employed to sample from the posterior distribution of the regression model's parameters.

2.3 Bayesian Ordinal Logistic Regression for ordinal response

An ordinal variable is a categorical variable in which the levels have a natural ordering (e.g., depression is categorized as Minimal, Mild, Moderate, Moderately Severe, and Severe). Ordinal logistic regression can be used to assess the association between predictors and an ordinal outcome.
Proportional-odds cumulative logit model is possibly the most popular model for ordinal data and it will be used in this work. This model uses cumulative probabilities up to a threshold, thereby making the whole range of ordinal categories binary at that threshold.

- **Proportional-odds cumulative logit model**

Let the response be $Y = 1, 2, ..., J$ and $x = x_1, .., x_p$ the set of p independent variables. where the ordering is natural. The associated probabilities are $(\pi_1, \pi_2, ..., \pi_j)$, and a cumulative probability of response less than or equal to a specific category, j, is

$$P(Y \leqslant j) = \pi_1 + ... + \pi_j = \sum_{k=1}^{j} \pi_k \quad for \quad j = 1, ...J-1$$

The odds of being less than or equal to a particular category is $\frac{P(Y \leqslant j)}{P(Y > j)}$.

Then a **cumulative logit** is defined as:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right) \tag{2.6}$$

This is the log-odds of the event that $Y \leq j$ and measures how likely the response is to be in category j or below versus in a category higher than j.

The ordinal logistic regression model can be defined as:

$$logit(P(Y \leq j)) = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p,$$

where $\beta_{j0}, \beta_{j1}, \cdots + \beta_{jp}$ are model coefficient parameters (i.e., intercepts and slopes) with p predictors for$j = 1, \cdots, J - 1$.

The **Proportional Odds Assumption** required the intercepts to be different for each category but the slopes are constant across categories, which simplifies the equation above to:

$$logit(P(Y \leq j)) = \beta_{j0} + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.7}$$

, the intercepts are different for each category but the slopes are constant across categories.

So, the cumulative probabilities are given:

$$P(Y \leq j) = \frac{exp(\beta_{j0} + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + exp(\beta_{j0} + \beta_1 x_1 + \cdots + \beta_p x_p))} = expit(\beta_{j0} + \beta_1 x_1 + \cdots + \beta_p x_p) \tag{2.8}$$

We will provide a simple example that aids in understanding the intuition behind calculating the probabilities for each category:
Suppose we have N data samples and k=3 categories for the outcome Y. Every y_i will take the values (0, 1, 2) with probabilities ($p_{yi0}, p_{yi1}, p_{yi2}$). We assume the ordinal regression of the outcome Y on X and Z attributes. Then the probability of each category will be:

- $p_{yi0} = P(Y_i = 0) = P(Y_i \leq 0) = expit(a_{i0} + b_2 X_i + b_3 Z_i)$
- $p_{yi1} = P(Y_i = 1) = P(Y_i \leq 1) - P(Y_i \leq 0) = expit(a_{i1} + b_2 X_i + b_3 Z_i) - expit(a_{i0} + b_2 X_i + b_3 Z_i)$
- $p_{yi2} = P(Y_i = 2) = 1 - P(Y_i \leq 1) = 1 - expit(a_{i1} + b_2 X_i + b_3 Z_i)$

One tricky part of interpreting these cumulative logit model parameters lies in the fact that the cumulative probability applies to response **j or, less** . If a β coefficient is positive, it means that increasing its predictor is associated with an increase in the probability of j or less, which means that the response is more likely to be small if the predictor is large (the odds of the response being small is higher). This is why in most R and Python packages the ordinal logistic regression model parametrized as

$$logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - \cdots - \eta_p x_p$$

, where $\eta_i = -\beta_i$.
Now, positive b implies that higher values of x are associated with increased odds of being in a high-response category rather than a low-response category.[17] [10]

In this case, the **likelihood** function is based on the probability mass function of a **Categorical distribution**. Assume that π_k represents the probability of a sample to be in category k and $\sum_{k=1}^{j} \pi_k = 1$. Then the probability mass function of a Categorical distribution with probabilities $(\pi_1, ..., \pi_j)$ for each category is:

$$Categorical_{PMF}(Y, \pi_k) = \prod_{k=1}^{j} (\pi_k)^{[y=k]}$$

,where [y = k] evaluates to 1 if y=k, and 0 otherwise.
For data with sample size N, the **likelihood** will be:

$$Likelihood = \prod_{i=1}^{N} [\prod_{k=1}^{J}]((\pi_{yik})^{[y_i=k]}) = \prod_{i=1}^{N} Categorical_{PMF}(Y_i, \pi_{yi}) \tag{2.9}$$

$$LogLikelihood = \sum_{i=1}^{N} [log(Categorical_{PMF}(Y_i, \pi_{yi}))] \tag{2.10}$$

'The Bayesian framework for ordinal regression provides several advantages and is particularly relevant in healthcare research and decision-making.

Firstly, the flexibility of the Bayesian ordinal regression model allows for the effective modeling of ordinal outcomes. By considering the order and categories of the variables, the model can capture the shading relationships between independent variables and health outcomes. This is crucial in healthcare, where understanding the severity or progression of diseases, patient satisfaction, or functional limitations is essential.

One important benefit of Bayesian ordinal inference in healthcare is its ability to quantify uncertainty. In healthcare, there is often uncertainty due to factors like variability in the data, measurement errors, or missing information. Bayesian inference tackles this issue by providing posterior distributions that capture a range of likely values for the model parameters. These distributions help us understand the level of uncertainty associated with the results, enabling more informed decision-making in healthcare.

Lastly, healthcare data frequently exhibit complex relationships among predictors. Bayesian ordinal regression can handle complex models that incorporate interactions, non-linear effects, and random effects, allowing for a comprehensive understanding of the factors influencing health outcomes. This flexibility helps uncover hidden relationships and provides a more accurate representation of the underlying mechanisms in healthcare.

2.4 Markov Chain Monte Carlo Sampling

Let $D_n = \{X_1, ..., X_n\}$ be the observed data. Suppose that $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$ with some prior distribition $\pi(\boldsymbol{\theta})$. In Bayesian probability theory, if the posterior distribution $P(\boldsymbol{\theta}|D_n)$ is in the same probability distribution family as the prior probability distribution $P(\boldsymbol{\theta})$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $P(D_n|\boldsymbol{\theta})$. A conjugate prior is an algebraic convenience, giving a closed-form

expression for the posterior; otherwise, numerical integration may be necessary. If we had conjugate priors, we could estimate the posterior probability distribution from the likelihood retrieved from the regression model and the priors of the parameters in closed form:

$$p(\boldsymbol{\theta}|D_n) \propto \mathcal{L}_n(\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Definition 2.4.1. *(Conjugate priors)*
A prior distribution is conjugate if it is closed under-sampling. That is, suppose that $\mathcal{P}$ is a family of prior distributions, and for each $'theta$, we have a distribution $p(\cdot|\theta) \in \mathcal{F}$ over a sample space $\mathcal{X}$. Then if the posterior

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta}$$

satisfies $p(\cdot|\theta) \in \mathcal{P}$, we say that family $\mathcal{P}$ is conjugate to the family of sampling distributions $\mathcal{F}$. For this to be a meaningful notion, the family $\mathcal{P}$ should be sufficiently restricted and is typically taken to be a specific parametric family.

In our study, **we do not assume conjugate priors**. Thus, we do not know the family of the posterior distribution.
The question now arises of how to extract inferences about one single parameter. The key is to find the marginal posterior density for the parameter of interest. Suppose we want to make inferences about θ_1. The marginal posterior for θ_1 is:

$$p(\theta_1|D_n) = \int \cdots \int p(\theta_1, ..., \theta_d|D_n)d\theta_2...d\theta_d \tag{2.11}$$

In practice, it might not be feasible to do this integral. In those instances, we need to draw randomly from the posterior. As mentioned in Section 2.2, Markov Chain Monte Carlo(MCMC) methods allow for sampling from the posterior distribution or approximating it to obtain the desired inference.

In this work, **Markov Chain Monte Carlo (MCMC) sampling** is utilized to sample from the posterior distribution, which represents our updated beliefs about the regression's model parameters given the observed data and prior knowledge, to calculate marginals. However, MCMC is not the main objective of this study. So in this section, we provide a concise introduction to the methods that will be employed later, emphasizing their relevance to the research at hand.

The key idea behind MCMC sampling is to construct a Markov chain that explores the parameter space and converges to the target posterior distribution. The Markov chain is built by iteratively generating a sequence of parameter values, where each value is dependent only on the previous value according to a transition rule. The key advantage of MCMC is that it allows us to sample from high-dimensional and complex distributions without requiring explicit knowledge of their functional form.

In the context of feature selection or finding the Markov boundary, MCMC sampling is used to obtain samples from the joint posterior distribution of the model parameters and the variable selection indicators. This allows us to estimate the posterior probabilities of including or excluding each feature in the model. The Markov chain explores different configurations of the selected features and converges to a stationary distribution that represents the posterior distribution of interest.

Two commonly used methods in Bayesian inference that will be applied in this analysis later, are Sequential Monte Carlo (SMC) and the No-U-Turn Sampler (NUTS) with gradient-based sampling:

Sequential Monte Carlo (SMC):
Sequential Monte Carlo is a Monte Carlo-based method that aims to approximate the posterior distribution by iteratively updating a set of weighted samples. SMC is particularly useful when dealing with models that involve dynamic or sequential data, where the posterior distribution may not be tractable.

Sequential Monte Carlo (SMC) is a method used for sampling from complex distributions, particularly in situations where the target distribution is high-dimensional or the posterior distribution is difficult to sample from directly. SMC algorithms are a broad family of sampling methods where a tempered target distribution is sampled and then the samples' plausibility weights are calculated and used to re-sample the tempered distribution until the temperance factor is 1. At this point, we have samples of the target distribution. The idea behind SMC is to avoid the problem of sampling from complex target probability density functions (PDFs) but sampling from a series of intermediate PDFs that converge to the target PDF and are easier to sample. The Metropolis-Hastings algorithm, implemented with a Metropolis kernel in the Python package that we will use, PyMC3, is one way to perform the sampling step within SMC. It proposes new parameter values based on the current particle values. These proposed values are then evaluated using the likelihood of the observed data and accepted or rejected based on the acceptance criterion. The acceptance criterion typically involves comparing the posterior probability of the proposed values with the posterior probability of the current particle values.

No-U-Turn Sampler (NUTS):
One such alternative is the No-U-Turn Sampler (NUTS), which is a variant of the Hamiltonian Monte Carlo (HMC) algorithm. NUTS uses the gradient information of the target distribution to propose and explore new samples in a more efficient manner. By leveraging the derivative information, NUTS can automatically tune its parameters, adaptively adjust the step size, and explore the posterior distribution more effectively. This makes NUTS particularly suitable for high-dimensional spaces and complex target distributions. Compared to methods like SMC, NUTS often provides faster convergence and more reliable estimates.

Thus, SMC with a Metropolis kernel and NUTS are both sampling methods commonly used in Bayesian inference. SMC, implemented with a Metropolis kernel, constructs a sequence of intermediate distributions and updates samples based on acceptance/rejection criteria. NUTS, on the other hand, utilizes gradient information to propose and explore new samples efficiently. The choice between these methods depends on the specific characteristics of the target distribution, the dimensionality of the problem, and the desired trade-off between computational efficiency and robustness.

In summary, MCMC sampling is a versatile and widely used technique for sampling from posterior distributions. It enables us to explore complex parameter spaces and estimate posterior probabilities in feature selection or causal inference problems. By iteratively generating samples from the posterior distribution, MCMC allows us to make statistical inferences and draw conclusions about the model parameters and selected features.

Chapter 3

Modeling Causality with Causal Graphical Models

3.1 Definition of causality

Causality refers to the relationship between cause and effect, where one event or factor (the cause) brings about, influences, or determines another event or outcome (the effect). It is the concept of understanding how one event or variable is responsible for producing a particular outcome.

The concept of causality poses considerable complexity when applied to reality. Most individuals share a common objective of safeguarding their health. However, the pursuit of a healthy lifestyle is influenced by various factors, such as engaging in unhealthy habits, experiencing high levels of anxiety, and neglecting regular medical check-ups. We often speculate that these choices and circumstances may lead to a shorter lifespan compared to those who prioritize their diet and receive regular medical attention. Nevertheless, it is impossible to definitively assert that engaging in specific actions or adopting all of the aforementioned measures will guarantee a long and fulfilling life. While tracing the path from cause to effect appears relatively straightforward, determining the outcomes or discerning the underlying causes proves considerably more intricate.

In the realm of research and data analysis, understanding the distinction between association and causation is vital for drawing accurate conclusions and making informed decisions. While these terms are often used interchangeably, they represent distinct concepts with significant implications. Most students and researchers have once heard the expression: *"Association does not imply Causation"*. But is it so easy to make this distinction in our everyday life?

Association refers to the statistical relationship or connection between two variables. When two variables are associated, they tend to co-occur or vary together, but this does not necessarily imply a cause-and-effect relationship. On the other hand, causation denotes a cause-and-effect relationship, where one variable directly influences or brings about changes in another variable. Differentiating between these two terms is crucial because mistaking correlation for causation can lead to misleading interpretations or misguided actions.

Questions like:

- *"I have taken an aspirin an hour ago. How likely am I to get a headache?"*

indicate an association relationship between aspirin and headache that can be represented as

marginal or conditional distributions over observable quantities (e.g., *P(headache|aspirin))*, and can be computed from the joint distribution over all variables in the domain.
But if we can answer a question that involves the effects of interventions, like:

- *"I am about to take an aspirin. Will it help my headache?"*

, this means that we know that between aspirin and headache, there is a causal relationship. By implementing an intervention, we disrupt the typical flow of influence from causes to effects, by setting some set of variables to specific values.

Interventions are indicated using the do(.) notation, where do(**x**) means that a set of variables **X** is set to values **x**. The effects of interventions will be represented using **interventional distributions** denoted with either the do(.) operator past the conditioning bar or a subscript denoting a set of intervened values (e.g., $P(\mathbf{y}|do(\mathbf{x}))$, or $P_{\mathbf{x}}(\mathbf{y})$. The effect of intervention do(**x**) on a variable set **Y** is often called the **causal effect** of do(**x**) on **Y**.

To illustrate the difference between association and causation, let's consider an example involving the attribute of someone having yellow teeth and lung cancer.

Suppose a study reveals a strong positive correlation between someone having yellow teeth and lung cancer. Maybe most of the people who smoke, have yellow teeth and it is most possible to have lung cancer because of smoking and vice versa most people who have yellow teeth may be smokers, so have a bigger probability of lung cancer. But the sentence:
If I bleach my teeth then the probability of getting lung cancer will be reduced, is not rational.

Thus, this association does not necessarily imply a causal relationship between the two variables. The association was induced due to **confounding**. A confounder is a variable that influences both the dependent and the independent variable, causing a spurious association. This means that two or more variables are associated but not causally related due to confounding factors.

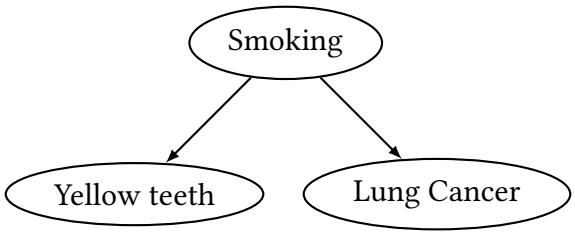


Figure 3.1: Bayesian Network that visualizes confounding.

To bring everything together, if two variables, A and B, are correlated, A causes B OR B causes A OR they share a latent common cause.

3.2 Causal Graphical Models

Model causality effectively requires establishing a unified approach that connects available data to the underlying causal mechanisms. This involves the definition of the operational mechanisms of these models and their relationship to the probability distributions of the variables that are being modeled.
Every Probabilistic Graphical model (PGM) can be described by two components:

1. the qualitative specification : **graph**

2. and the quantitative specification : **Joint Probability distribution (JPD)**.

These two components are linked with **Causal Markov Condition** and **Causal Faithfulness Condition** in the framework of semi-Markovian causal models used in this analysis.

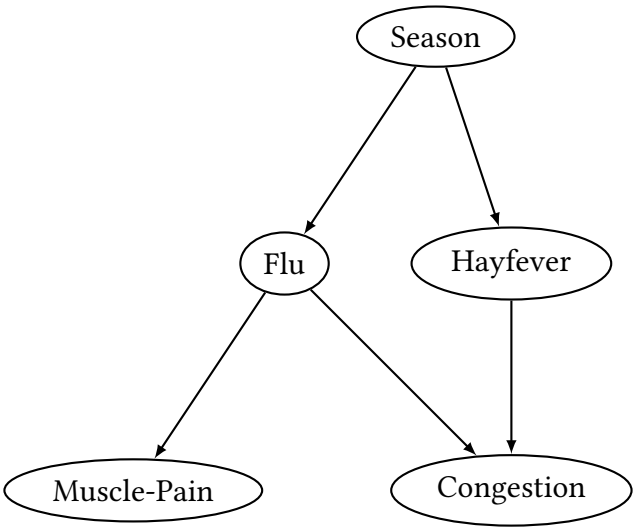


Figure 3.2: DAG and the independencies that entailed by CMC.

Let $\mathcal{G}$ be a graph presented in Figure 3.2. $\mathcal{G}$ fully describes the direct and indirect (probabilistic) causal relations among the set of measured quantities **V**. Suppose that we want to estimate the Joint Probability Distribution (JPD) $\mathcal{P}$ over **V** given a data set measuring **V**. Before presenting the **causal assumptions** connect the graph $\mathcal{G}$ with the JPD $\mathcal{P}$, some basic graph terminology must be revised.

In causally insufficient systems, where latent confounders are possible, the most common causal models are **Semi Markov causal modes (SMCMs)** or **Acyclic directed mixed graphs (ADMGs)** [36]. In this analysis, the framework of SMCMs is used. They are **mixed graphs**, meaning that they can have both directed ($\rightarrow$) and bi-directed ($\leftrightarrow$) edges.

A graph $\mathcal{G}$ is an ordered pair (**V, E**), where **V** is a set of nodes(or vertices), and **E** is a set of edges. In a mixed graph $\mathcal{G}$, a path is a sequence of distinct nodes $\langle V_0, V_1, ..V_n\rangle$ s.t for $0 \leq i < n, V_i$ and V_{i+1} are adjacent in $\mathcal{G}$. X is called a **parent** of Y and Y a **child** of X in $\mathcal{G}$ if $X \rightarrow Y$ in $\mathcal{G}$. A path from V_0 to V_n is **directed** if for $0 \leq i < n, V_i$ is a parent of V_{i+1}. X is called an **ancestor** of Y and Y is called a **descendant** of X in $\mathcal{G}$ if $X = Y$ in $\mathcal{G}$ or there exists a directed path from X to Y in $\mathcal{G}$. $\mathbf{Pa}_{\mathcal{G}}(\mathbf{X}), \mathbf{Ch}_{\mathcal{G}}(\mathbf{X}), \mathbf{An}_{\mathcal{G}}(\mathbf{X})\ and\ \mathbf{De}_{\mathcal{G}}(\mathbf{X})$ are used to denote the set of parents, children, ancestors and descendants of nodes **X** in $\mathcal{G}$, respectively. The set of variables that are connected with a variable Y through a bidirected path (i.e., a path that only has bidirected edges) is called the **district** of Y and denoted $\mathbf{Dis}_G(Y)$. A **directed cycle** in $\mathcal{G}$ occurs when $X \rightarrow Y \in \mathbf{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}(X)$. A *directed graph* is called acyclic (**DAG**) if it contains no directed cycles. Given a path p in a mixed graph, a non-endpoint node V on p is called a **collider** if the two edges incident to V on p are both into V. Otherwise, V is called a **non-collider**. A path $p = \langle X, Y, Z\rangle$, where X and Z are not adjacent in $\mathcal{G}$ is called an **unshielded triple**. If Z is a collider on this path, the triple is called an **unshielded collider**.

In our analysis, two variables in a DAG may share a latent common cause, so the **Causal Sufficiency Condition(CSC)** fails. Graphically this means that in Figure 3.2, the variables "Flu" and

"Muscle-Pain" may have an unmeasured common cause, "Covid" for example. Then there is no causal relationship between "Flu" and "Muscle-Pain", but a "non-causal dependence," or confounding and a bi-directed edge used for this representation.

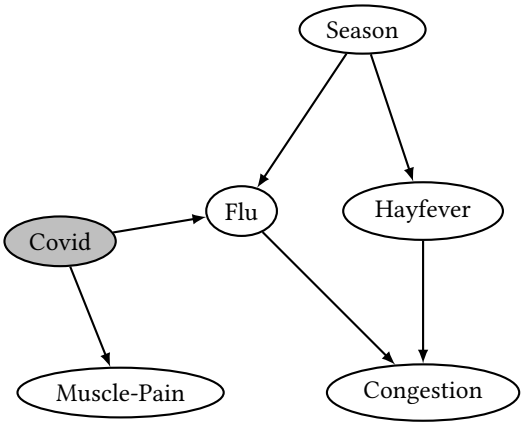


Figure 3.3: DAG with an unmeasured confounder.

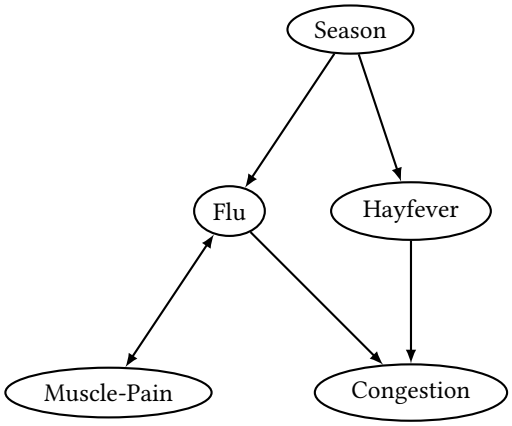


Figure 3.4: SMCMs representation.

Directed edges in a SMCM denote a causal relation that is direct in the context of observed variables. A bi-directed edge $X \leftrightarrow Y$ denotes that X does not cause Y and Y does not cause X, but (under the faithfulness assumption) the two share a latent confounder.

Now, we can proceed to the definitions of **Causal Markov Condition (CMC)** and **Faithfulness Condition (FC)** that connect a graph $\mathcal{G}$ to the JPD $\mathcal{P}$. The notation $\langle \mathbf{X}, \mathbf{Y} | \mathbf{Z} \rangle$ is used to denote that variables in **X** are independent of variables in **Y** given **Z**.

Definition 3.2.1. ***(Causal Markov Condition -CMC)****[35] Let $\mathcal{G}$ be a causal graph with node set **V** and $\mathcal{P}$ be a probability distribution over the nodes in **V** generated by the causal structure represented by $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{P}$ satisfy the Causal Markov Condition if and only if for every W in **V**, W is independent of $\mathbf{V} \setminus (\mathbf{De}_{\mathcal{G}}(W) \cup \mathbf{Pa}_{\mathcal{G}}(W))$ given $\mathbf{Pa}_{\mathcal{G}}(W)$.*

Shortly, CMC states that a variable is independent of its non-effects given its direct causes. Under the Causal Markov Condition, and according to the chain rule of probability, the joint probability distribution for a set **V** can be factorized:

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V | \mathbf{Pa}_{\mathcal{G}}(V)) \tag{3.1}$$

, where $P(V|\mathbf{Pa}_{\mathcal{G}}(V))$ denotes the probability of **V** given the set of nodes that are direct causes (parents) of V. For a given graph the CMC yields a set of independence relations.

Definition 3.2.2. ***(Faithfulness Condition)****[35] Let $\mathcal{G}$ be a causal graph and $\mathcal{P}$ a probability distribution generated by $\mathcal{G}$. $\langle \mathcal{G}, \mathcal{P} \rangle$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in $\mathcal{P}$ is entailed by the Causal Markov Condition applied to $\mathcal{G}$.*

When $\langle \mathcal{G}, \mathcal{P} \rangle$ are faithful to each other, the independencies that hold in $\mathcal{P}$ are all and only those entailed by the Causal Markov Condition. However, some of the entailed independencies are

not obvious by the CMC. In DAGs, the d-separation criterion is proposed for deciding, from a given causal graph, whether a set **X** of variables is independent of another set **Y**, given a third set **Z**. The extension of d-separation to mixed causal graphs is called **m-separation criterion**:

Definition 3.2.3. *(m-**connection**, m-**separation**) In a mixed graph $\mathcal{G} = (\mathbf{E}, \mathbf{V})$, a path p between X and Y is **m-connecting** given (conditioned on) a set of nodes* **Z**, $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$ *if*

1. *Every non-collider on p is not a member of Z.*
2. *Every collider on the path is an ancestor of some member of Z.*

*X and Y are said to be m-**separated** by Z if there is no m-connecting path between X and Y relative to Z. Otherwise, they are said to be m-**connected** given Z.*

Under the Causal Markov (**CMC**) and Faithfulness (**FC**) conditions the set of m-separations that hold in $\mathcal{G}$ correspond to the set of conditional independencies that hold in $\mathcal{P}$, the joint probability distribution (JPD), over the same set of variables and vice versa.

Causal Markov Condition	Faithful Condition
MSep(X, Y | **Z**) $\Rightarrow$ $\langle X, Y \vert \mathbf{Z} \rangle$	MSep(X, Y | **Z**) $\Leftrightarrow$ $\langle X, Y \vert \mathbf{Z} \rangle$

For example, for the graph of Figure 3.2, the CMC yields a set of independence relations that allow the factorization of its JPD. Under faithfulness assumption, knowing this factorization means that the graph in Figure 3.2 fits this JPD. This graph may not be unique. Often, multiple graphs fit the data equally well and are called **Markov equivalent (ME)**.
Independencies

$$(F \perp H|S)$$
$$(C \perp S|F, H)$$
$$(M \perp H, C|F)$$
$$(M \perp C|F)$$

Factorization of JPD :

$$P(S, F, H, C, M) = P(S)P(H|S)P(C|F, H)P(M|F)$$

If a causal DAG $\mathcal{G}$ and a JPD $\mathcal{P}$ satisfy the CMC then the tuple $\langle \mathcal{G}, \mathcal{P} \rangle$ are called a **Causal Bayesian Network**. In causal Bayesian networks, every missing edge in $\mathcal{G}$ corresponds to a conditional independence that holds in JPD $\mathcal{P}$, meaning there exists a subset of the variables in the model that renders the two non-adjacent variables independent. Respectively, every conditional independence in JPD of $\mathcal{P}$ corresponds to a missing edge in the DAG $\mathcal{G}$. This is not always true for SMCMs. Figure 3.5 illustrates an example of an SMCM where two non-adjacent variables are not independent given any subset of observed variables.
The variables A and D are m-connected given any subset of observed variables, but they do not share a direct relationship in the context of observed variables. [38]

3.3 Modeling intervention with Causal Models

The causal interpretation views the arrows in $\mathcal{G}$ as representing causal influences between the corresponding variables. In this interpretation, the factorization of CMC (3.1) still holds, but the factors are further assumed to represent autonomous data generation processes, that is, each parent-child relationship characterized by a conditional probability $P(V|\mathbf{Pa}_{\mathcal{G}}(V))$ represents

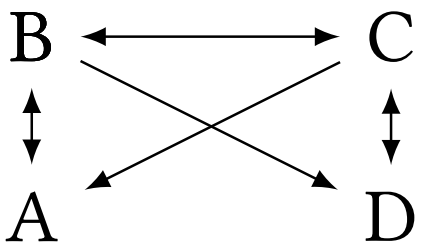


Figure 3.5: A semi-Markov causal model over variables $\{A, B, C, D\}$.

a stochastic process by which the values of V are assigned in response to the values $\mathbf{Pa}_{\mathcal{G}}(V)$ (previously chosen for V's parents), and the stochastic variation of this assignment is assumed independent of the variations in all other assignments in the model. Moreover, each assignment process remains invariant to possible changes in the assignment processes that govern other variables in the system. In order to get many causal identification results, the main assumption we will make is that interventions are local, known as **modularity assumption**:

Assumption 3.1. *(Modularity / Independent Mechanisms / Invariance) If we intervene on a set of nodes $S \subset 1, 2, .., n$, setting them to constants, then for all i, we have the following:*

1. *If $i \notin S$, then $P(x_i|\mathbf{Pa}_{\mathcal{G}i})$ remains unchanged.*
2. *If $i \in S$, then $P(x_i|\mathbf{Pa}_{\mathcal{G}i})$ = 1 if x_i is the value that X_i was set to by the intervention; otherwise, $P(x_i|\mathbf{Pa}_{\mathcal{G}i})$ =0.*

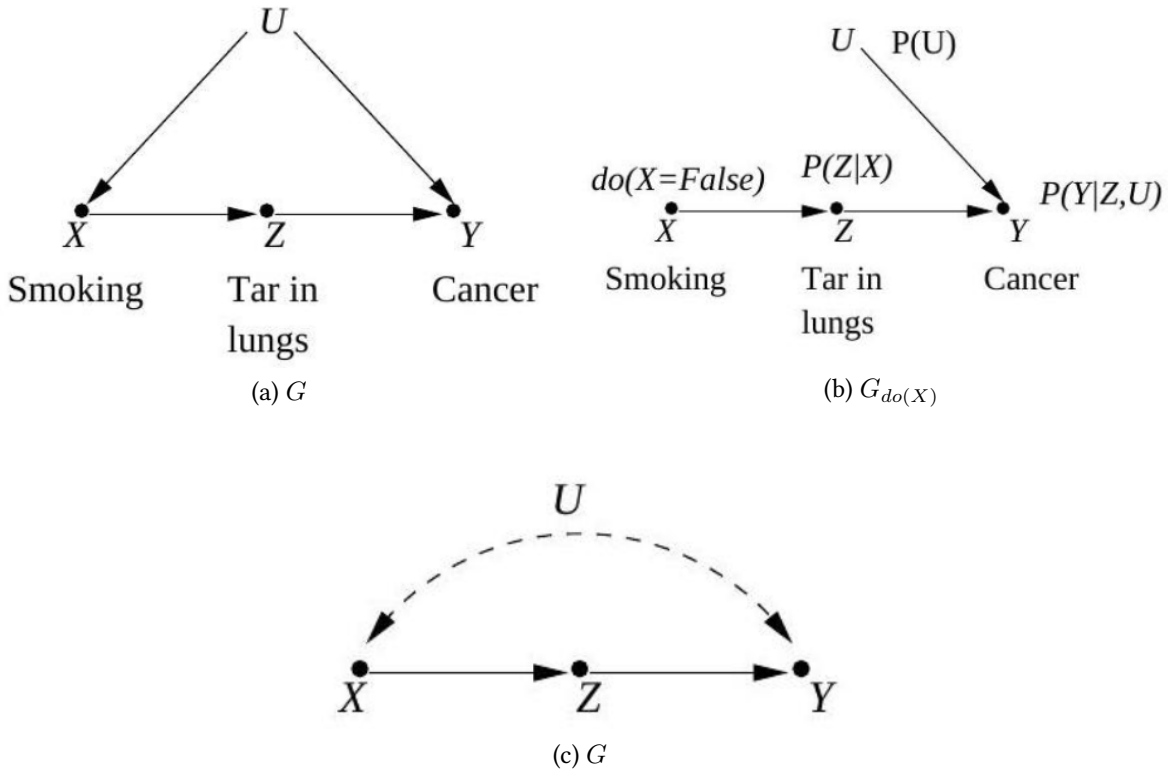
This modularity assumption [21] enables us to infer the effects of interventions, such as policy decisions and actions, whenever interventions are described as specific modifications of some factors in the product of (3.1).

The simplest such intervention called **atomic** or **hard**, involves fixing a set **T** of variables to some constants **T = t** denoted by do(**T = t**) or do(**t**), which yields the **post-intervention distribution** for the set **V** of variables:

$$P_{\mathbf{t}}(\mathbf{V}) = \begin{cases} \prod_{\{V \in \mathbf{V} \, and \, V \notin \mathbf{T}\}} P(V|\mathbf{Pa}_{\mathcal{G}}(V)) & if \quad V \, consistent \, with \, t. \\ 0 & otherwise \end{cases} \tag{3.2}$$

Eq. (3.2) represents a truncated factorization of (3.1), with factors corresponding to the manipulated variables removed. This truncation follows immediately from (3.1) since, assuming modularity, the post-intervention probabilities $P(V|\mathbf{Pa}_{\mathcal{G}}(V))$ corresponding to variables in T are either 1 or 0, while those corresponding to unmanipulated variables remain unaltered. If **T** stands for a set of treatment variables and Y for an outcome variable in **V T**, then Eq. (3.2) permits us to calculate the probability $P_t(y)$ that event Y = y would occur if treatment condition T = t were enforced uniformly over the population. This quantity, often called the "causal effect" of T on Y, is what we normally assess in a **controlled experiment with T randomized**, in which the distribution of Y is estimated for each level t of T.

An example of intervention will be given below to illustrate this condition. For the causal structure of Figure 3.6 (a) , Study describes an intervention: A scientist wants to check whether quitting smoking can prevent Lug Cancer, and therefore samples a group from the population, and



(a) G (b) $G_{do(X)}$

(c) G

Figure 3.6: A causal diagram illustrating the effect of smoking on lung cancer

then randomly assigns people into two groups: The first group is forced to quit smoking, while the latter is forced to become smokers. What happens to the system under study is that the smoking habit in the graph is no longer causally dependent on the experimental design. Graphically this is equivalent to removing the edge from the (unmeasured) U variable. This procedure is called a **Randomized Control Trial (RCT)** and was first described by Peirce and Jastrow (1885). The purpose of randomization is to minimize bias and ensure that the groups being compared are as similar as possible, except for the intervention being studied.

Graph G, in Figure 3.6(a) from [Pearl, 2000] [23] concerns the relation between smoking (X) and lung cancer (Y), mediated by the amount of tar (Z) deposited in a person's lungs. The model makes qualitative causal assumptions that the amount of tar deposited in the lungs depends on the level of smoking (and external factors) and that the production of lung cancer depends on the amount of tar in the lungs but smoking has no effect on lung cancer except as mediated through tar deposits. There might be (unobserved) factors (say some unknown carcinogenic genotype) that affect both smoking and lung cancer, but the genotype nevertheless has no effect on the amount of tar in the lungs except indirectly (through smoking). Quantitatively, the model induces the joint distribution factorized as:

$$P(u, x, z, y) = P(u)P(x \mid u)P(z \mid x)P(y \mid z, u). \tag{3.3}$$

Assume that we could perform an ideal intervention on variable X that the study describes, by banning smoking and forcing people to smoke. Then the effect of those actions is given by:

$$\begin{aligned} P_{X=False}(u, z, y) &= P(u)P(z \mid X = False)P(y \mid z, u). \\ P_{X=True}(u, z, y) &= P(u)P(z \mid X = True)P(y \mid z, u). \end{aligned} \tag{3.4}$$

which are represented by the model in Figure 3.6(b). [32]

3.4 Identifiability with causal models

Another crucial concept to comprehend is **identifiability**, which pertains to the capability of converting interventional quantities into observational quantities. Unlike interventional quantities, observational quantities can be obtained merely from standard data without the need for conducting any experiments. In causal inference, identifiability can be thought of as the condition that permits measuring causal quantity from observed data. So in identification, a causal estimand that has a do operator in it can be turned into a statistical estimand that does not have the do operator in it: $P(Y \mid do(X), Z_1, Z_2) = P(Y \mid X, Z_1, Z_2)$.

Identifiability is one of the reasons that RCT is so important. As already shown, $Y|do(X)$ or Y_X is used to denote a variable Y after the hard intervention on variable X. Also, if $\mathcal{G}$ denote a causal graph, $\mathcal{G}_{\overline{X}}$ denote the causal graph if all the incoming edges on X will be deleted, as seen in Figure 3.7 and Figure 3.8. If the graph of Figure 3.7 is the causal graph of the data-generating process, where X represents the treatment and Y is the outcome then Z is a confounder of the effect of X on Y.

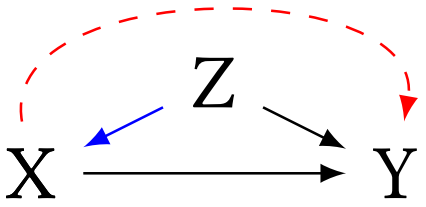


Figure 3.7: Causal Graph $\mathcal{G}$

A randomized control trial would mean that the way the treatment is assigned is just a function of a coin flip. So there should be no edge from Z to X, as seen in Figure 3.8. By randomizing treatment, we remove the blue edge which removes confounding in this case, and now the post-interventional probability of the outcome Y is identifiable:
$P(Y \mid do(X), Z) = P(Y \mid X, Z)$.

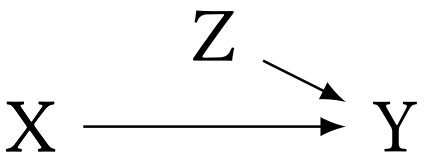


Figure 3.8: Causal Graph $\mathcal{G}_{\overline{X}}$

RCTs are widely regarded as the gold standard for establishing causal relationships between interventions and outcomes. However, conducting RCTs is not always feasible or ethical due to various practical constraints and ethical considerations, and even when experimental data are available, they usually have small sample sizes.

So, in several cases, we do not have the interventional distribution, but we have the causal graph and the observational distribution, and we still want to predict the post-interventional outcome. Here comes the formal definition of Causal-Effect Identifiability:

Definition 3.4.1. *(**Causal-Effect Identifiability**) The causal effect of X on Y is identifiable from graph G if the quantity $P(Y|do(X))$ can be computed uniquely from any positive probability of the observed variables – that is if $P_{M_1}(Y|do(X)) = P_{M_2}(Y|do(X))$ for every pair of models M_1 and M_2 with $P_{M_1}(\mathbf{V}) = P_{M_2}(\mathbf{V}) > 0$ and $G(M_1) = G(M_2) = G$.*

In this section we identify – given a causal graph - under such assumptions and graphical criteria we can estimate the effects of interventions from passive observations, using the truncated factorization formula of (3.2). Yet the more challenging problem is to derive causal effects in situations where some members of $\mathbf{Pa}_{\mathcal{G}(V)}$ are unobservable and so prevent estimation of $P(V|\mathbf{Pa}_{\mathcal{G}}(V))$ or as in most real-world problems, the causal graph is unknown.

- **Problem definition**
  To illustrate and clearly explain our problem, assume we are given a causal diagram G, together with non-experimental data on a subset **V** of observed variables in $\mathcal{G}$, and suppose we wish to estimate what effect the interventions (treatments) do(X=x) would have on a set of response variables Y, where X and Y are two subsets of **V**. In other words, we seek to estimate P(Y| do(X)) from a sample estimate of P(Y), given the assumptions encoded in $\mathcal{G}$.

In general, the identifiability of causal effects can be decided using a set of inference rules - **Pearl's do-calculus** - by which probabilistic sentences involving interventions and observations can be transformed into other such sentences. Those rules allow us to get post-intervention probabilities from pre-intervention probabilities:

Let **X**, **Y**, and **Z** be arbitrary disjoint sets of nodes in $\mathcal{G}$. We denote by $\mathcal{G}_{\overline{X}}$ the graph obtained by deleting from $\mathcal{G}$ all arrows pointing to nodes in X. We denote by $\mathcal{G}_{\underline{X}}$ the graph obtained by deleting from $\mathcal{G}$ all arrows emerging from nodes in X. Similarly, $G_{\overline{X}\underline{Z}}$ will represent the deletion of both incoming and outgoing arrows.

Theorem 3.4.1. *(Rules of do-Calculus). [Pearl, 1995] For any disjoint sets of variables X, Y, Z, and W we have the following rules.*
***Rule 1**(Insertion/deletion of observations) :*

$$P_x(y \mid z, w) = P_x(y \mid w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}. \tag{3.5}$$

***Rule 2**(Action/observation exchange) :*

$$P_{x,z}(y \mid w) = P_x(y \mid z, w) \quad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}. \tag{3.6}$$

***Rule 3**(Insertion/deletion of actions) :*

$$P_{x,z}(y \mid w) = P_x(y \mid w) \quad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\bar{X},Z(W)}} \tag{3.7}$$

, where Z(W) is the set of Z-nodes that are not ancestors of any W-node in $\mathcal{G}_{\overline{X}}$.

Theorem 3.4.2 (Shpitser and Pearl, 2006a). *Do-calculus is complete for identifying causal effects of the form $P_x(y|z)$:*

If we can identify a post-intervention probability from the pre-intervention probability, we can do this using some combination of do-calculus rules and the axioms of probability.

In practice, do-calculus may be challenging to apply manually in complex causal diagrams, and a number of **graphical criteria** have been developed for identifiability by looking at the causal

graph $\mathcal{G}$. The most influential are Pearl's **back-door** and front-door criteria. A path from X to Y is called back-door (relative to X) if it starts with an arrow pointing at X.

Definition 3.4.2. ***(Backdoor paths)****: A backdoor path is a non-causal path from X to Y. This is a path that would remain if we were to remove any arrows pointing out of X. Backdoor paths between X and Y generally indicate common causes of X and Y. The simplest possible backdoor path is the common confounding situation:* $X \leftarrow Z \rightarrow Y$.

Definition 3.4.3. ***(Back - Door)****: A set of variables* $\mathbf{Z}$ *satisfies the back-door criterion relative to an ordered pair of variables* (X_i, X_j) *in a DAG* $\mathcal{G}$ *if :*

1. *no node in* $\mathbf{Z}$ *is a descendant of* X_i*, and*
2. $\mathbf{Z}$ *blocks every back-door path from* X_i *to* X_j.

Similarly, if **X** and **Y** are two disjoint sets of nodes in $\mathcal{G}$, then **Z** is said to satisfy the back-door criterion relative to (**X**, **Y**) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

- Condition "1" in Def.3.4.3 reflects the prevailing practice that the variables that are observed parallel or after the treatment (the concomitant observations) should be quite unaffected by the treatment.
- Condition "2" in Def.3.4.3, requires that only paths with arrows pointing at X_i be blocked; these paths can be viewed as entering X_i through the back door.

Theorem 3.4.3. ***(Back-Door Criterion or Adjustment Criterion)*** *[Pearl, 1995] If a set of variables* $\mathbf{Z}$ *satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula:*

$$P_x(Y) = P(Y|do(X = x)) \sum_{\mathbf{z}} P(Y|x, \mathbf{z})P(\mathbf{z}) \tag{3.8}$$

Eq. 3.8 is called the ***adjustment formula****, and* $\mathbf{Z}$ *is an* ***adjustment set*** *for X and Y. If we know the causal SMCM* $\mathcal{G}$*, we can identify all valid adjustment sets using a sound and complete graphical criterion called the* ***adjustment criterion*** *(Shpitser, VanderWeele, and Robins 2012)[33]*

For example, in Figure 3.6(c), X satisfies the back-door criterion relative to (Z, Y):

$$P_z(y) = \sum_{x} P(y|x, z)P(x) \tag{3.9}$$

To illustrate the backdoor criterion and its role in identification the example from Pearl[2000,2009] [22] is given:

Example 3.4.1. *In Figure 3.9, the sets* $\mathbf{Z_1} = \{X_3, X_4\}$ *and* $\mathbf{Z_2} = \{X_4, X_5\}$ *meet the back-door criterion, but* $\mathbf{Z_3} = \{X_4\}$ *does not because* X_4 *does not block the path* $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$*. The same is for the* $\mathbf{Z_4} = \{\mathbf{X_6}\}$*. So, adjusting for variables of* ***Z1*** *or* ***Z2*** *sets, yields a consistent estimate of* $P(X_j|do(X_i))$ *as this probability is identifiable. But the sets* ***Z3*** *and* ***Z4*** *do not and would yield a biased estimate.*

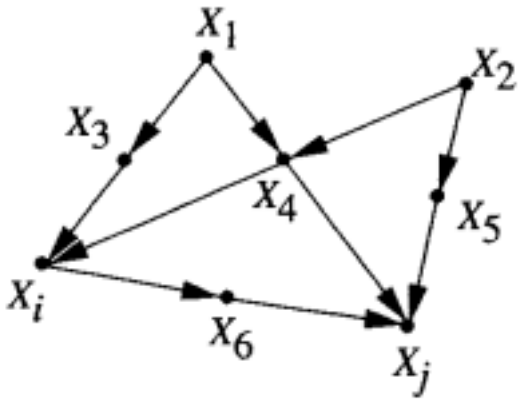


Figure 3.9: A diagram representing the back-door criterion.

In our analysis as mentioned in the introduction, we want to discover the optimal way to find the effect of a treatment on an outcome, and we assume that all variables are **pre-treatment**. The front-door criterion[Pearl,1995], illustrates how concomitants that are affected by the treatment can be used to facilitate causal inference and we do not use it in our work. So, it is important to keep in mind that there are several graphical criteria that enable the identification of causal effects of a variable X on an outcome Y, but due to space limitations and to maintain focus on the main subject matter, those aspects could not be further explored or addressed in this work. A combination of the syntactic rules of do-calculus with the graphical criterion of the back-door path is used as the main tool for causal effect identification. As an extension of the example in Figure 3.6 where the intervention mechanism was illustrated and a truncated factorization was given, it will be shown how the do-calculus rules and back-door definition can be used to derive an indicative causal effect in the same causal graph. [22]

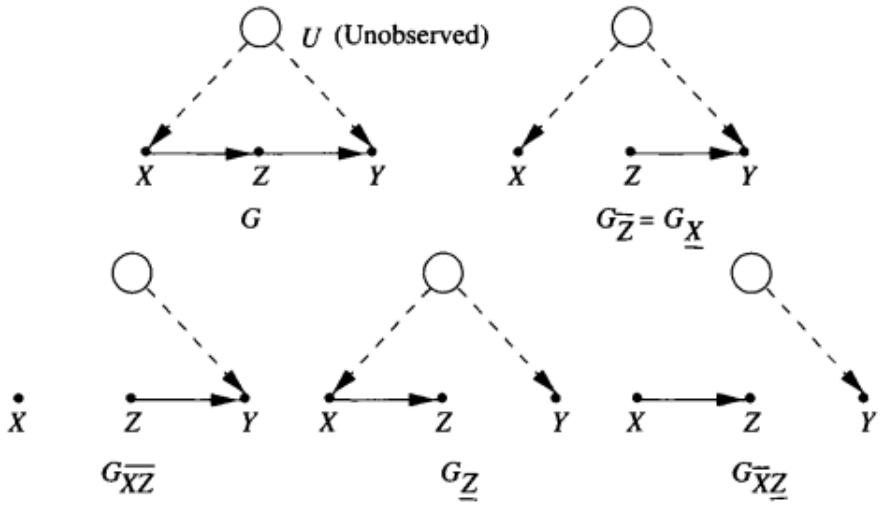


Figure 3.10: Subgraphs of $\mathcal{G}$ used in the derivation of causal effects.

Example 3.4.2. *The task in this example is to find the causal effect of tar in lugs (Z) on cancer (Y).*

- *Task: Compute P(Y |do(Z))*

Rule 2 could not be applied to exchange do(Z) with Z because $\mathcal{G}_{\underline{Z}}$ contains a back-door path from Z to Y: $Z \leftarrow X \leftarrow U \rightarrow Y$. Naturally, we would like to block this path by measuring variables (such as X) that reside on that path. This involves conditioning and summing over all values of X:

$$P(Y|do(Z)) = \sum_X P(Y|do(Z), X)P(X|do(Z)) \tag{3.10}$$

By applying Rule 3 for action deletion:

$$P(X|do(Z)) = P(X) \; if \; (Z \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{Z}}} \tag{3.11}$$

, since X and Z are d-separated in $\mathcal{G}_{\overline{Z}}$. (Intuitively, manipulating Z should have no effect on X, because Z is a descendant of X in G.) By consult Rule 2:

$$P(Y|do(Z), X) = P(Y|Z, X) \; if \; (Z \perp\!\!\!\perp Y|X)_{\mathcal{G}_{\underline{Z}}} \tag{3.12}$$

, noting that X d-separates Z from Y in $\mathcal{G}_{\underline{Z}}$. This allows us to express P(Y|do(Z)):

$$P(Y|do(Z)) = \sum_X P(Y|Z, X)P(X) \tag{3.13}$$

So, the causal effect of Z on Y is expressed in terms of observational data and thus we can say that the probability P(Y| do(Z)) is identifiable.

In conclusion, this section showed that there exists a simple graphical test, named the “back-door criterion” in Pearl (1993b), that can be applied directly to the causal diagram in order to test if a set $\mathbf{Z} \subseteq \mathbf{V}$ of variables is sufficient for identifying the causal effect of X on Y, P(Y | do(X)) and that ”do-calculus” is complete for identifying causal effects of the form P(Y|do(X), Z).

Chapter 4

Feature Selection and Markov boundaries

Feature selection is a crucial problem within the domain of machine learning, as it aims to identify the most relevant variables that contribute to an optimal predictive model. This research addresses two distinct perspectives of the feature selection problem.

- *In this chapter, our focus lies on feature selection in the context of observational data, where the objective is to identify the minimal set of features that result in the optimal prediction of a target variable Y. This minimal set corresponds to the Markov Boundary for the observational distribution. Consequently, we refer to this specific problem as the* ***Feature selection problem****.*
- *Moving forward, the subsequent chapter will shift its attention to the selection of minimal sets (if exists) of pre-treatment covariates that are maximally informative for the post-interventional distribution, P(Y| do(X)), from experimental and observational data to improve feature selection and effect estimation. So, we define the feature selection for post-interventional prediction as the* ***Causal Feature selection problem****.*

4.1 Definition of Feature selection

Definition 4.1.1. *(**Feature Selection**) The problem of feature selection in supervised learning tasks can be defined as the problem of selecting a minimal-size subset of the variables that leads to an optimal predictive model for a target variable of interest.*

Thus, the task of feature selection is to find this set of variables that exhausts the predictive information for the state of variable Y and filter out irrelevant variables or variables that are superfluous given the selected ones.(Tsamardinos and Aliferis, 2003)[43]. For observational distributions, this set is the **Markov boundary** of Y , MB(Y).

Solving feature selection problem has several advantages with one of the most significant being **knowledge discovery**. When we remove unnecessary or irrelevant variables through feature selection, it enhances our intuition and understanding of the underlying data-generating mechanisms. This helps us gain a deeper understanding of how different variables interact and influence each other, revealing important insights into the underlying causal structure. This is no accident as solving the feature selection problem has been linked to the data-generating causal network[43].

Usually, feature selection is a primary goal of the analysis and the predictive model is only a side product. This holds especially true in the healthcare domain, where the features selected through the process of feature selection are of utmost importance as they possess the potential to provide guidance and shape future experiments and studies. A motivation for employing feature selection is to **reduce the cost** associated with measuring or collecting the features. Cost-aware feature selection in healthcare involves considering the financial implications and resource constraints associated with selecting specific features. For instance, consider a scenario where a healthcare provider wants to predict the risk of a particular disease in a patient population. They have a set of potential predictors, including medical history, laboratory tests, and imaging results. However, conducting all possible tests for every patient can be costly and may not be feasible within the available resources. Without the features selected from feature selection, conducting such experiments and studies would be impractical or unethical, if not impossible.

Another impact that feature selection has is that it **reduces the computational complexity** by eliminating irrelevant or redundant features. This improves the efficiency and scalability of the knowledge discovery process, allowing for faster analysis and exploration of large datasets. By selecting a smaller set of features, the computational resources required for processing and modeling the data are optimized. Also often improves the **predictive performance** of the resulting model in practice, especially in high-dimensional settings. This is because a good-quality selection of features facilitates modeling, particularly for algorithms vulnerable to the curse of high dimensionality.

4.1.1 Why Bayesian Feature selection?

In the context of feature selection, Bayesian and frequentist approaches provide distinct methodologies. These approaches diverge in their underlying principles and computational techniques for selecting relevant features from a given set. The choice between these approaches depends on the specific requirements and assumptions of the problem at hand.

Frequentist approach to feature selection typically involves statistical techniques that evaluate the relationship between individual features and the target variable. Common methods include hypothesis testing, p-values, and model selection criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). These approaches assess the statistical significance or goodness of fit of individual features, and select those that show the strongest evidence of association with the target variable. Frequentist methods are computationally efficient and widely used in practice. However, they may not capture complex relationships or account for prior knowledge in an explicit way.

On the other hand, Bayesian approach to feature selection considers the uncertainty associated with model parameters and incorporates prior knowledge through the use of probability distributions. Bayesian methods provide a principled framework for model selection and feature selection by balancing the data-driven evidence and prior beliefs. Bayesian feature selection techniques, such as Bayesian model averaging (BMA) or reversible jump Markov chain Monte Carlo (RJMCMC), explore different feature subsets and estimate their posterior probabilities based on the data and prior information. Bayesian methods can handle small sample sizes, account for uncertainty, and allow for more flexibility in modeling complex relationships and incorporating prior knowledge.

Frequentist and Bayesian approaches to feature selection offer different perspectives and trade-offs. Frequentist methods are computationally efficient and widely used, but they may lack flex-

ibility and struggle with incorporating prior knowledge explicitly. Bayesian methods provide a more principled framework that incorporates uncertainty and prior information, but they can be computationally demanding. The choice between these approaches depends on the specific problem, available data, prior knowledge, and computational resources.

In this work, where we are motivated by healthcare settings, where the goal is often to select the treatment that will maximize a specific patient's outcome, the Bayesian approach selected due to several reasons:

- Bayesian methods allow for the explicit **incorporation of prior beliefs** and knowledge into the feature selection process. In medical research, where prior information from domain experts or existing literature is often available, Bayesian feature selection can effectively leverage this knowledge to guide the selection of relevant features.
- In medical research datasets, especially in rare diseases or specialized populations, may have **limited sample sizes**. Bayesian methods are well-suited for handling small sample sizes by providing more stable estimates through the use of prior distributions. This can lead to more reliable and robust feature selection results compared to constraint-based or score-based methods that may be sensitive to small sample sizes.
- Bayesian feature selection provides a probabilistic framework that allows for the **quantification of uncertainty** associated with the selected features. This is particularly important in medical research, where the identification of potential biomarkers or risk factors requires a measure of confidence or uncertainty. Bayesian methods can provide credible intervals or posterior probabilities to quantify the uncertainty in feature selection results.
- Can also handle complex models that involve interactions, non-linear relationships, or high-dimensional feature spaces. They offer **flexibility** in specifying the model structure and capturing intricate relationships between features and the outcome variable. This is especially relevant in medical research, where the underlying mechanisms and relationships may be complex and not easily captured by simple constraint-based or score-based methods.
- Bayesian feature selection methods, such as Bayesian model averaging, can explore multiple models and average their predictions to obtain **more robust and stable results**. This can mitigate the risk of overfitting and improve the generalization performance of the selected features. In medical research, where the goal is often to develop predictive models or risk assessment tools, Bayesian model averaging can provide more reliable and accurate predictions.

4.2 Markov Boundaries

4.2.1 Markov Boundary - Observational Markov Boundary (OBS)

Feature selection as described above, is a fundamental problem in machine learning that aims to select the minimal set of features that lead to the optimal prediction of a target variable Y. For observational distributions, this set is the Markov boundary of Y, MB(Y). This set can be used to obtain the best (and minimal) predictive model P(Y |MB(Y)) for Y as it encompasses all the

available predictive information regarding the state of variable Y. In causal graphical models, this set can be identified from the causal graph $\mathcal{G}$.

Definition 4.2.1. ***(Markov Blanket)*** *Markov blanket of a variable Y in a set of variables* $\boldsymbol{V}$ *is a subset* $\boldsymbol{Z}$ *of* $\boldsymbol{V}$ *conditioned on which other variables are independent of Y:* $Y \perp\!\!\!\perp \boldsymbol{V} \backslash \boldsymbol{Z} | \boldsymbol{Z}$.

The **Markov boundary** of Y is the Markov blanket that is also minimal (i.e., no subset of the Markov boundary is a Markov blanket) [Pearl, 2000]. In this work, for convenience, we often use the terminology **observational Markov boundary (OMB)** to denote the Markov boundary of a variable.

Thus, the **OMB** of a random variable is the set of variables that, if known, would make the variable conditionally independent of all other variables in the network. In a Bayesian network representing the joint distribution of the variables, the Markov Boundary of a node is the set of its **parents**, **children**, and the **parents of its children** (spouses).

- So for a DAG $\mathcal{G}$, the OMB of a variable Y in any distribution **faithful** to $\mathcal{G}$ is: MB(Y)= $Pa_{\mathcal{G}}(Y) \bigcup Ch_{\mathcal{G}}(Y) \bigcup Pa_{\mathcal{G}}(Ch_G(Y))$.
- And for an SMCM $\mathcal{G}$, the OMB of a variable Y in any distribution **faithful** to $\mathcal{G}$ is: MB(Y)= $Pa_{\mathcal{G}}(Y) \bigcup Dis_{\mathcal{G}}(Y) \bigcup Pa_{\mathcal{G}}(Dis_G(Y))$.

Therefore, it represents the minimal set of variables that contain all the information needed to determine the probability distribution of a variable and provides a way to characterize the dependencies between variables in a Bayesian network.

As this work is focused on healthcare settings, post-intervention covariates are not included in this model. This is because prior to the treatment assignment, these variables are not known and as a result, they cannot affect the assignment.

Assumption 4.1. *Covariates* $\boldsymbol{V}$ *are pre-treatment.*

Knowing the OMB and taking into account this assumption, the expression of OMB is simplified as the children of Y and their districts are no longer needed, and a more efficient representation of the conditional distribution of Y given **V** holds:

$$P(Y|\boldsymbol{V}) = P(Y|MB(Y)). \tag{4.1}$$

Assume that X represents the treatment and Y the outcome. For the Bayesian network seen in Figure 4.1, the Markov boundary of a variable Y in any distribution faithful to this graph contains the set of variables that are parents, children, and parents of its variables children: MB(Y) = { X, Z_2, Z_3, Z_4 }. Taking into account Assumption 3.1, the OMB is simplified: MB(Y) = { X, Z_2 }, and for this small network, two of the four variables no longer need, as only the pre-treatment covariates **V** = { X, Z_1, Z_2 } are needed for the representation of the conditional distribution of Y. [Figure 4.2]

Theorem 4.2.1. ***(Intersection Property)*** *Let X, Y, Z, and W be any four subsets of variables from V. The intersection property holds in any joint probability distribution P over variables V:*

• Intersection: $X \perp\!\!\!\perp Y|(Z \bigcup W)$ *and* $X \perp\!\!\!\perp W|(Z \bigcup Y) \Rightarrow X \perp\!\!\!\perp (Y \bigcup W)|Z$.

If P is faithful to G, then P satisfies the above property. [Stated to the work by Pena et al. (2007) and its proof is given in the book by Pearl (1988).] [24]

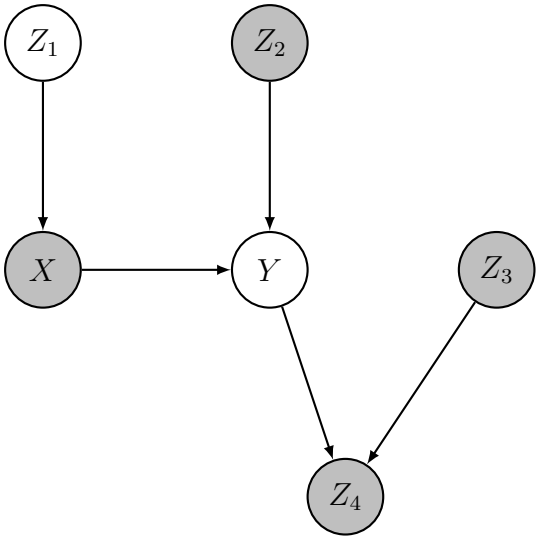


Figure 4.1: Markov Boundary without Assumption 3.1

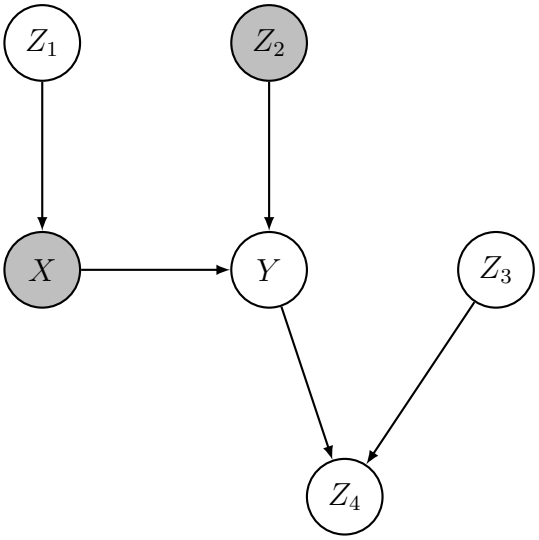


Figure 4.2: Markov Boundary and Assumption 3.1

Theorem 4.2.2. ***(Uniqueness of Markov boundaries)*** *If a joint probability distribution P over variables* $\mathbf{V}$ *satisfies the intersection property, then for each* $Y \in \mathbf{V}$*, there exists a unique Markov boundary of Y. [book by Pearl (1988)]*

Since every joint probability distribution P that is **faithful** to G satisfies the intersection property *(Theorem 3.1.1)*, then there is a unique Markov boundary in such distributions according to *Theorem 3.1.2.* Theorem 3.1.2 does not guarantee that Markov boundaries will be unique in distributions that do not satisfy the intersection property. [1]

4.2.2 Importance of Markov Boundary

This work aims to find the Markov boundary from a set of observational data and find the model that gives the optimal prediction of the post-intervention distribution. There are many algorithms and ways in the bibliography to find the set that leads to the optimal predictions[2][11]. But what is the need that leads so many people in research to deal with finding this content?
In our contemporary society, an abundance of data is pervasive. Large-scale datasets containing vast amounts of information are increasingly prevalent, presenting a challenge in terms of how to effectively manage and analyze such a colossal quantity of data. The Markov boundary is one of the key concepts, especially in probabilistic graphical models, which are used to model complex systems with multiple interdependent variables. Its ability to identify the minimal set of variables that are relevant for predicting the value of a target variable led many researchers to focus on the Markov boundary for several reasons:

- Feature selection: The Markov boundary can be used for feature selection, which is crucial in machine learning and data analysis. Feature selection can help in identifying the most relevant variables for predicting the target variable, leading to improved model accuracy and interpretability.
- Making predictions about the values of variables: In a Bayesian network, the values of some variables may be observed, while others are unknown. The Markov boundary of each unknown variable provides a set of variables that must be considered when

making predictions about its value, and ignoring any variable outside the Markov boundary can lead to biased predictions.

- Efficient computation: In large datasets, the Markov boundary can help in reducing the computational complexity by identifying the minimal set of variables needed for inference, thus speeding up the computation process.
- Learning the structure of a Bayesian network: When constructing a Bayesian network from data, it is important to identify the conditional dependencies between variables. The Markov boundary of each variable provides a compact summary of its dependencies, which can be used to guide the construction of the network.
- Performing causal inference: In some applications, it is important to identify causal relationships between variables. The Markov boundary of a variable can provide insights into the causal mechanisms that underlie its behavior and can help distinguish between direct and indirect causal effects. Also like this work trying to show, finding the Markov boundary gives a more efficient representation of the conditional distribution of the outcome variable Y given the other system's covariates.

As an extension, the concept of the Markov Boundary has found wide-ranging applications in solving real-world problems. An example focusing on the interests in this work is the effect of finding MB in medical research, where it can be used to identify risk factors for diseases. In a Bayesian network representing a medical system, identifying the Markov boundary of a disease node can help identify the minimal set of risk factors that are most predictive of the disease. Another example that involves causal inference in healthcare is the evaluation of the effectiveness of medical treatments or social interventions. By identifying the minimal set of variables that need to be adjusted for, we can try to estimate the causal effect of a treatment or intervention more accurately.
Overall, the need to deal with the Markov boundary arises from its potential to simplify the modeling process, improve model accuracy, and provide insights into the interdependencies among variables in complex systems.

4.3 Our contribution: Bayesian Feature Selection in cases of binary and ordinal outcomes

Before we start our analysis, it is important to know that the following assumptions are made throughout the entire document:

- X causes Y
- all variables **V** are pre-treatment.

As already described in Sections 4.1 and 4.2 the feature selection problem is intricately linked to the concept of the Markov Boundary, particularly in the context of observational distributions. There are many constraint-based algorithms that can find the Markov Boundary in several data settings. However as we mentioned in section 4.1, Bayesian methods sometimes can outperformed those methods for feature selection. There are, also, several constraints and challenges associated with using causal discovery algorithms to find the Markov Boundary, especially when we have to deal with ordinal target variables.

- **Problem Setting**:
  Find **Markov boundary** from observational data (**MB**(Y) or **OMB**) for data sets with mixed explanatory variables (both continuous and binary) with:
  1. **binary**
  2. and **ordinal** responses.

When the causal graph is known, under the faithfulness assumption, we can discover the Markov boundary of an outcome variable Y, as the set of its parents, children, and the parents of its children (spouses). Unfortunately, in most real-world applications, the true graph is unknown, and selecting the Markov boundary will be from the observational distribution.

We present a model-based Bayesian method that utilizes regression models, Markov Chain Monte Carlo (MCMC) sampling, and marginal distributions to find the Markov boundary of a target variable based on available observational data. The method aims to make feature selection feasible for any data structure by leveraging MCMC sampling, which allows for the calculation of posterior distributions without relying on the assumption of conjugate priors or closed-form solutions.

In the two perspectives of finding the Markov Boundary (MB) for both binary and ordinal outcome variables, the overall pipeline of the procedure remains the same. However, the distinction lies in the specific choice of the regression model employed in each case. This recognition is crucial as it ensures that the feature selection procedure is tailored to the specific characteristics of the outcome variable, leading to accurate and meaningful results in both scenarios.

Pipeline for finding MB(Y)

Our method, called FIND MB(Y), uses observational data (D_o) measuring treatment X, outcome Y, and pre-treatment covariates **V** to estimate the MB(Y) and return traces from the posterior distributions of models' parameters. The method is presented in Alg. 1.

Algorithm 1: FIND MB(Y)

Input: D_o, treatment X, outcome Y, pre-treatment covariates **V**, $number\_of\_samples$
Output: MB(Y), $traces$

$var\_subsets \leftarrow Find\_sub(X, \mathbf{V})$;
foreach *subset* $\boldsymbol{Z}$ *of var_subsets* **do**
  for $number\_of\_samples$ **do**
    Sample $\boldsymbol{\theta}$ from an uninformative $p(\boldsymbol{\theta})$;
  Fit regression model $\hat{Y} = f(\mathbf{Z}, \boldsymbol{\theta})$;
  Compute $P(D_o|\boldsymbol{\theta})$ using Eq. (4.2) or (4.5) ;
  Compute marginal likelihood: $\hat{P}(D_o) \approx \sum_{\boldsymbol{\theta}} P(D_o|\boldsymbol{\theta})p(\boldsymbol{\theta})$;
  marginals[**Z**] $\leftarrow \hat{P}(D_o)$;
  Sample from the posterior: $\boldsymbol{\theta'} \sim P(\boldsymbol{\theta}|D_o)$ using MCMC sampling, for $j = number\_of\_samples$;
  Keep traces: $traces[\boldsymbol{Z}] \leftarrow \boldsymbol{\theta'}$;
MB(Y) $\leftarrow argmax_{\mathbf{Z}}$(marginals) ;

The procedure starts by identifying all possible subsets of independent variables, ensuring that

each subset includes the treatment variable X (Line 1). For every subset, It takes samples for models coefficients from an uninformative prior distribution (Line 3). Then, it applies the appropriate regression model for computing the likelihood of the data given each model's parameters, $P(D_o|\theta)$, with random sampling for the coefficients from an uninformative prior (Line 6). The empirical marginal likelihood was calculated using the likelihood and the uninformative priors (line 7). A list keeps the marginal likelihood for every subset. The subset with the maximum marginal probability gives us the MB(Y). After identifying the MB(Y), we want to have some predictions for the model's coefficients. It uses Markov Chain Monte Carlo sampling for taking draws from the posterior of every subset's coefficients given the observational data and keeping the traces (lines 9,10).

In this part, we will provide a comprehensive explanation of the workflow, elucidate essential concepts like MCMC sampling, and delineate the differentiation among the regression methods based on the distinct nature of the outcome variable.

4.3.1 Binary Outcome

Estimating $P(D_o|\boldsymbol{\theta})$:

For computing the likelihood of the D_o given model's parameters $\boldsymbol{\theta}$, we need to set which Regression model (Line 3) will be used. In this case, when we have a binary target variable, the **Bayesian Logistic Regression for a binary response model** is used as described in Section 2.2.

Let's assume for example that $\boldsymbol{\theta}$ consists of a set of parameters $\boldsymbol{\theta} = (b_0, b_1, .., b_d)$, each of which is the coefficient of a predictor respectively. We also assume that these parameters are independent and their priors $P(b_0), .., P(b_d)$, are flat as we do not have any prior knowledge of these parameters ($P(\boldsymbol{\theta}) = P(b_0)P(b_1)\cdots P(b_d)$). In Bayesian inference procedures, in order not to set a uniform prior (E.g $P(b_1) = 1$), we use a Normal distribution with large variance instead, (E.g $P(b_1) = Normal(0, 100)$) to illustrate our non-informative prior distributions. We take random draws for each model's parameters from their uninformative - Normal distributions and given the appropriate regression model we calculate the likelihood for those draws.

As already mentioned in Section 2.2, the likelihood of a data point of a target variable Y_i will be equal to 1, is:

$$\pi_i = P(Y_i = 1) = expit(b_0 + b_1 X_{i1} + ... + b_d X_{id})$$

, given N samples, the likelihood of the data given the set of parameters will be:

$$P(D_o|\boldsymbol{\theta}) = \prod_{i=1}^{N} [\, \pi_i^{Y_i} \, (1 - \pi_i)^{(1-Y_i)} \,] = \prod_{i=1}^{N} Bernoulli_{PMF}(Y_i, \pi_i)$$

and the Log-Likelihood:

$$Log\_P(D_o|\boldsymbol{\theta}) = \sum_{i=1}^{N} [log(Bernoulli_{PMF}(Y_i, \pi_i))] \tag{4.2}$$

In our implementations, we use log_likelihood for computational and numerical stability. The likelihood function often involves the multiplication of many probabilities, which can lead to very small values. Taking the logarithm of the likelihood allows us to convert the multiplication operation into an addition operation, which is computationally more efficient and less prone

to numerical underflow. Additionally, the logarithm function has the useful property of transforming products into sums, which simplifies the mathematical calculations involved in statistical inference. It also helps to reduce the sensitivity to extreme values and improves the interpretability of the results.

In our work, we do not have conjugate priors to calculate the posterior in a closed-form solution, and numerical integrations are necessary.

Calculate Marginal likelihood $\hat{P}(D_o)$

Assume that *"d"* is the number of model regression parameters. Our goal is to calculate the marginal :

$$\begin{aligned} P(D_{\mathbf{o}}) &= \int_{\boldsymbol{\theta}_1} \int_{\boldsymbol{\theta}_2} ... \int_{\boldsymbol{\theta}_d} P(D_{\mathbf{o}}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_d) p(\boldsymbol{\theta}_1) p(\boldsymbol{\theta}_2) ... p(\boldsymbol{\theta}_d) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 ... d\boldsymbol{\theta}_d \\ &= \sum_{\boldsymbol{\theta}} P(D_o|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \ \ for \ \boldsymbol{\theta} \backsim N(0, 100) \end{aligned}$$

We take the random samples from the uninformative prior ($Normal(0, 100)$) and as we have already calculated the likelihood, we can find the empirical marginal likelihood for each model.

Sample from the posterior $P(\boldsymbol{\theta}|D_o)$:

In order to calculate posterior distribution $P(\boldsymbol{\theta}|D_o)$ for the model's parameters, we need the likelihood of the observational data given each regression model's parameters and the set of uninformative priors. Since we do not make assumptions about distributions with closed forms, calculating the posterior is not a straightforward task. Using the Bayes rule for inference:

$$P(\boldsymbol{\theta}|D_o) \propto P(D_o|\boldsymbol{\theta}) P(\boldsymbol{\theta})$$

We use Sequential Monte Carlo Sampling, as described in Section 2.4 for sampling from the posterior distribution.

Let $D_n = \{X_1, ..., X_n\}$ be the observed data. Suppose that $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$ are independent and identically distributed (iid) variables with prior distributions $\pi(\boldsymbol{\theta}_1), ..., \pi(\boldsymbol{\theta}_d)$.

A simulation may help. We draw from the posterior:

$$\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^B \sim (\boldsymbol{\theta}|D_o)$$

where the superscripts index different draws. Each $\boldsymbol{\theta}^j$ is a vector $\boldsymbol{\theta}^j = (\theta_1^j, ..., \theta_d^j)$. Now collect together the first component of each draw: $\theta_1^1, ..., \theta_1^B$. These are a sample from $p(\theta_1|D_o)$. So for example, we have three coefficients $\theta = (b_0, b_1, b_2)$ and we want B = 1000 samples for each coefficient. We will end up with an array (trace): [1000x3] where its column represents the different draws of coefficients b_0, b_1, b_2 respectively. So we can have predictions for our model's coefficients. We keep those traces, as in the next section (Causal Feature selection), we will use it again.

Software used for implementations:

- Python programming language is used.
- Bayesian logistic regression model defined in two ways:

1. Mathematical types implemented by hand.
2. Using PyMC3 library.

- Sampling implementation with PyMC3.

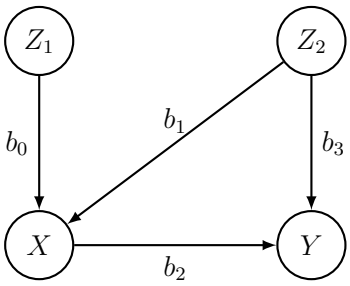
PyMC3 is a Python package for Bayesian statistical modeling and probabilistic machine learning which focuses on advanced Markov chain Monte Carlo and variational fitting algorithms. It provides users with a straightforward Python API to construct Bayesian models and perform parameter inference using Markov chain Monte Carlo (MCMC) methods. [15]

Illustrative Workflow Example: A Simple Binary Implementation

In order to enhance understanding, let us consider a straightforward example. Suppose the objective is to identify the minimum set of features that optimally predict a target variable Y. The target variable in this case is binary and represents in-hospital mortality, with the value "Yes" indicating death and "No" denoting survival. Additionally, the data set contains three other variables: the treatment variable X (steroids), as well as two continuous variables, lifestyle Z_1 and blood pressure Z_2. The real data is not yet available and you want to create a simulation of the procedure that you will follow. The implementation workflow is outlined below:

1. **Data Generating process**.
   To verify the accuracy of your implementation and assess if it correctly predicts the Markov boundary, it is essential to generate a data set that adheres to your model assumptions and parameter relationships. This data set should have fixed coefficients, allowing you to evaluate the performance of your implementation. By comparing the predicted Markov boundary with the ground truth, you can determine the correctness of your implementation and validate its effectiveness in identifying the desired set of variables.

   Assume that the true coefficients will be a = 1 and b = $[b_0, b_1, b_2, b_3]$ = [0.3, 1.25, 1.4, 1.15], and the real graph is given below. Then the observational data will be created among the expressions below following the assumptions of Bayesian Logistic Regression for a binary response model:



- $noise : e \sim N(0, 1)$
- $Z_1 \sim N(0, 15)$
- $Z_2 \sim N(0, 10)$
- $p_{xi} = expit(b_0 * Z_1 + b_1 * Z_2), X \sim Bernoulli(p_{xi})$
- $p_{yi} = expit(a + b_2 * X + b_3 * Z_3 + e), Y \sim Bernoulli(p_{yi})$

2. **Regression Cases**
   Let us consider a scenario where we generate a data set consisting of 1000 samples. We define the number of observational data as $N_o = 1000$.

   Case 1: Regression of Y on (X, Z_1, Z_2) : $Y = a + b_0X + b_1Z_1 + b_2Z_2$

(a) Define priors:

$$\left.\begin{array}{l} a \sim N(0, 100) \\ b_0, b_1, b_2 \sim N(0, 100) \end{array}\right\} \textit{big standard deviation to approach flat priors}$$

(b) Sample from the uninformative prior distribution of every parameter and calculate the marginal likelihood:

$$P(D_{\mathbf{o}}) = \int_a \int_{b_0} \int_{b_1} \int_{b_2} P(D_{\mathbf{o}}|a, b_0, b_1, b_2) p(a) p(b_0) p(b_1) p(b_2)\, da\, db_0\, db_1\, db_2 = -126.9027$$

(c) Estimate and simulate the posterior distributions of parameters after seeing the observational data, $P(\boldsymbol{\theta}|D_o)$:

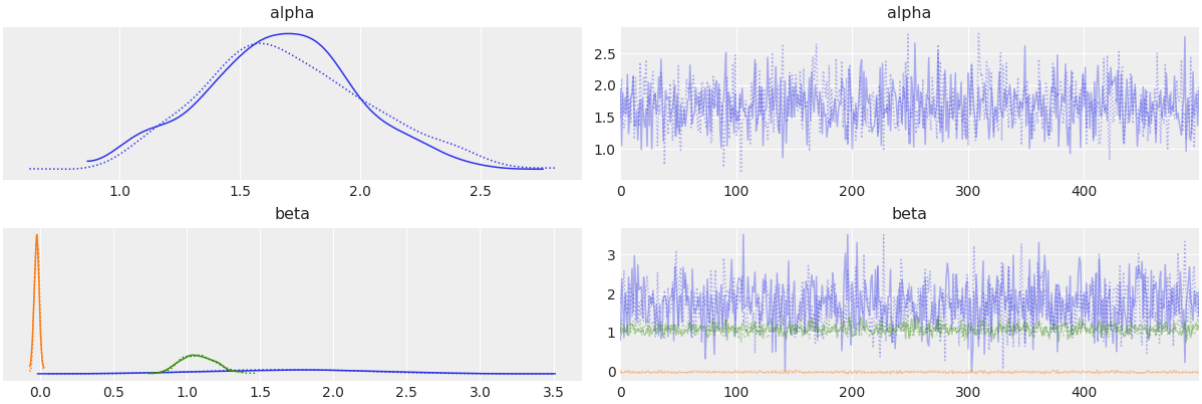


Figure 4.3: Parameters posterior simulation in case 1.

Case 2: Regression of Y on (X, Z_2) : $Y = a + b_0 X + b_1 Z_2$

(a) Define priors:

$$\left.\begin{array}{l} a \sim N(0, 100) \\ b_0, b_1 \sim N(0, 100) \end{array}\right\} \textit{big standard deviation to approach flat priors}$$

(b) Sample from the uninformative prior distribution of every parameter and calculate the marginal likelihood:

$$P(D_{\mathbf{o}}) = \int_a \int_{b_0} \int_{b_1} P(D_{\mathbf{o}}|a, b_1) p(a) p(b_0) p(b_1))\, da\, db_0\, db_1 = -121.361$$

(c) Estimate and simulate the posterior distributions of parameters after seeing the observational data, $P(\boldsymbol{\theta}|D_o)$:

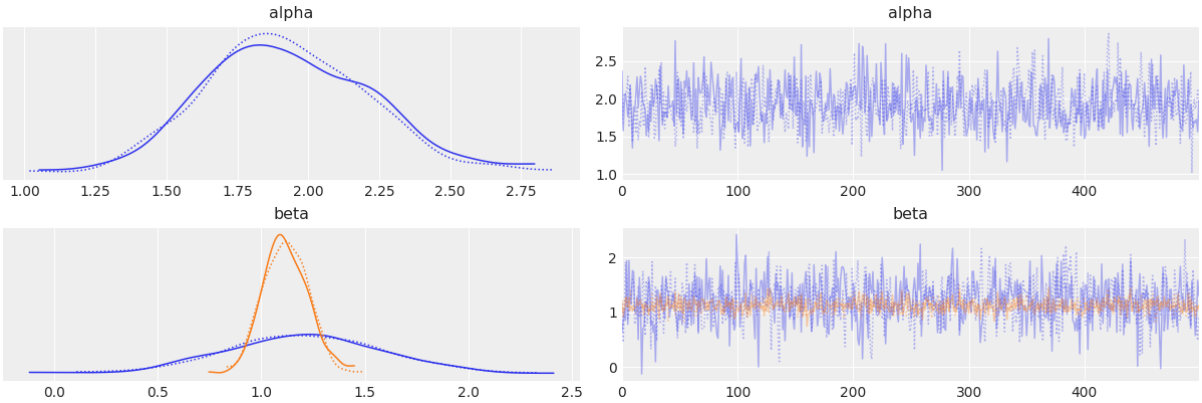


Figure 4.4: Parameters posterior simulation in case 2.

3. **Results**:
   The same process followed and for (X, Z_1) and (X) subsets and the results illustrated in the table below:

Regression Set	**Marginal Likelihood**
(X, Z_1, Z_2)	−126.90273631529186
(X, Z_1)	-319.4019659250364
(X, Z_2)	−121.36199194461011
(X)	−366.2949400457388

   - The maximum marginal likelihood gives us that the Markov Boundary from observational data is (X, Z_2).

From the DAG that represents the joint distribution of the variables in our observational data, we know the true Markov Boundary of the target variable, Y is the set of its parents, children, and the parents of its children (spouses), which is the set (X, Z_2). In this example, our method discovers from the observational distribution the correct Markov boundary.

4.3.1.1 Experiments and Results - Binary

In this section, we aim to provide further insights into our contribution to feature selection methods for determining the Markov boundary of a binary target variable in the presence of only observational distribution. We will present a series of experiments and corresponding results, focusing on the variability of the regression variable coefficients ($\boldsymbol{\theta}$) and the manipulation of the number of observational data points (N_o). Through this systematic approach, we aim to assess the method's performance under various scenarios, enabling a comprehensive understanding of its behavior and effectiveness.

Assume that the SMCM, G, and the data-generating process followed, are the same as described above.

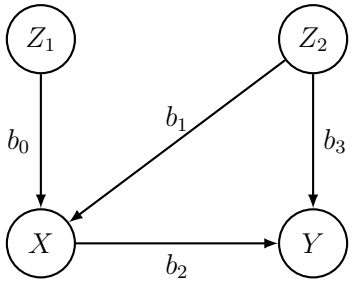


Figure 4.5: SMCM G used for experiments

As we can see from the graph G, the coefficients that mostly affect the resulting Markov boundary are b_0, b_3, and the product of $b_1 b_3$. We expect, for example, that when we use a big b_3 coefficient it will be easier for our model to understand the direct impact of Z_2 on Y or in the case when the b_0 coefficient is significantly large and in the same experiment b_3 is much smaller, it will be easier for our model to give a wrong prediction.

All these thoughts and assumptions are particularly relevant when dealing with small sample sizes of observational data. In such cases, the challenges and limitations associated with feature selection methods become more pronounced. However, as the number of observational data increases, we anticipate that our method will demonstrate greater effectiveness in identifying the correct Markov boundary. We investigate the performance of our method in several scenarios:

Scenario 1:

- For **No = 100** and 60 runs, the algorithm found the correct MB(Y) at a rate of:

  - **98%** with coefficients: a = 1 and b = [0.4 1.25 1.4 1.1]
  - **66%** with coefficients: a = 1 and b = [1.3 1.25 1.4 0.15]

The performance disparity becomes notable when the coefficient b_3, which represents the direct impact of Z_2 on X, decreases in value, while simultaneously the coefficient b_1, describing the effect of Z_1 on X, increases in magnitude.

- For **No = 1000** the algorithm found the correct MB(Y) at a rate of ≈**100**% in 60 runs in both cases.

Scenario 2:

We provide a table that illustrates the probability of finding the correct MB(Y) for different sets of ”beta” coefficients and different numbers of observational data (No) in 10 runs and the related box plot.

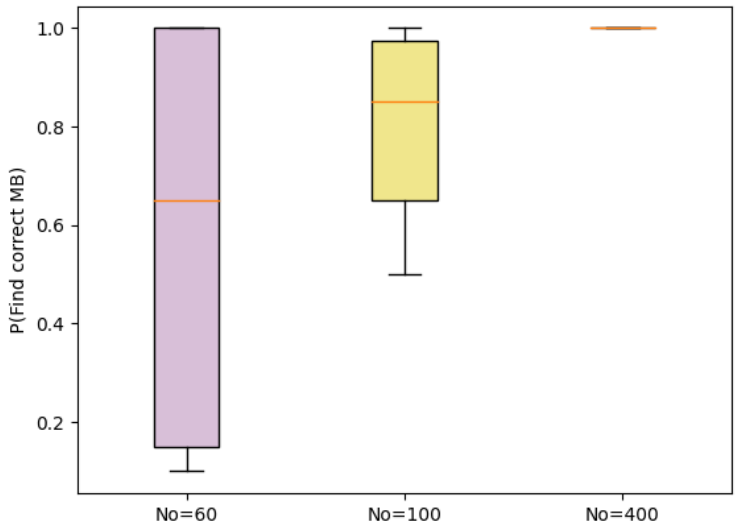


Figure 4.6: Box plot of algorithm performance for different coefficients and sample sizes.

It is readily apparent from Figure 4.7 and Figure 4.6 for larger data sets, the performance of the algorithm increased significantly. Also, from Figure 4.7 it is clear that when we have only 60 samples, the value of b_3 coefficient has an important impact on the prediction of the correct MB(Y) as we expected.

beta	No=60	No=100	No=400
[0.4 1.25 1.4 1.15]	1	0.9	1
[1.3 1.25 1.4 1.15]	1	1	1
[0.4 1.25 1.4 0.5]	0.9	1	1
[0.4 0.6 1.4 0.5]	1	1	1
[0.3 0.6 1.4 0.15]	0.1	0.5	1
[0.3 0.6 0.4 0.15]	0.1	0.6	1
[1.3 1.25 1.4 0.15]	0.4	0.8	1
[1.3 0.25 1.4 0.15]	0.3	0.8	1
[0.3 0.25 2 0.15]	0.1	0.6	1
[1.3 1.25 2 1.15]	1	0.9	1

Figure 4.7: Table of algorithm performance for different coefficients and sample sizes.

Scenario 3:

In this scenario, we aim to demonstrate the variation in performance as the value of the coefficient b_3 increases.

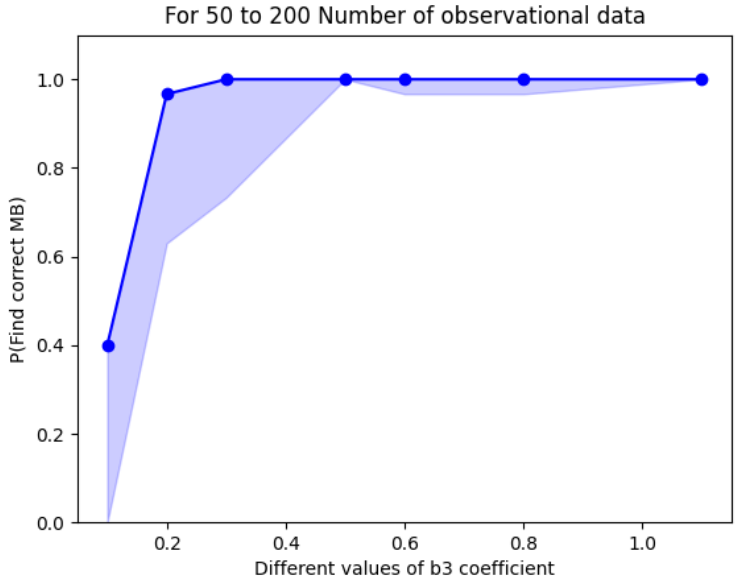


Figure 4.8: Performance as the value of the coefficient b3 increases.

4.3.2 Ordinal Outcome

Estimating $P(D_o|\boldsymbol{\theta})$:

In this case, when we have an ordinal target variable, the **Bayesian Regression for an ordinal response model** or the **Cumulative-logit Model for Ordinal Responses** is used as described in Section 2.2.

Let the response be $Y = 1, 2, ..., J$ and $x = x_1, ...x_d$ the set of d independent variables. The associated probabilities are $(\pi_1, \pi_2, ...\pi_j)$. We made the same assumptions for the set of parameters $\boldsymbol{\theta} = (b_0, b_1, ...b_d)$ and for their prios, E.g $P(b_1) = Normal(0, 100)$.

As already mentioned in Section 2.2, the cumulative probability of response less than or equal to a specific category, j, is:

$$logit(P(Y \leq j)) = \beta_{j0} + \beta_1 x_1 + \cdots + \beta_d x_d.$$

, the intercepts are different for each category but the slopes are constant across categories.

The **likelihood** function is based on the probability mass function of a **Categorical distribution**:

$$Categorical_{PMF}(Y, \pi_k) = \prod_{k=1}^{j} (\pi_k)^{[y=k]}$$

,where π_k represents the probability of a sample to be in category k and [y = k] evaluates to 1 if y=k, and 0 otherwise.

Given N samples, the **likelihood** of the data given the set of parameters will be:

$$P(D_o|\boldsymbol{\theta}) = \prod_{i=1}^{N} [\prod_{k=1}^{J}]((\pi_{yik})^{[y_i=k]}) = \prod_{i=1}^{N} Categorical_{PMF}(Y_i, \pi_{yi})$$

and the log-likelihood:

$$\boldsymbol{Log\_P(D_o|\theta)} = \sum_{i=1}^{N}[log(Categorical_{PMF}(Y_i, \pi_{yi}))] \tag{4.3}$$

Calculate Marginal, $P(D_o)$ and Sample from the posterior $P(\theta|D_o)$:

We made the same assumptions as in the binary case for the coefficients $\boldsymbol{\theta}$ and the way that the sampling is illustrated is almost the same. However, in case of ordinal data, we used the No-U-Turn Sampler (NUTS), which is a variant of the Hamiltonian Monte Carlo (HMC) algorithm.

Software used for implementations:

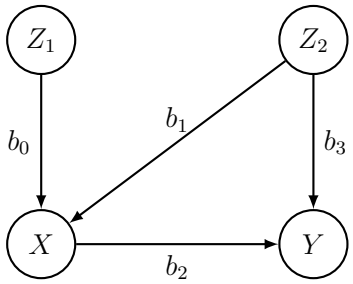
- Python programming language is used.
- Bayesian logistic regression model defined in two ways:
  1. Mathematical types implemented by hand.
  2. Using NumPyro and Jax.
- Sampling implementation with NumPyro.

NumPyro is a lightweight probabilistic programming library that provides a NumPy backend for Pyro (Pyro enables flexible and expressive deep probabilistic modeling, unifying the best of modern deep learning and Bayesian modeling). Rely on JAX for automatic differentiation and Just-in-time (JIT) compilation to GPU / CPU. It, also, supports a number of inference algorithms, with a particular focus on MCMC algorithms like Hamiltonian Monte Carlo, including an implementation of the No U-Turn Sampler.

Illustrative Workflow Example: A Simple Ordinal Implementation

Suppose again, that the objective is to identify the minimum set of features that optimally predict a target variable Y. But the target variable, in this case, is ordinal and represents in-hospital mortality in three ordered categories with the values of "0", "1" and "2" indicating the days of survival in the hospital. Additionally, the data set contains three other variables: the treatment variable X (steroids), as well as two continuous variables, lifestyle Z_1 and blood pressure Z_2. You do not have the real data in this case either and you want to create a simulation of the procedure that you will follow. The implementation workflow is outlined below:

1. **Data Generating process**.
   Assume that the true coefficients will be **a** = $[a_0, a_1]$ = [-4, 4] as cutpoints, the constant slopes will be **b** = $[b_0, b_1, b_2, b_3]$ = [0.9, 0.5, 1.4, 1.8], and the real graph is given below. Then the observational data will be created among the expressions following the assumptions of the Proportional-odds cumulative logit model as described in Sec. 2.3 :



- $noise : e \sim N(0, 1)$
- $Z_1 \sim N(0, 15)$
- $Z_2 \sim N(0, 10)$
- $X \sim Bernoulli(p_{xi}),\ p_{xi} = expit(b_0 * Z_1 + b_1 * Z_2)$
- $Y \sim Categorical(p_{yi0}, p_{yi1}, p_{yi2})$ with,
- $p_{y0} = P(Y \leq 0) = expit(a_0 + b_2 * X + b_3 * Z2 + e)$
- $p_{y1} = P(Y \leq 1) - P(Y \leq 0)$
  $= expit(a_1 + b_2 * X + b_3 * Z2 + e) - p_{y0}$
- $p_{y2} = 1 - expit(a_1 + b_2 * X + b_3 * Z2 + e)$

2. **Regression Cases**
   Let us consider a scenario where we generate a data set consisting of 1000 samples. In this scenario, we will have 397 samples in the first category "0", 175 samples in the second category "1" and 428 samples in the third category "2". We define the number of observational data as $N_o = 1000$ and the number of samples in each category as (Cat_0, Cat_1, Cat_2) = (397, 175, 428).

   Case 1: Regression of Y on (X, Z_1, Z_2)

   (a) Define priors:

   $$\left.\begin{array}{l} a_0, a_1 \sim N(0, 100) \\ b_X, b_{Z_1}, b_{Z_2} \sim N(0, 100) \end{array}\right\} big\ standard\ deviation\ to\ approach\ flat\ priors$$

   At this stage, we have executed a prior predictive simulation analysis as suggested by the *numpyro* and *PyMC* Python packages. As cited in this package, *prior predictive checks are also a crucial part of the Bayesian modeling workflow. Basically, they have two main benefits*:

   i. *They allow you to check whether you are indeed incorporating scientific knowledge into your model – in short, they help you check how credible your assumptions before seeing the data are.*
   ii. *They can help to sample considerably, especially for generalized linear models, where the outcome space and the parameter space diverge because of the link function.*

   More information on this topic can be found in [8] [28] [16]

   Prior predictive checks implementation

   Suppose that we do not see the data yet and we have no prior knowledge about our coefficients. We want to create an unopinionated model. We can try to see what our model thinks before it's conditioned on the data. The typical thing is to use wide uninformative priors. We use Normal(0, 100) for every coefficient. Suppose that we have two fictitious responders, one who has taken the treatment X=1 and one who knot X=0 and we know that the Z1 and Z2 variables had

measured 3 and 1 in both cases. We want to see what outcome our model gonna predict for these two responders without seeing the data yet. So it is only gonna use the priors that we have set.

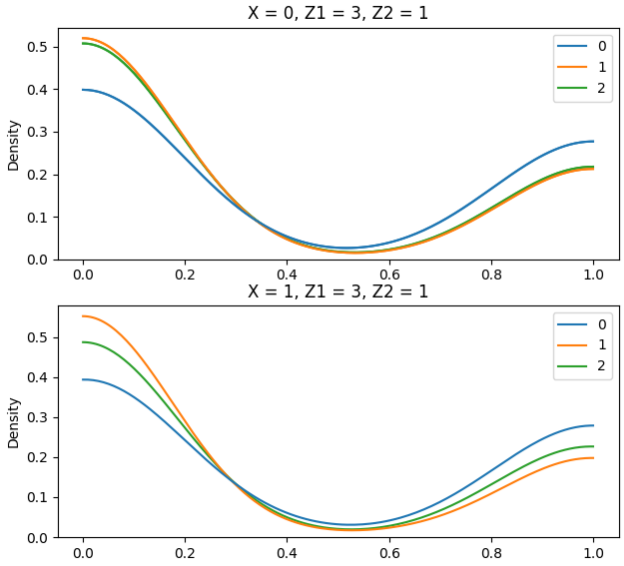


Figure 4.9: Prior predictive simulations for N(0, 100)

For each participant, this model is extremely confident for each category, the probability of the responder being in this category, is either very close to zero or very close to one. The model is not giving much weight to intermediate values. This is not expected, as we started with very wide, uninformative priors, thinking that we end up with an unopinionated model, but we have ended up with a very opinionated model, despite not seeing any data yet. The reason for that is the link function in the regression model and the transformation that it imposed. We put our parameters through the inversed logit function and just because our parameter is flat in our parameter space, does not mean it is going to be flat in the outcome space. In this case, setting a narrower prior for our coefficients and cutpoints, we will have a less opinionated model which gives more probabilities to intermediate values (Figure 4.10 and Figure 4.11).

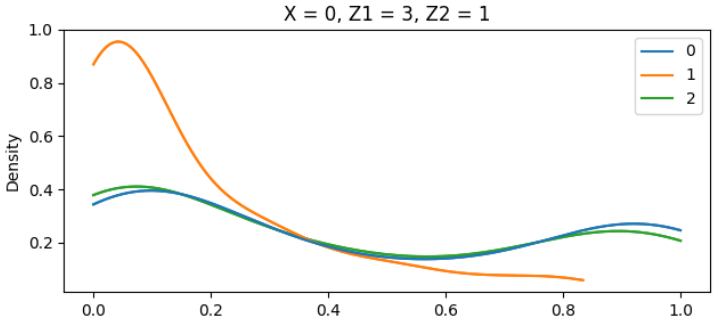


Figure 4.10: Prior predictive simulations for N(0, 1) for the first responder.

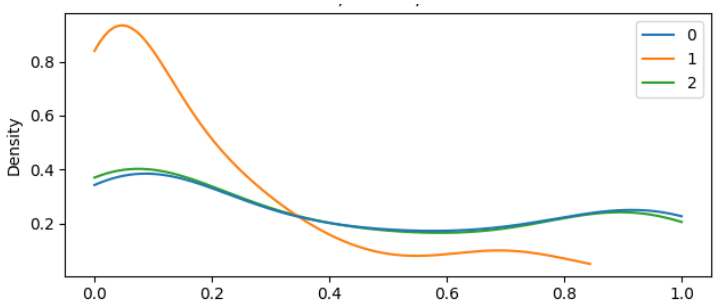


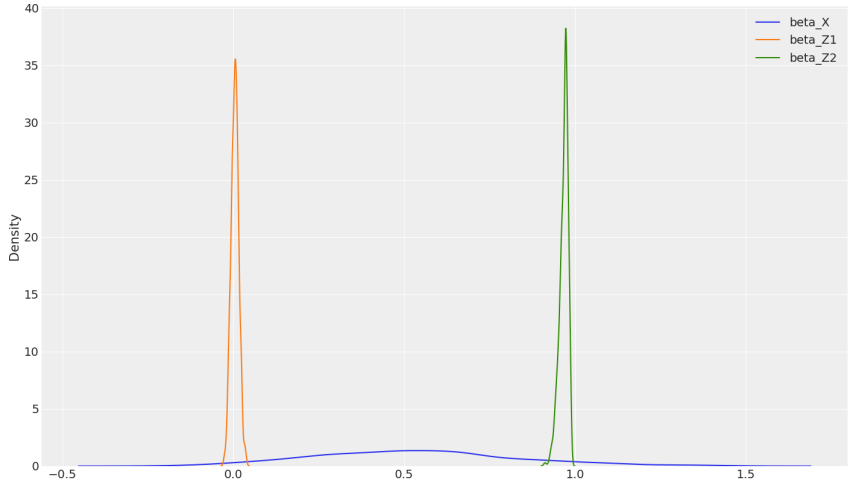
Figure 4.11: Prior predictive simulations for N(0, 1) for the second responder.

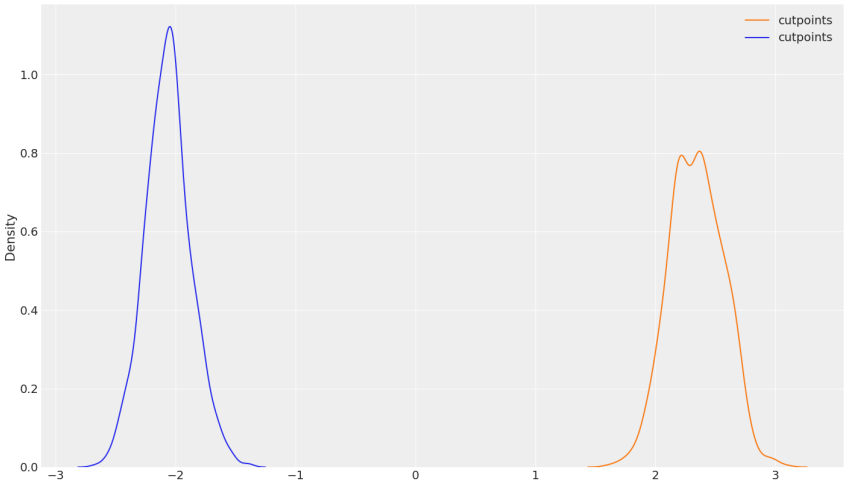
Using N(0, 1) seems that provide a more unopinionated model and this is the reason that we will use this prior from now on. The proper prior selecting is something that it needs more efficient investigation and how the logit transformation and the use of prior like N(0,100) can affect our model, especially when we have a very small data set (future work and extensions).

(b) Sample from the uninformative prior distribution of every parameter and calculate the marginal likelihood:

$$P(D_{\mathbf{o}}) = \int_{a_0}\int_{a_1}\int_{b_X}\int_{b_{Z_1}}\int_{b_{Z_2}} P(D_{\mathbf{o}}|a_0, a_1, b_X, b_{Z_1}, b_{Z_2})p(a_0)p(a_1)p(b_X)p(b_{Z_1})p(b_{Z_2})$$
$$da_0\, da_1\, db_X\, db_{Z_1}\, db_{Z_2} = -166521.4416$$

(c) Estimate and simulate the posterior distributions of parameters after seeing the observational data, $P(\boldsymbol{\theta}|D_o)$:



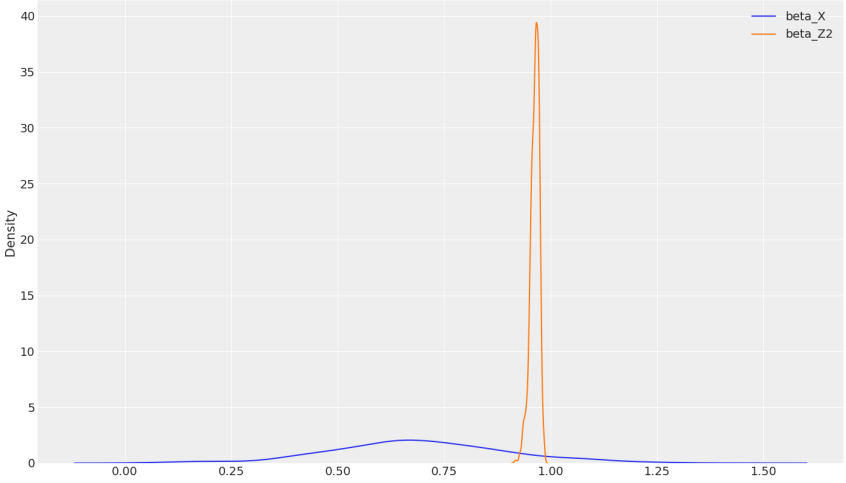


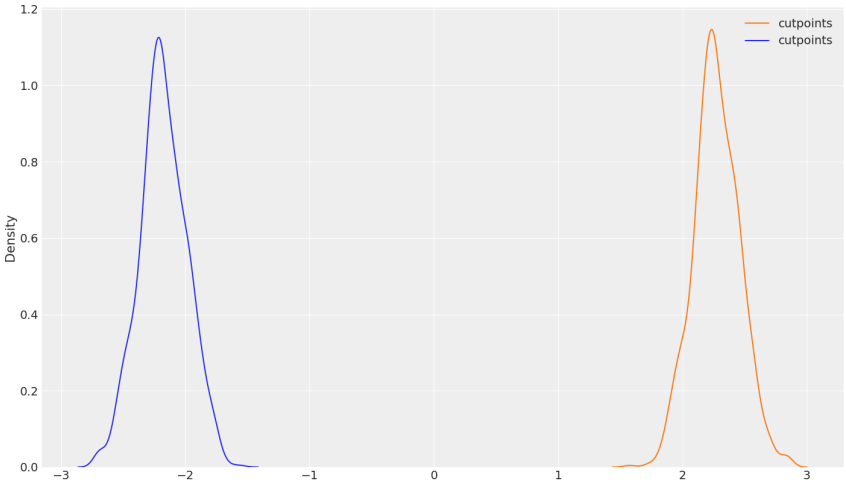
Case 2: Regression of Y on (X, Z_2)

(a) Sample from the posterior of every parameter and calculate the marginal likelihood:

$$P(D_{\mathbf{o}}) = \int_{a_0}\int_{a_1}\int_{b_X}\int_{b_{Z_2}} P(D_{\mathbf{o}}|a_0, a_1, b_X, b_{Z_2})p(a_0)p(a_1)p(b_X)p(b_{Z_2})$$
$$da_0\, da_1\, db_X\, db_{Z_2} = -165098.129973$$

(b) Estimate and simulate the posterior distributions of parameters after seeing the observational data, $P(\boldsymbol{\theta}|D_o)$:





3. **Results:**
   The same process followed and for (X, Z_1) and (X) subsets and the results illustrated in the table below:

Regression Set	Marginal Likelihood
(X, Z_1, Z_2)	−166521.44164852
(X, Z_1)	-939590.74730954
(X, Z_2)	-165098.129973
(X)	-989633.93837287

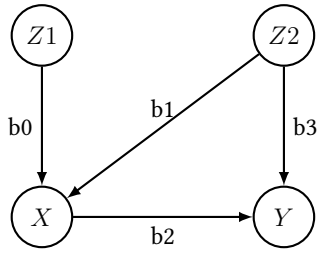
- The maximum marginal likelihood gives us that the Markov Boundary from observational data is (X, Z_2).

4.3.2.1 Experiments and Results - Ordinal

Now the performance of our " algorithm 1 " when we have mixed explanatory variables and **ordinal** outcomes will be investigated in several scenarios:

Scenario 1:

This scenario represents the performance of algorithm 1 when dealing with ordinal data for different numbers of observational data in 10 runs for coefficients:



- **alpha = [-4, 4]**
  beta = [0.9, 0.5, 1.4, 1.8]

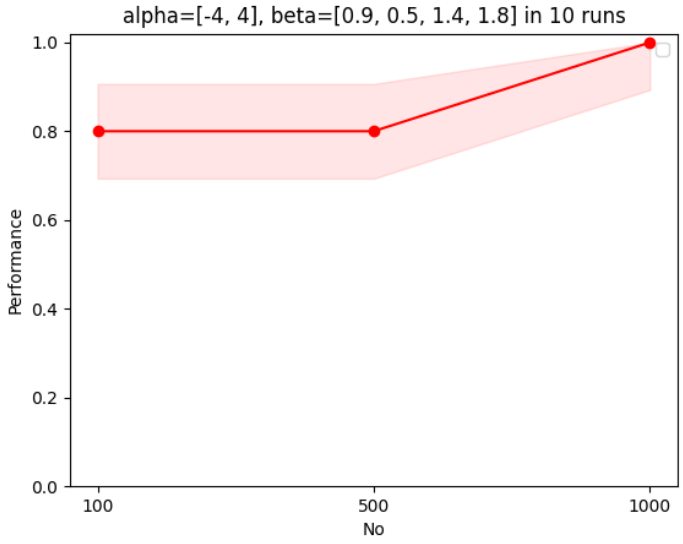
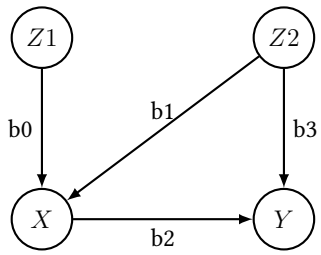


Figure 4.12: Probability of finding the correct MB(Y) for beta = [0.9, 0.5, 1.4, 1.8]

Scenario 2:

This scenario represents the performance of algorithm 1 when dealing with ordinal data for different numbers of observational data in 10 runs for coefficients:



- **alpha = [-4, 4]**
  beta = [1.3, 1.25, 1.4, 1.15]

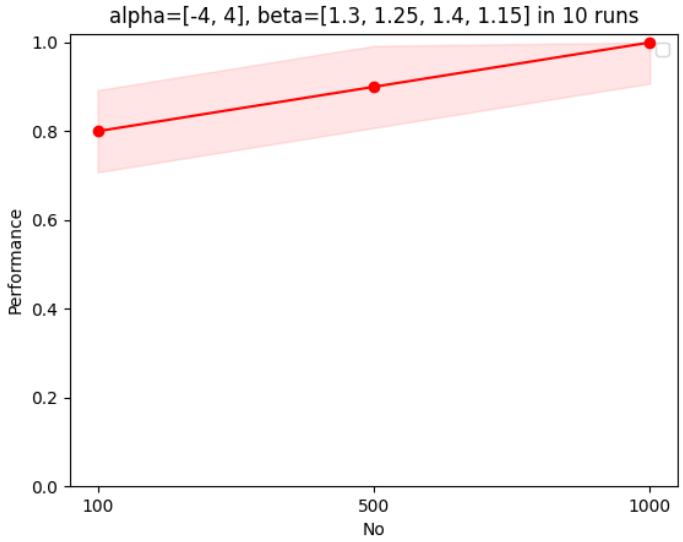


Figure 4.13: Probability of finding the correct MB(Y) for beta = [1.3, 1.25, 1.4, 1.15]

Chapter 5

Causal Feature Selection

As previously highlighted, the primary motivation for this research stems from clinical settings, where the objective is to make informed decisions regarding the optimal treatment for each patient. In such scenarios, it is of paramount importance to determine the most effective course of action tailored to the individual needs of each patient. In this section, our focus lies in exploring the optimal set of covariates that, when incorporated into a model, results in the most accurate prediction for the **post-intervention outcome**.

It is crucial at this point to reiterate that throughout the entire document, the following assumptions are made:

- X (treatment) causes Y (outcome)
- All variables **V** are pre-treatment.

5.1 Definition

Until this point, we have seen Feature Selection Methods that can find **Markov boundary** from observational data. To go back to our motivation example in Section 1.1, we can find from the **observational distribution**, the minimal set of pre-treatment covariates that makes all other variables independent of in-hospital mortality. This is just an observational prediction of the outcome Y, that answers, for example, a question like: "Whether a patient will die in hospital, given the information that you have for him, including whether you have given him steroids or not".

However, as one can envision, the ultimate objective of a doctor would be to predict the probability of in-hospital mortality for a new patient based on whether they receive a specific treatment or not. This translates in our example with this question: *Which variables should we include in a model to have the best prediction for in-hospital mortality given that we administer steroids to the patient?*

In cases where we have access to the **experimental distribution** , P_X, the set of variables that we should include in a model to get the best prediction for $P(Y|do(X))$, is the Markov boundary of Y in the post-intervention graph $G_{\overline{X}}$. (minus X). This set makes all other variables redundant for predicting Y|do(X) and is called **Interventional Markov Boundary**, $MB_X(Y)$.

Definition 5.1.1. ***(Interventional Markov Boundary - IMB)*** *Interventional Markov Boundary is the set of variables that lead to the optimal model for the post-intervention distribution of a target Y relative to a specific treatment X. IMB denoted as* $MB_X(Y)$.

It is evident that $MB_X(Y) \subseteq MB(Y)$, and with sufficient experimental data, the Markov boundary in the post-interventional distribution can be identified using the feature selection approach discussed in section 4.3. Alternatively, other statistical methods for identifying the Markov boundary MB(Y) can be employed to obtain the Interventional Markov boundary (IMB) of Y with respect to X.

However, in most real-world problems, obtaining a substantial amount of experimental data can be challenging. For instance, due to ethical concerns, it is not feasible to force a person to smoke in a clinical trial to obtain desired results on the impact of cigarettes on a specific disease. As a result, experimental data typically have much smaller sample sizes and are not powered to identify conditional distributions.

On the other hand, there are cases where we are interested in finding the optimal predictive model for the **post-intervention distribution** of an outcome Y after intervening on a treatment X, when we **only** have **observational data**. Ideally, our objective is to incorporate the Markov boundary **Z** of Y into our model. However, under causal insufficiency in which latent confounding may exist, the conditional post-interventional distribution P(Y |do(X), **Z**) may not be identifiable.

So, in cases where we can not use the Markov boundary due to not identifiability from observational distribution, we are interested in finding the set **Z** for which the post-intervention distribution P(Y |do(X), **Z**) is identifiable from observational data.

There are already some methods for identifying IMB from observational data. When the causal graph $\mathcal{G}$ is known, Shpitser and Pearl [2006] and Tian and Shpitser [2003] provide sound and complete identifiability results for estimating conditional post-intervention distributions from observational distributions induced by $\mathcal{G}$.[31][29][37]Hyttinen et al. [2015] and Jaber et al. [2019] provide similar identifiability results when the graph is unknown, using the Markov equivalence class of graphs that are consistent with the observational data.

In the work of Triantafillou's, Jabbari's, and Cooper's [40], the set **Z** for which the post-intervention distribution P(Y |do(X), **Z**) is identifiable from observational data was called **Causal Markov boundary** and it will be analytically explained in the Sections below.

The limited availability of large sample sizes in experimental data, which hinders the identification of conditional distributions, along with the issue of non-identifiability arising from observational data, as mentioned earlier, has motivated us to describe a method that provides estimates of the post-interventional distribution that are based on **both observational and experimental data**, when possible. Formally, at this stage, we define the problem of "**Causal Feature Selection**".

Definition 5.1.2. ***(Causal Feature Selection)*** *The problem of Causal Feature selection can be defined as the task of identifying the optimal set **Z** that when included in a model, yields the best prediction for the post-intervention outcome $P(Y|do(X), \boldsymbol{Z})$.*

The research paper of Sofia's Triantafillou, Fattaneh Jabbari, and Gregory F. Cooper [2021] with the title "Causal and Interventional Markov Boundaries", which is the work that motivates and initiates this thesis, provides (a) a way to predict Y_X from the observational distribution when the causal graph is known by defining the concept of causal Markov boundaries of an outcome Y relative to a treatment X, and (b) a way to combine observational and experimental data in a Bayesian manner that may help to improve the prediction of Y_X by estimating the probability of a set being $MB_X(Y)$ when the causal graph is unknown.

In this work, we present the Bayesian method that combines observational and experimental data to learn interventional Markov boundaries. Our approach builds upon the work presented in [40] and extends their findings from multinomial distributions with closed-form marginals to datasets containing both continuous and binary explanatory variables, while considering binary or ordinal outcomes. The method provides estimates of the post-interventional distribution that are based on both observational and experimental data, when possible, in which case the IMB is a CMB. The proposed method leads to a better causal effect estimation.

In order to enhance intuition and establish a solid understanding of the "Causal Feature Selection" problem, it is now appropriate to revisit our motivational example in the upcoming sections. This will allow us to elucidate the fundamental concepts that need to be defined and connect our problem with the concepts of identifiability or the adjustment set.

5.2 Connection to Identifiability

In this work, we assume that we do not know the causal graph but we have the observational and the experimental data. However, we will use graphical representation to illustrate our problem, as when observational and experimental distributions are available you can identify the $MB(Y)$ and $MB_Y(X)$ based solely on the conditional independences that these distributions entail.

We assume that we have observational data D_o (associated with distribution $\mathcal{P}$) and experimental data D_e (associated with distribution $\mathcal{P}_X$) measuring treatment X, outcome Y, and pre-treatment covariates **V**. We use N_o, N_e to denote the number of samples in D_o, D_e, respectively. Assume, also, the hypothetical causal structure of Section 1.1 Figure 1.1, an Acyclic directed mixed graph with unmeasured confounders (bi-directed edges) as illustrated in the SMCMs below:

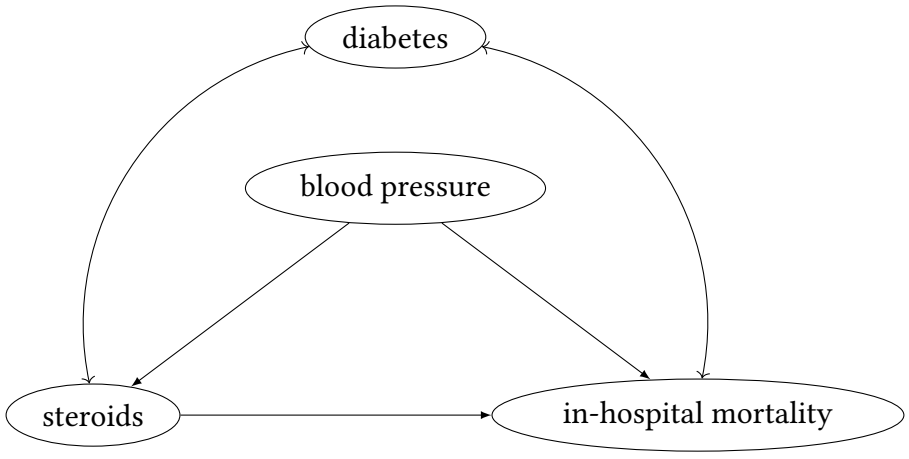


Figure 5.1: Hypothetical SMCM

Assume that we want to predict the post-intervention outcome. So the graph that describes causal relationships and the factorization of our distribution is now the SMCM post-manipulation, with all incoming edges in the intervention (treatment) variable deleted. [Figure 5.2]

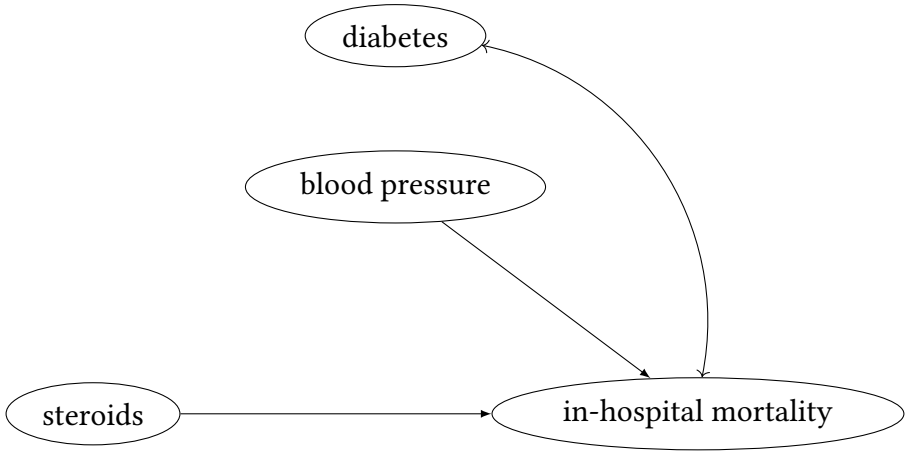


Figure 5.2: Post-Intervention SMCM

We want to predict the chance of in-hospital mortality for a patient if we give them steroids and we want to know which variables should we include in a model to get the best prediction for **In-hospital Mortality|do(steroids)**.

- If we have the experimental distribution ($\mathcal{P}_X$) we can find the Interventional Markov Boundary, $MB_X(Y)$.
- If we only have access to the observational distribution $\mathcal{P}$ and we still want to predict our post-intervention outcome, we know by looking in Figure 5.1 that the maximal amount of information is given by "blood pressure(BP)" and "diabetes(D)". The problem now is that we can **not identify P(IM | do(S), BP, D)** from observational distribution. What we aim to achieve here is to obtain a conditional distribution for predicting Y| (do(X) and some additional covariates **V**). Our objective is to include as many covariates as possible to enhance the accuracy of the prediction for Y. However, it is crucial that this probability can be expressed as a formula based on observational probabilities, considering that we only have access to observational data.

Thus, we want the smallest subset $\mathbf{Z} \in \mathbf{V}$ of variables that carries the most information for $Y|do(X)$, such that $P(Y|do(X, \mathbf{Z}))$ is identifiable from $\mathcal{P}$. Shpitser and Pearl, 2006[[30]] provide a sound and complete algorithm for identification or return "not identifiable". Let us now examine the concept of identifiability in the two distinct cases. In Figure 5.3 we do not have data for blood pressure and it is an unmeasured confounder:

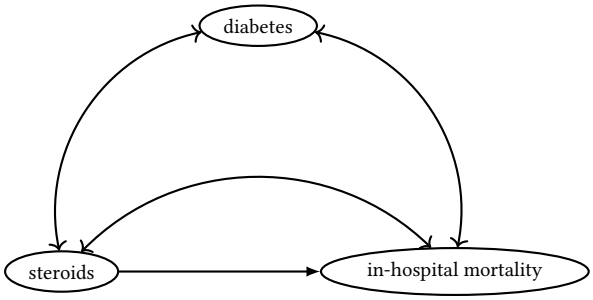


Figure 5.3: Observational distribution without measuring blood pressure.

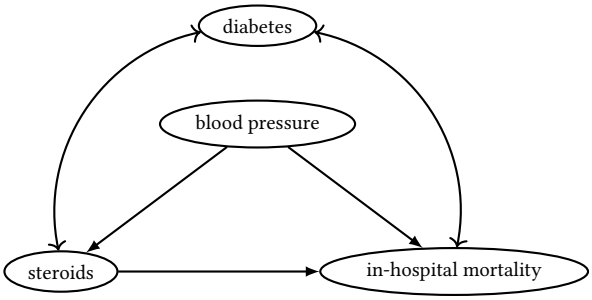


Figure 5.4: Observational distribution when measuring blood pressure.

	Graph of Figure 5.3	Graph of Figure 5.4
P(IM|do(S))	not identifiable	$\sum_{bp} P(IM\|S,bp)P(bp)$
P(IM|do(S),BP)	N/A	P(IM|S, BP)
P(IM|do(S),D)	not identifiable	not identifiable
P(IM|do(S),BP,D)	N/A	not identifiable

It is evident that the role of identifiability in causal feature selection is of utmost importance and holds significant prominence. Identifiability refers to the ability to accurately estimate the causal effects of interventions based on observed data. In the context of causal feature selection, identifiability determines whether we can ascertain the causal relationships between the treatment variable (X), the outcome variable (Y), and the set of covariates (Z) that should be included in the model.

5.3 Connection to Adjustment set

Whenever we undertake to evaluate the effect of one factor (X) on another (Y), the question arises as to whether we should **adjust** (or "standardize") our measurements for possible variations in some other factors (Z), otherwise known as "covariates," "concomitants," or "confounders" [5].

Adjustment involves dividing the population into distinct groups based on the similarity of Z, analyzing the impact of X on Y within each group, and subsequently taking the average of the findings. The illusive nature of such adjustment was recognized as early as 1899 when Karl Pearson discovered what is now called *Simpson's paradox*: Any statistical relationship between two variables may be reversed by including additional factors in the analysis. For example, we may find that students who smoke obtain higher grades than those who do not smoke but, adjusting for age, smokers obtain lower grades in every age group and, further adjusting for family income, smokers again obtain higher grades than nonsmokers in every income–age group, and so on.

Back to our problem, we want the minimal set of covariates that have the maximal amount of information for our post-intervention outcome, while also ensuring identifiability. **Covariate adjustment** stands as the primary approach for estimating causal effects using observational data. Extensive research has been conducted in the domains of potential outcomes and causal graphs to identify appropriate sets of covariates for adjustment.

In trying to find the best model for predicting Y_X from the observational distribution, when the causal graph is known, Triantafillou, Jabbari, Cooper in "Causal and Interventional Markov Boundaries"[2021][40] gave three conditions that conduct the context of **Causal Markov Boundary**.

Definition 5.3.1. ***(Causal Markov Boundaries** - $CMB_X(Y)$):*
Let $\mathbf{Z} \subseteq (\mathbf{V} \cup X)$*, and* $\mathbf{W} = \mathbf{Z} \backslash X$*. Then* $\mathbf{Z}$ *is a causal Markov boundary (CMB) for* Y *relative to* X *if it satisfies the following properties:*

1. $P(Y \mid do(X), \mathbf{W})$ *is identifiable from* $P(X, Y, \mathbf{V})$*.*
2. *For every subset* $\mathbf{W}'$ *of* $\mathbf{V} \backslash \mathbf{W}$ *either* $P(Y \mid do(X), \mathbf{W}, \mathbf{W}') = P(Y \mid do(X), \mathbf{W})$ *or* $P(Y \mid do(X), \mathbf{W}, \mathbf{W}')$ *is not identifiable from* $P(X, Y, \mathbf{V})$*.*

3. $\nexists \mathbf{W}' \subset \mathbf{W}$ *s.t.* $P(Y \mid do(X), \mathbf{W}') = P(Y \mid do(X), \mathbf{W})$.

The notation $\mathbf{CMB}_X(Y)$ *used to denote the set of causal Markov boundaries of Y relative to X.*

Condition1:

- Ensures that the post-intervention conditional probability of Y_X given a CMB is identifiable.

Condition 2:

- States that the covariates that are not in that CMB are either redundant for the prediction of Y_X given the CMB, or they make the post-intervention distribution non-identifiable.

Condition 3:

- No variable from causal Markov boundary, **Z**, can be removed without losing some information for Y_X.

It is important to keep in mind that Causal Markov boundaries do not make all remaining variables redundant for predicting Y, as the Markov boundary does. Also, $\mathbf{CMB}_X(Y)$ may not be unique or can be empty; thus either multiple sets satisfy Definition 3.2 or none. For example if $X \rightarrow Y$ and $X \leftrightarrow Y$ in G, there is no $CMB_X(Y)$.

Some definitions and lemmas are given below that explain some useful concepts and provide some important results in determining a minimal set of maximally predictive variables for which we can use observational data to predict post-interventional distributions. In Section 3.4 we gave the definition of Backdoor paths. For pre-treatment covariates, sets that m-separates X and Y in $G_{\underline{X}}$ called **backdoor sets**, since they block all backdoor paths between X and Y.

Lemma 1. *Let Z be a set for which* $P(Y \mid do(X), Z)$ *is identifiable from* $P(Y \mid X, Z)$*, then Z is a subset of a backdoor set.[proof provided in the supplementary of [40]]*

Lemma 2. *Every causal Markov boundary is a backdoor set. [proof provided in Appendix A]*

Those lemmas end up with this important Theorem:

Theorem 5.3.1. $CMB_X(Y)$ *satisfied the adjustment criterion with reference to X, Y.*

Adjustment amounts to selecting a proper set of variables **V** and "adjusting" for their effect to obtain the Interventional Distribution:

$$P_X(Y) = P(Y|do(X = x)) = \sum_{\mathbf{z}} P(Y|do(x), \mathbf{z}) P(\mathbf{z}|do(x)) \tag{5.1}$$

$$= \sum_{\mathbf{z}} P(Y|x, \mathbf{z}) P(\mathbf{z}) \tag{5.2}$$

Example Graph:

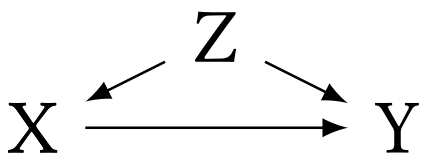


Figure 5.5: Example Graph for adjustment set.

Equation (5.1-2) is called the **adjustment formula**, and set **Z** is an **adjustment set** for X and Y.

In this work, as we deal only with pre-treatment covariates the adjustment set is equivalent to the set of Causal Markov boundary:

- $CMB \sim adjustment\,set$

Thus, we only need to look for sets **Z** where $P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z})$. If we block all backdoor paths with a backdoor set **W**, taking into account only pre-treatment variables, then by condition on this set, the identifiability is secured: $P(Y \mid do(X), \boldsymbol{W}) = P(Y \mid X, \boldsymbol{W})$

Example 5.3.1. *Example in backdoor path criterion*

To enhance intuition about the identification, the backdoor path criterion and the adjustment sets, the following example is given.[3]
Suppose that we want to identify the causal effect of X on Y and calculate the conditional probability $P(Y \mid do(X))$.
As we can see from the graph in Figure 5.7, the probability $P(Y \mid X) \neq P(Y \mid do(X))$ *is not identifiable from the observational data.*
If we draw the interventional distribution in Figure 5.6, we remove every backdoor path from X to Y. So when we intervene on X is easy to isolate the causal associational and calculate $P(Y \mid do(X))$. *But this means that we have the interventional distribution.*

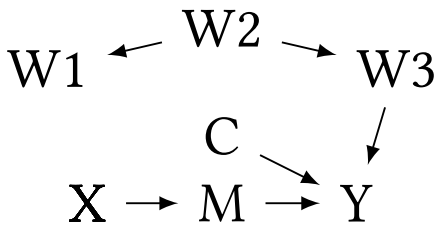


Figure 5.6: Graph of Interventional data

If we have only observational data, the way to do that is by including a backdoor set in the conditioning set of $P(Y \mid X)$. *As the backdoor paths in Figure 5.7 are* $X \leftarrow W_1 \leftarrow W_2 \rightarrow W_3 \rightarrow Y$ *and* $X \leftarrow C \rightarrow Y$, *we can block those paths with more than one way. By including one of the sets* $\{C, W_1\}$ *or* $\{C, W_2\}$ *or* $\{C, W_1, W_3\}$ *in conditional probabilities,* $P(Y \mid do(X), C, W_1)$, $P(Y \mid do(X), C, W_2)$, $P(Y \mid do(X), C, W_1, W_3)$ *can be identified from observational data. Thus, we can identify the causal effect of X on Y from observational data.*

The main two theorems that help us to proceed with the implementation afterward tell us that for the Causal Markov boundary, we need to search only for subsets of the Markov boundary

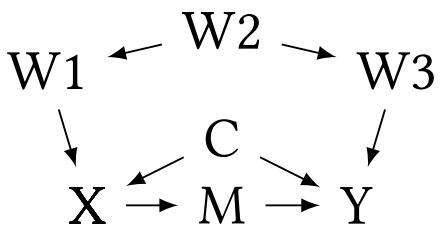


Figure 5.7: Graph of Observational data

and that if the IMB is a CMB, so we have identifiability for the IMB, this only happens when the IMB is the same with Markov Boundary:

Theorem 5.3.2. *We assume that P_x and $G_{\overline{X}}$ are faithful to each other. Every CMB **Z** of an outcome variable Y w.r.t a treatment variable X is a subset of the MB(Y):*

$$CMB_X(Y) \subseteq MB(Y)$$

Theorem 5.3.3. *If $MB_X(Y)$ is a causal Markov boundary, then $MB_X(Y) = MB(Y)$.*

Proofs of Theorem 5.3.2 and 5.3.3 are in Appendix A.

Example 5.3.2. *In the causal graph $\mathcal{G}$ bellow, $\{X, Z_2\}$. is the causal Markov boundary for Y_X.*

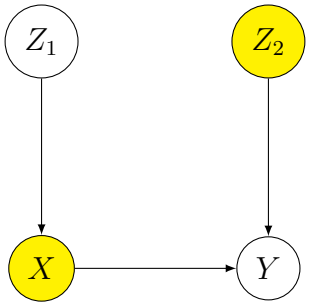


Figure 5.8: Causal Markov boundary of $\mathcal{G}$

For the graph in Figure 5.8 the $CMB_X(Y) = MB_X(Y) = MB(Y)$. In this work, we are interested in identifying the optimal set Z, the optimal adjustment set, for which the post-intervention distribution P(Y |do(X), Z) is identifiable from observational data, which we call the causal Markov boundary.

The aim of this thesis is not to explain the way of founding causal Markov boundaries with graphical criteria but to use this notion to improve feature selection and effect estimation when we do not know the causal graph but have observational and limited experimental data. Graphical examples and proofs are given in the paper and the Supplementary of "Causal and Interventional Markov Boundaries". [40]. The most useful results to keep in mind for this work are that CMBs are subsets of the observational Markov boundary and that if the $MB_X(Y)$ is a CMB then $MB_X(Y) = MB(Y)$.

Table 5.1: Different Markov boundaries.

Markov Boundary MB(Y)	Optimal prediction of Y from the observational distribution $\mathcal{P}$.
Interventional Markov Boundary $MB_X(Y)$	Optimal prediction of $Y\|do(X)$ from the experimental distribution P_X.
Causal Markov Boundaries $\mathbf{CMB}_X(Y)$	Optimal prediction of $Y\|do(X)$ from the observational distribution $\mathcal{P}$

Table 5.1 summarizes the types of Markov boundaries discussed in this thesis.

5.4 Learning optimal feature sets from observational and experimental data

When we have knowledge of the causal graph, it is possible to obtain Causal Markov boundaries by examining subsets of the Markov Boundary that adhere to Definition 5.3.1. However, it is important to note that in many real-world scenarios, the true causal graph is unknown to us. Consequently, it becomes challenging to select the precise causal or interventional Markov boundary solely based on observational data.

Because of the limited availability of large sample sizes in experimental data, the conditional effects that can be derived from them, have high variance and may not be reliable. Also, the problem of non-identifiability that arises from observational data, as previously mentioned, poses a challenge in estimating the post-interventional distribution based solely on observational data. This lack of identifiability makes it unreliable to calculate the post-interventional distribution in certain cases using observational data alone.

In the context of healthcare settings, which is the focus of this thesis, we usually have a large amount of observational data (e.g. from Electronic Health Records) but experimental data are typically limited in comparison, primarily due to factors like cost or ethical considerations. This scenario frequently arises in embedded trials, where the number of non-randomized patients exceeds that of trial participants.

This motivated us to propose a methodology that allows to leverage both observational and experimental data, whenever feasible, for the purpose of predicting the post-interventional outcome. Combining all available data in a Bayesian Manner may help to improve the prediction of Y_x.

- **Problem Setting**
  Without having the causal graph: Find the post-interventional outcome using both observational and experimental data for data sets with mixed explanatory variables when the response variable is **binary** or **ordinal** using Bayesian Regression methods.

Triantafillou, Jabbari, and Cooper in the work of "Causal and Interventional Markov Boundaries"[2021] [40] presented their result on this problem for multinomial distributions. In this thesis, we are trying to extend these results to distributions without closed-form marginals. We investigate and provide experiments and results in distributions with both continuous and binary explanatory variables when the outcome variable is **binary** or **ordinal**.

The underlying concept that drives this method is that, when the IMB is a CMB, we can use both the observational and the experimental data to estimate the conditional post-intervention distribution. Otherwise, we can use only experimental data to derive the estimate. The complete algorithm, "*FindIMB*", that provides a solution to this problem is derived and analyzed in [40]. Due to limitations in space and time, we refrain from providing a comprehensive analysis of the entire algorithm.

Instead, we focus on identifying the cases where a subset of the Markov boundary is IMB and CMB, so both observational and experimental data can be used for predicting the post-interventional distribution. We present a Bayesian method that mimics the way described in Section 4.3 for Feature selection and uses the Regression models for binary and ordinal variables as well.

We assume that we have observational data D_o and experimental data D_e measuring treatment X, outcome Y, and pre-treatment covariates **V**. We use N_o, N_e to denote the number of samples in D_o, D_e, respectively. We use the following notation to express our hypothesis:

- $H^c_{\boldsymbol{Z}}$ is a binary variable denoting the hypothesis that **Z** is the IMB $MB_X(Y)$, and it is also a CMB: $\boldsymbol{Z} = MB_X(Y) \wedge \boldsymbol{Z} \in \boldsymbol{CMB}_X(Y)$.
- $H^{\bar{c}}_{\boldsymbol{Z}}$ is a binary variable denoting the hypothesis that **Z** is the IMB $MB_X(Y)$, but it is not a CMB: $\boldsymbol{Z} = MB_X(Y) \wedge \boldsymbol{Z} \notin \boldsymbol{CMB}_X(Y)$

For a set **Z*** if either $H^c_{\boldsymbol{Z^*}}$ or $H^{\bar{c}}_{\boldsymbol{Z^*}}$ is true, **Z*** is an IMB and therefore:

$$P(Y|do(X), \boldsymbol{V}) = P(Y|do(X), \boldsymbol{Z^*} \backslash X).$$

Under $H^c_{\boldsymbol{Z^*}}$, , however, **Z*** is also a CMB and therefore the pre- and post-intervention distributions are the same, i.e.,

$$P(Y|do(X), \boldsymbol{Z^*} \setminus X, H^c_{\boldsymbol{Z^*}}) = P(Y|X, \boldsymbol{Z^*} \setminus X) \tag{5.3}$$

In contrast, under $H^{\bar{c}}_{\boldsymbol{Z}}$ (i.e., **Z*** is an IMB but not a CMB), $P(Y|do(X), \boldsymbol{Z^*} \backslash X)$ is not identifiable from observational data.

This means that $H^c_{\boldsymbol{Z^*}}$ is true only when $\boldsymbol{Z^*}$ is the IMB, and form the definition on CMB.

Specifically, if $P(Y|do(X), MB_X(Y) \backslash X)$ was identifiable from observational data, then it would satisfy all conditions in Definition 5.3.1, and it would therefore be a CMB. Therefore, if $H^{\bar{c}}_{\boldsymbol{Z^*}}$ holds, $P(Y|do(X), MB_X(Y) \setminus X)$ is not identifiable from $\mathcal{P}$, and we cannot use D_o to estimate $P(Y|do(X), \boldsymbol{Z^*} \setminus X)$. Thus, if $H^c_{\boldsymbol{Z^*}}$ holds, we can use both D_o and D_e in our estimation of $P(Y|do(X), \boldsymbol{Z^*} \setminus X)$, while if $H^{\bar{c}}_{\boldsymbol{Z^*}}$ holds we can only use D_e.

The algorithm *FindIMB* [40] constitutes a Bayesian method, that uses both D_o and D_e to estimate the probability of a set being the $MB_X(Y)$ and estimate $P(Y|do(X), \boldsymbol{V}) = P(Y|do(X), MB_X(Y) \backslash X)$. We will provide this algorithm but its step will not be analyzed. The analytical explanations and the closed-form expressions for multinomial distribution can be found in Supplementary Material of the "Causal and Interventional Markov Boundaries" [40]. Algorithm "FindIMB" first estimates the OMB of Y in observational data MB(Y) (Line 1), and then looks among subsets of MB(Y) for sets that are IMBs (Line 2). It uses D_e and D_o to evaluate the probability that a set is an IMB (Line 3), and then returns a weighted average for P(Y |do(X), **V**) based on these probabilities (Line 5).

Algorithm 2: FindIMB

Input: D_o, D_e, treatment X, outcome Y, pre-treatment covariates **V**
Output: :Post-intervention distribution $P(Y|do(X), \boldsymbol{V})$

1 $MB(Y) \leftarrow MarkovBoundary(Y, D_o)$;
2 **foreach** *subset* $\boldsymbol{Z}$ *of MB(Y) and* $C = c, \bar{c}$ **do**
3 Compute $P(H^C_{\boldsymbol{Z}}|D_o, D_e)$;
4 Compute $P(Y|do(X), \boldsymbol{V}, D_e, D_o, H^C_{\boldsymbol{Z}})$;
5 $P(Y|do(X), \boldsymbol{V}) \leftarrow \sum_{\boldsymbol{Z}} \sum_{C=c,\bar{c}} P(Y|do(X), \boldsymbol{V}, D_e, D_o, H^C_{\boldsymbol{Z}}) P(H^C_{\boldsymbol{Z}}|D_e, D_o)$;

Every line of this algorithm is clearly explained and analyzed in [40] and theoretical guarantees are given on how this method provides a better prediction for the post-interventional outcome using both D_o and D_e for multinomial distributions. We do not provide the extended expressions for lines 4,5 as the individual probabilities of line 4 can be estimated as posterior expectations of P(Y |do(X), **W**) from the data, for **W**= **Z**\X. Specifically, under given $H^c_{\boldsymbol{Z}}$, $P(Y|do(X), \boldsymbol{W}) = P(Y|X, \boldsymbol{W})$ and therefore we can use both D_e and D_o for the posterior expectation. In contrast, under $H^{\bar{c}}_{\boldsymbol{Z}}$, we only use D_e.

In this thesis, we will simplify the problem of predicting Y|do(X) when we have $N_o \gg N_e$, in the problem of finding the set **Z*** that optimizes the prediction of Y|do(X). If **Z*** is also an **adjustment set**, we want to use both D_o and D_e to estimate P(Y |do(X), **Z***). Otherwise, we only want to use D_e to estimate P(Y| do(X), **Z***).

Thus, we want to compute $P(H^c_{\boldsymbol{Z}}|D_e, D_o)$ and $P(H^{\bar{c}}_{\boldsymbol{Z}}|D_e, D_o)$ for possible IMBs **Z**. These probabilities tell us both how likely it is that **Z** is an IMB, and if we can include observational data in the estimation of $P(Y|do(X), \boldsymbol{V})$.

Thus we focused on line 3 of "*FindIMB*" algorithm and using Bayes' rule we obtain:

$$P\left(H^c_{\boldsymbol{Z}'} \mid D_e, D_o\right) = \frac{P\left(D_e \mid D_o, H^c_{\boldsymbol{Z}'}\right) P\left(D_o \mid H^c_{\boldsymbol{Z}'}\right) P\left(H^c_{\boldsymbol{Z}'}\right)}{\sum_{\boldsymbol{Z}} \sum_{C=c,\bar{c}} P\left(D_e \mid D_o, H^C_{\boldsymbol{Z}}\right) P\left(D_o \mid H^C_{\boldsymbol{Z}}\right) P\left(H^C_{\boldsymbol{Z}}\right)}. \tag{5.4}$$

$P\left(H^{\bar{c}}_{\mathbf{Z}} \mid D_e, D_o\right)$ similarly derived by replacing each appearance of c with $\bar{c}$ in the numerator.

The denominator is the same for all sets. $P\left(H^c_{\mathbf{Z}'}\right)$ and $P\left(H^{\bar{c}}_{\mathbf{Z}'}\right)$ are our priors that $H^c_{\mathbf{Z}'}$ and $H^{\bar{c}}_{\mathbf{Z}'}$ hold, respectively. We set this to be uniform over both values of C and all **Z**.

- **Our simplified problem is**: We have the observational, D_o, and the experimental data, D_e and we want to use them to estimate the probability that **Z** is an adjustment set(or CMB).

As we can see from Baye's rule e, we can estimate the posterior probabilities for the set of hypotheses $H^c_{\mathbf{Z}'}$ and $H^{\bar{c}}_{\mathbf{Z}'}$ using marginal likelihoods of the experimental and observational data.

Estimating marginal likelihood of observational data: $P\left(D_o \mid H^c_{\mathbf{Z}}\right), P\left(D_o \mid H^{\bar{c}}_{\mathbf{Z}}\right)$:

We just refer to two methods for the calculation of these probabilities but in the formula of Eq.(5.4), it does not make much difference because even for small experimental sample sizes, the part of the marginal likelihood of the experimental data, $P\left(D_e \mid D_o, H^c_{\boldsymbol{Z}'}\right)$ dominates in the product of $P\left(D_e \mid D_o, H^c_{\boldsymbol{Z}'}\right) P\left(D_o \mid H^c_{\boldsymbol{Z}'}\right)$.

These probabilities score how well the observational data fit with the hypotheses $H^c_{\mathbf{Z}}, H^{\bar{c}}_{\mathbf{Z}}$. Express the hypothesis that a set $\mathbf{U}$ is the OMB of Y : Let $H^o_{\mathbf{U}}$ denote this hypothesis; thus, for any $\mathbf{U} \subseteq \mathbf{V} \cup X$, $H^o_{\mathbf{U}}$ is true iff $\mathbf{U}$ is the OMB for Y. Then we can write

$$P\left(D_o \mid H^C_{\mathbf{Z}}\right) = \sum_{\mathbf{U} \subseteq \mathbf{V} \cup X} P\left(D_o \mid H^o_{\mathbf{U}}\right) P\left(H^o_{\mathbf{U}} \mid H^C_{\mathbf{Z}}\right),$$

for $C = c, \bar{c}$. Under $H^c_{\mathbf{Z}}$, Theorem 5.3.3, implies that $P\left(H^o_{\mathbf{U}} \mid H^c_{\mathbf{Z}}\right) = 1$ if $\mathbf{U} = \mathbf{Z}$, and zero otherwise. Under $H^{\bar{c}}_{\mathbf{Z}}$, the IMB is not a CMB. Instead, the IMB has to be a subset of $\mathbf{U}$, therefore $P\left(H^o_{\mathbf{U}} \mid H^C_{\mathbf{Z}}\right) = 0$ for any $\mathbf{Z} \supset \mathbf{U}$. $P\left(D_o \mid H^o_{\mathbf{U}}\right)$ is the marginal likelihood of Y in D_o, under the hypothesis that $\mathbf{U}$ is the data-generating OMB for Y in the observational data. We can obtain this likelihood using a Bayesian scoring algorithm like FGES [Ramsey et al. 2017], by scoring a DAG where Y is a child of variables U [40].

Instead, we can express

$$P\left(H^c_{\boldsymbol{Z}'} \mid D_e, D_o\right) \propto P\left(D_e \mid D_o, H^c_{\boldsymbol{Z}'}\right) P\left(H^c_{\boldsymbol{Z}'} \mid D_o\right)$$

and calculate the probability that the IMB $\mathbf{Z}$ is an adjustment set given the observational data. Due to time and space constraints, we will not analyze the way of doing it. In the work "Learning Adjustment Sets from Observational and Limited Experimental Data " [39], they provide a way to calculate this probability, by considering $H^c_{\boldsymbol{Z}'}$ based on causal graphs that are plausible given D_o. This requires an additional assumption analogous to faithfulness for the adjustment criterion. Specifically, we need to assume that the adjustment sets for (X, Y) are exactly those for which the adjustment criterion holds(*adjustment faithfulness*).

Estimating marginal likelihood of experimental data: $P\left(D_e \mid D_o, H^c_{\boldsymbol{Z}'}\right), P\left(D_e \mid D_o, H^{\bar{c}}_{\boldsymbol{Z}'}\right)$:

The probability $P\left(D_e \mid D_o, H^c_{\boldsymbol{Z}'}\right)$ expresses the probability that you see your experimental data if $\mathbf{Z}$ is an adjustment set and if you have already seen your observational data, D_o.

Let $\mathbf{W} = \mathbf{Z} \backslash X$, and let :

- $\theta_{Y_x|\mathbf{W}}$ be a set of parameters expressing the conditional probabilities for $P(Y \mid do(X), \mathbf{W})$, also called experimental parameters.
- $\theta_{Y|X,\mathbf{W}}$ denote the observational parameters for $P(Y \mid X, \mathbf{W})$.

By integrating over all $\theta_{Y_x|\mathbf{W}}$, we obtain

$$\begin{aligned} &P\left(D_e \mid D_o, H^c_{\mathbf{Z}}\right) = \\ &\int_{Y_x|\mathbf{W}} P\left(D_e \mid \theta_{Y_x|\mathbf{W}}\right) f\left(\theta_{Y_x|\mathbf{W}} \mid D_o, H^c_{\mathbf{Z}}\right) d\theta_{Y_x|\mathbf{W}}, \end{aligned} \tag{5.5}$$

where $P\left(D_e \mid \theta_{Y_x|\mathbf{W}}\right)$ is the outcome of your experimental data (likelihood of your post-intervention outcome) and $f\left(\theta_{Y_x|\mathbf{W}} \mid D_o, H^c_{\mathbf{Z}}\right)$ is the posterior for interventional parameters, $\theta_{Y_x|\mathbf{W}}$, given the observational data when $\mathbf{Z}$ is the IMB and the CMB.
In this case, $P(Y \mid do(X), \mathbf{W}) = P(Y \mid X, \mathbf{W})$, and therefore $f\left(\theta_{Y_x|\mathbf{W}} \mid D_o, H^c_{\mathbf{Z}}\right) = f\left(\theta_{Y|X,\mathbf{W}} \mid D_o\right)$. Eq.(5.5) can then be rewritten in terms of the observational parameters as

$$\begin{aligned} &P\left(D_e \mid D_o, H^c_{\mathbf{Z}}\right) = \\ &\int_{\theta_{Y|X,\mathbf{W}}} P\left(D_e \mid \theta_{Y|X,\mathbf{W}}\right) f\left(\theta_{Y|X,\mathbf{W}} \mid D_o\right) d\theta_{Y|X,\mathbf{W}}. \end{aligned} \tag{5.6}$$

, where $P\left(D_e \mid \theta_{Y|X,\mathbf{w}}\right)$ is the likelihood for your post-intervention outcome (your experimental data) given the observational parameters.

Eq.(5.6) is the marginal likelihood of Y in experimental data, with parameter density $f\left(\theta_{Y|X,\mathbf{w}} \mid D_o\right)$ being equal to the parameter posterior given D_o. In other words, the observational and experimental parameters coincide, under $H^c_{\mathbf{Z}}$. Therefore, D_o gives us a strong "prior" for D_e.

- Under $H^c_{\mathbf{Z}}$, $\theta_{Y_x|\mathbf{w}} = \theta_{Y|X,\mathbf{w}}$

Under $H^{\bar{c}}_{\mathbf{Z}}$, the equality of the observational and experimental parameters does not hold, and we cannot use the $\theta_{Y|X,\mathbf{w}}$ to inform $\theta_{Y_X|\mathbf{w}}$, at least not in a straightforward way. Instead, we model that $f\left(\theta_{Y_X|\mathbf{w}} \mid D_o\right) = f\left(\theta_{Y_X|\mathbf{w}}\right)$. Then $P\left(D_e \mid D_o, H^{\bar{c}}_{\mathbf{Z}}\right)$ corresponds to the marginal likelihood of Y in the experimental data, using a prior that we model as being non-informative:

$$\begin{aligned} &P\left(D_e \mid D_o, H^{\bar{c}}_{\mathbf{Z}}\right) = \\ &\int_{Y_x|\mathbf{w}} P\left(D_e \mid \theta_{Y_x|\mathbf{w}}\right) f\left(\theta_{Y_x|\mathbf{w}} \mid D_o, H^{\bar{c}}_{\mathbf{Z}}\right) d\theta_{Y_x|\mathbf{w}} = \\ &\int_{Y_x|\mathbf{w}} P\left(D_e \mid \theta_{Y_x|\mathbf{w}}\right) f\left(\theta_{Y_x|\mathbf{w}}\right) d\theta_{Y_x|\mathbf{w}}, \end{aligned} \tag{5.7}$$

- Under $H^{\bar{c}}_{\mathbf{Z}}$, $\theta_{Y_x|\mathbf{w}} \neq \theta_{Y|X,\mathbf{w}}$

5.5 Our contribution: Causal Bayesian Feature Selection in cases of binary and ordinal outcomes

As elucidated in our motivational example and expounded upon in Chapter 4, **Causal Feature Selection** is about to find the minimal set (if exists) of pre-treatment covariates that are maximally informative for the post-interventional distribution P(Y| do(X)) using both observational and experimental data.

We translate this problem in the problem of investigating if a set **Z** of pre-treatment covariates is an adjustment set (CMB). In the papers titled "Learning Adjustment Sets from Observational and Limited Experimental Data "[39] and "Causal and Interventional Markov Boundaries"[40] the authors presented results on this problem for discrete variables and multinomial distributions with Dirichlet priors and assumed that the results can be extended to other distributions for which marginal likelihoods can be computed in closed form.

Our contribution is that we propose a way to solve this problem when we do not have conjugate priors and thus closed-form marginals, but for data sets with **mixed explanatory variables** when the response variable is **binary** or **ordinal** using Bayesian Regression methods and both observational and experimental data, without the limitation of distributions with closed-form marginals.

We present a Bayesian method for combining observational (Do) and experimental data (De) to score possible adjustment sets, under the assumption that they come from the same population and they measure. This method utilizes regression models, Markov Chain Monte Carlo (MCMC) sampling, and marginal distributions and proposes a more accurate prediction of Y|do(X) based on the available large samples of Do and limited De. This approach aims to enable causal

feature selection feasible for any data structure by leveraging MCMC sampling, s it allows for the estimation of posterior distributions without relying on the assumption of conjugate priors or closed-form solutions.We provide the *"CFS(Y)"* (algorithm 2) that describes this procedure.

Notation:

- $\theta_e \backsim \theta_{Y_x|\mathbf{W}}$: The parameters of the interventional distributions $P(Y|do(X), \boldsymbol{W})$.
- $\theta_o \backsim \theta_{Y|X,\mathbf{W}}$: The parameters of the observational distribution $P(Y|X, \boldsymbol{W})$.

Algorithm 2: CFS(Y)

Input: $D_\mathbf{o}$, $D_\mathbf{e}$, treatment X, outcome Y, pre-treatment covariates **V**
Output: $H^c_{\boldsymbol{W}}$
For D_o:
1 $var\_subsets \leftarrow Find\_sub(X, \mathbf{V})$;
2 **foreach** *subset **Z** of var_subsets* **do**
3 **for** $number\_of\_samples$ **do**
4 Sample $\boldsymbol{\theta}$ from an uninformative $p(\boldsymbol{\theta})$;
5 Fit regression model $\hat{Y} = f(\mathbf{Z}, \boldsymbol{\theta})$;
6 Compute $P(D_o|\boldsymbol{\theta})$ using Eq. (4.2) or (4.5);
7 Compute marginal likelihood: $\hat{P}(D_o) \approx \sum_{\boldsymbol{\theta}} P(D_o|\boldsymbol{\theta})p(\boldsymbol{\theta})$;
8 marginals[**Z**] $\leftarrow \hat{P}(D_o)$;
9 Sample from the posterior: $\boldsymbol{\theta'} \sim P(\boldsymbol{\theta}|D_\mathbf{o})$ using MCMC sampling, for $j = number\_of\_samples$;
10 Keep traces: $traces[\boldsymbol{Z}] \leftarrow \boldsymbol{\theta'}$;
11 MB(Y) $\leftarrow argmax_\mathbf{Z}$(marginals) ;
For D_e:
12 **foreach** *subset **W** of MB(Y) and* $C = c, \overline{c}$ **do**
13 Compute $\hat{P}(D_e|D_o, H^{C=\overline{c}}_\mathbf{W}) \approx \sum_{\boldsymbol{\theta_e}} P(D_e|\boldsymbol{\theta_e})p(\boldsymbol{\theta_e}), \boldsymbol{\theta_e} \backsim$uninformative;
14 Compute $\hat{P}(D_e|D_o, H^{C=c}_\mathbf{W}) \approx \sum_{\boldsymbol{\theta_o}} P(D_e|\boldsymbol{\theta_o})p(\boldsymbol{\theta_o}), \boldsymbol{\theta_o} \backsim trace[\boldsymbol{W}]$;
15 **if** $P(D_e|D_o, H^{C=c}_\mathbf{W}) > P(D_e|D_o, H^{C=\overline{c}}_\mathbf{W})$ **then**
16 **W**$= IMB = CMB$ and $H^c_{\boldsymbol{W}} = 1$;
  else
17 **W**$= IMB \neq CMB$ and $H^c_{\boldsymbol{W}} = 0$;

In the two perspectives of determining whether a set **W** of pre-treatment covariates serves as an adjustment set (or a $CMB_X(Y)$) for binary and ordinal outcome variables, the overall procedure remains the same. However, the difference lies in the choice of the regression model utilized. Until line 11 of Algorithm 2, the procedure follows the description provided in Section 4.3, incorporating our contribution of identifying the MB(Y). Notably, we retain the *trace* for each subset **Z** (line 10) of pre-treatment covariates, as these traces will be used later for calculating marginal likelihoods in the experimental data (line 14).

Then for each subset of MB(Y), we calculate the probabilities $P(D_e|D_o, H^{C=c}_\mathbf{W})$ and $P(D_e|D_o, H^{C=\overline{c}}_\mathbf{W})$ for possible IMBs **W**. These probabilities tell us both how likely it is that **W** is an IMB, and if we can include observational data in the estimation of P(Y |do(X), **V**) (line 16,17).

Line 13: Estimating marginal likelihood of experimental data: $P\left(D_e \mid D_o, H_{\mathbf{W}}^{\bar{c}}\right)$ under $H_{\mathbf{W}}^{\bar{c}}$.

The probability $P\left(D_e \mid D_o, H_{\mathbf{W}}^{\bar{c}}\right)$ expresses the probability that you see your experimental data if **W** is an IMB but not an adjustment set and if you have already seen your observational data, D_o. Under the $H_{\mathbf{W}}^{\bar{c}}$, the equality of observational and experimental parameters does not hold, so we will use a prior that we model as being non-informative. Eq. (5.7) can not be computed in closed-form because we do not have distributions with conjugate priors. In order to calculate the marginal likelihood of Y in experimental data in this case, we follow the same procedure as we do before to calculate the marginal likelihood in observational data, but at this time using our experimental data:

- A Regression model for each type of data used and we calculate the posterior for each model's parameters using an uninformative prior, as we do not have any information about our D_e.
- We sample from this posterior distribution and we calculate the marginal likelihood of Y in experimental data from Eq.(5.7):

$$P\left(D_e \mid D_o, H_{\mathbf{W}}^{\bar{c}}\right) =$$
$$\int_{\boldsymbol{\theta_e}} P\left(D_e \mid \boldsymbol{\theta_e}\right) f\left(\boldsymbol{\theta_e} \mid D_o, H_{\mathbf{Z}}^{\bar{c}}\right) d\boldsymbol{\theta_e} =$$
$$\int_{\boldsymbol{\theta_e}} P\left(D_e \mid \boldsymbol{\theta_e}\right) f\left(\boldsymbol{\theta_e}\right) d\boldsymbol{\theta_e} = \sum_{\boldsymbol{\theta_e}} P(D_e|\boldsymbol{\theta_e})p(\boldsymbol{\theta_e}), \boldsymbol{\theta_e} \backsim uninformative,$$

Line 14: Estimating marginal likelihood of experimental data: $P\left(D_e \mid D_o, H_{\mathbf{W}}^{c}\right)$ under $H_{\mathbf{W}}^{c}$.

The probability $P\left(D_e \mid D_o, H_{\mathbf{W}}^{c}\right)$ expresses the probability that you see your experimental data if **W** is an adjustment set and if you have already seen your observational data, D_o.

In this case, where IMB=CMB and P(Y|do(X), **W**) = P(Y|X, **W**) and therefore $\boldsymbol{\theta_e} = \boldsymbol{\theta_o}$, we calculate the marginal of the experimental data using as prior, the posterior calculated in observational data for this set **W**. From Eq.(5.6):

$$P\left(D_e \mid D_o, H_{\mathbf{Z}}^{c}\right) =$$
$$\int_{\boldsymbol{\theta_e}} P\left(D_e \mid \boldsymbol{\theta_e}\right) f\left(\boldsymbol{\theta_e} \mid D_o, H_{\mathbf{Z}}^{c}\right) d\boldsymbol{\theta_e} =$$
$$\int_{\boldsymbol{\theta_o}} P\left(D_e \mid \boldsymbol{\theta_o}\right) f\left(\boldsymbol{\theta_o} \mid D_o\right) d\boldsymbol{\theta_o} = \sum_{\boldsymbol{\theta_o}} P(D_e|\boldsymbol{\theta_o})p(\boldsymbol{\theta_o}), \boldsymbol{\theta_o} \backsim trace[\boldsymbol{W}].$$

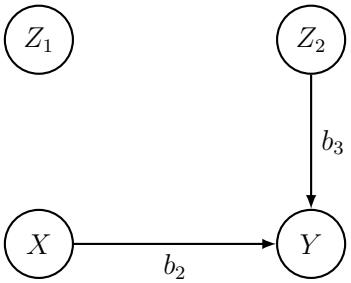
For our understanding, let's continue the illustrative example of section 4.3 (p.39) for a simple binary implementation:

Illustrative Workflow Example: A Simple Binary Implementation

Reminder: Suppose the objective is to identify the minimum set of features that optimally predict a target variable Y. The target variable, in this case, is binary and represents in-hospital mortality, with the value "Yes" indicating death and "No" denoting survival. Additionally, the data set contains three other variables: the treatment variable X (steroids), as well as two continuous variables, lifestyle Z1, and blood pressure Z2.

1. **Data Generating process**
   The real **experimental** data is not yet available and you want to create a simulation of the procedure that you will follow. The implementation workflow is outlined below:

   Assume that the true coefficients will be a = 1 and b = [b2, b3] = [1.4, 1.15], and the real graph is given below. Then the experimental data will be created among the expressions below following the assumptions of Bayesian Logistic Regression for a binary response model:



   - $noise : e \sim N(1, 0)$
   - $Z_1 \sim N(0, 15)$
   - $Z_2 \sim N(0, 10)$
   - $p_{xi} = 0.5, X \sim Bernoulli(0.5)$, Randomized Controlled Trial(RTC)
   - $p_{yi} = expit(a+b_2*X+b_3*Z_3+e), Y \sim Bernoulli(p_{yi})$

2. Assume that we create $N_e = 100$ experimental data and the observational data ($N_o = 1000$) are the samples as in the illustrative example of Sec 4.3 where we have found that the MB(Y) is the set **Z** = (X, Z_2):

   For every subset of MB that includes the treatment X; thus for (X, Z_2) and for (X), compute and compare:

   - For (X, Z_2):

$$P(D_e \mid D_o, H^{\bar{c}}_{\mathbf{Z}}) = \int_a \int_{b_2} \int_{b_3} P(D_{\mathbf{e}}|a, b_2, b_3)p(a)p(b_2)p(b_3)\, da\, db_2\, db_3 = -25.124, \tag{5.8}$$

     with $a, b_2, b_3 \backsim N(0, 100)$.

$$P(D_e \mid D_o, H^{c}_{\mathbf{Z}}) = \int_a \int_{b_2} \int_{b_3} P(D_{\mathbf{e}}|a, b_2, b_3)p(a)p(b_2)p(b_3))\, da\, db_2\, db_3 = -24.547, \tag{5.9}$$

     with $\boldsymbol{\theta_o} = (a, b_2, b_3)$ samples from the posterior $P(\boldsymbol{\theta_o}|D_o) \backsim trace[(X, Z_2)]$ from observational data.

   - For (X):

$$P(D_e \mid D_o, H^{\bar{c}}_{\mathbf{Z}}) = \int_a \int_{b_2} P(D_{\mathbf{e}}|a, b_2)p(a)p(b_2))\, da\, db_2 = -74.75, \tag{5.10}$$

     with $a, b_2 \backsim N(0, 100)$

$$P(D_e \mid D_o, H^c_{\mathbf{Z}}) = \int_a \int_{b_2} P(D_{\mathbf{e}}|a, b_2)p(a)p(b_2))\, da\, db_2 = -113.037, \quad (5.11)$$

with $\boldsymbol{\theta_o} = (a, b_2)$ samples from the posterior $P(\boldsymbol{\theta_o}|D_o) \backsim trace[(X)]$ from observational data.

Under $H^{\bar{c}}_{\mathbf{Z}}$, for (5.8) and (5.10), logistic regression was applied in D_e for the sets (X,Z_2) and (X), with almost flat priors and the likelihood functions, $P(D_{\mathbf{e}}|a, b_2, b_3)$ and $P(D_{\mathbf{e}}|a, b_2)$, were calculated from Eq (2.5) with samples from the prior distribution of D_e as described in Sec(4.3.1) for finding MB(Y).

Under $H^{c}_{\mathbf{Z}}$, for the calculation of marginals (5.9) and (5.11), we use as prior, the samples of the posterior distributions from observational data. Thus, using the trace of every set, (X,Z_2) and (X), we calculate the marginal likelihood of Y in D_e given the observational data and the assumption that the IMB=CMB.

To summarize, set **Z** = (X, Z_2) is IMB and CMB, or an adjustment set, thus for this set, we can use both observational and experimental data to calculate the post-intervention outcome: $P(Y|do(X), \boldsymbol{W}) = P(Y|X, \boldsymbol{W})$, where $\mathbf{W} = \mathbf{Z} \backslash X$.

The same process was applied to the ordinal data using the regression model specifically described in Chapter 4 (cumulative logit model) and the likelihood function for categorical data, Eq(2.10), for calculating the marginal likelihood in experimental data. However, due to space constraints and to maintain a focused analysis, we do not provide an illustrative example in ordinal data too, but we will continue presenting the experiments and results separately for both cases.

5.5.1 Experiments and Results

In this section, we aim to provide further insights into our contribution to causal feature selection methods for determining if a set **Z** of pre-treatment covariates is an adjustment set, and thus we can use both observational and experimental data for predicting the post-intervention outcome $P(Y|do(X), \boldsymbol{V})$.

We will present experiments and corresponding results for mixed data sets, which include both continuous and binary independent variables. In this scenario, we do not assume conjugate priors or closed forms for calculating the posterior distribution. Specifically, we focus on cases where the outcome variable is **binary** or **ordinal** in nature.

Assume that we do not know the graph and you want to predict your post-intervention outcome Y|do(X). We have observational data (D_o, N_o) and experimental data (D_e, N_e) measuring your treatment, steroids (X), your outcome, in-hospital mortality (Y), and two other variables: lifestyle (Z1) and blood pressure (Z2).
Assumptions:

- $N_o \gg N_e$.
- D_o and D_e comes from the same population and have those same variables.

We want to find the set **Z*** that optimizes the prediction of Y|do(X).

If **Z*** is also an adjustment set, we want to use both D_o and D_e to estimate P(Y|do(X), **Z***).

Otherwise, we only want to use D_e to estimate P(Y|do(X), **Z***).

5.5.1.1 Binary target variable.

In this case, the target variable is binary and represents in-hospital mortality, with the value "Yes" indicating death and "No" denoting survival.

Assume that we know that our real data is represented from the graphs below and the true coefficients are a=1 and b=[b_o, b_1, b_2, b_3].

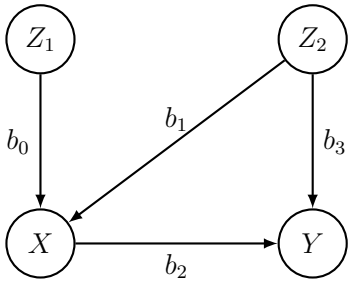


Figure 5.9: Observational distribution illustrates the D_o.

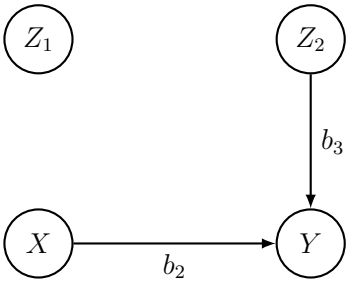


Figure 5.10: Experimental distribution illustrates the D_e.

As we assume $N_o = 1000$ or greater, in almost every case, our algorithm successfully identifies the correct Markov Boundary (MB(Y)). We present our results in scenarios where, after identifying the correct MB, we further investigate each subset of the MB to determine the probability of each subject being both an Interventional and Causal Markov boundary simultaneously.

In our study, we employed two distinct sampling algorithms, Sequential Monte Carlo (SMC) and the No-U-Turn Sampler (NUTS), implemented using two different software packages, PyMC and NumPyro. The objective was to mitigate the potential pitfalls associated with specifying non-informative priors, particularly in situations where the available sample size is severely limited, such as having only 50 samples derived from experimental data.

The outcomes of our investigation illustrate the average performance across various $beta$ parameters and N_e values. Notably, we observed substantial disparities in performance based on the choice of the sampler when dealing with limited experimental samples. Additionally, addressing the limitations associated with non-informative priors is an area of future research we aim to explore.

Scenario 1:
For standard coefficients, a = 1 and beta = [1.3, 1.25, 1.4, 1.15] we present our results for different numbers of experimental sample sizes: 50, 100, 150, 500.

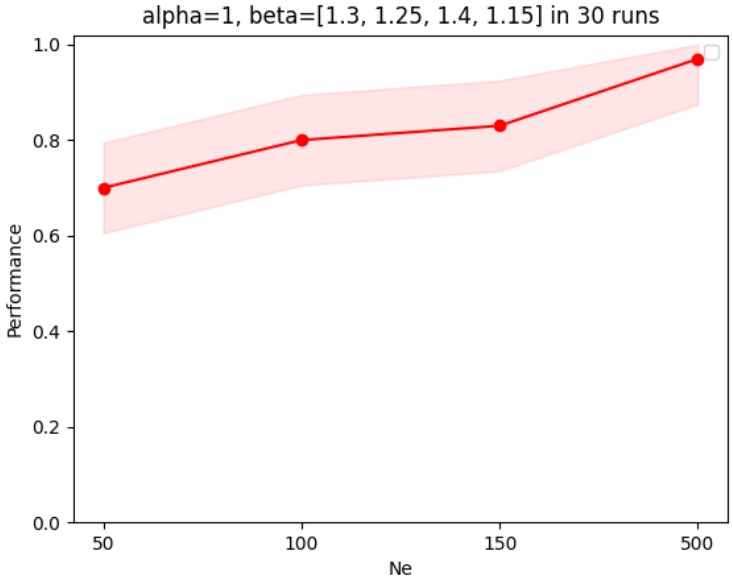


Figure 5.11: Binary outcome: Probability of identifying only the correct adjustment set for different N_e. The shaded area represents the 95% confidence interval.

Scenario 2:
For different "*beta*" coefficients, we present our results for 50 and 100 experimental sample sizes in 10 runs.

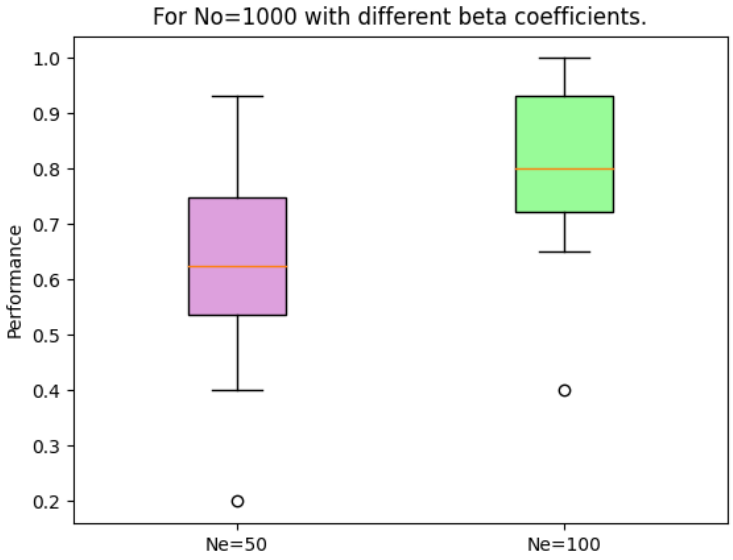


Figure 5.12: Binary outcome: Probability of identifying only the correct adjustment set for different "$beta$" coefficients and $N_e = 50, 100$.

Those cases present the most significant challenges due to the limited number of experimental data available. Conversely, for cases where the number of experimental data exceeds 500, our method consistently demonstrates excellent performance and provides highly accurate results. In the plot Figure 5.12 we embedded cases where the coefficient b_3 or the product $b_1 b_2$ is very small.

5.5.1.2 Ordinal target variable.

In this case, the target variable is ordinal and represents in-hospital mortality in three ordered categories with the values of "0", "1" and "2" indicating the days of survival in the hospital.

Assume that we know that our real data is represented from the graphs Figure 5.9 and Figure 5.10 and the true coefficients are a = [-4, 4] and b=[b_o, b_1, b_2, b_3].

Scenario 1:
For standard coefficients, a = [-4, 4] and beta = [1.3, 1.25, 1.4, 1.15] we present our results for different numbers of experimental sample sizes: 50, 100, 150, 500.

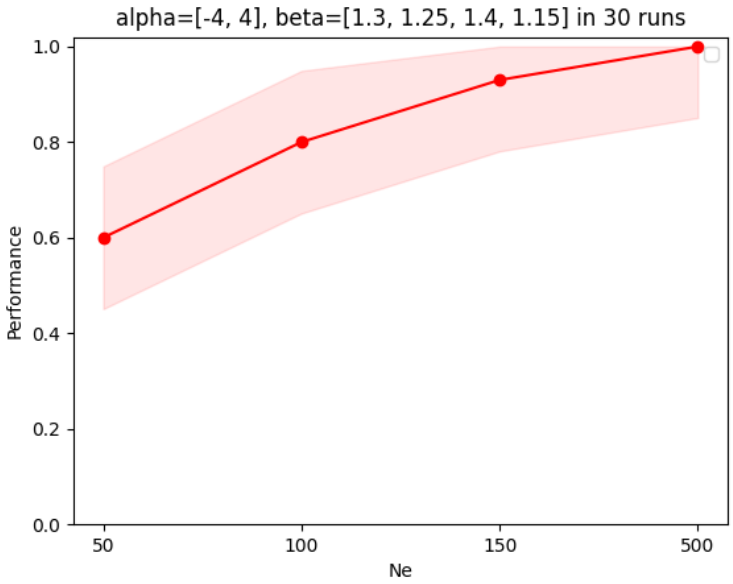


Figure 5.13: Ordinal outcome: Probability of identifying only the correct adjustment set for different N_e. The shaded area represents the 95% confidence interval.

Scenario 2:
For different "*beta*" and "*alpha*" coefficients, we present our results for 50 and 100 experimental sample sizes in 10 runs.

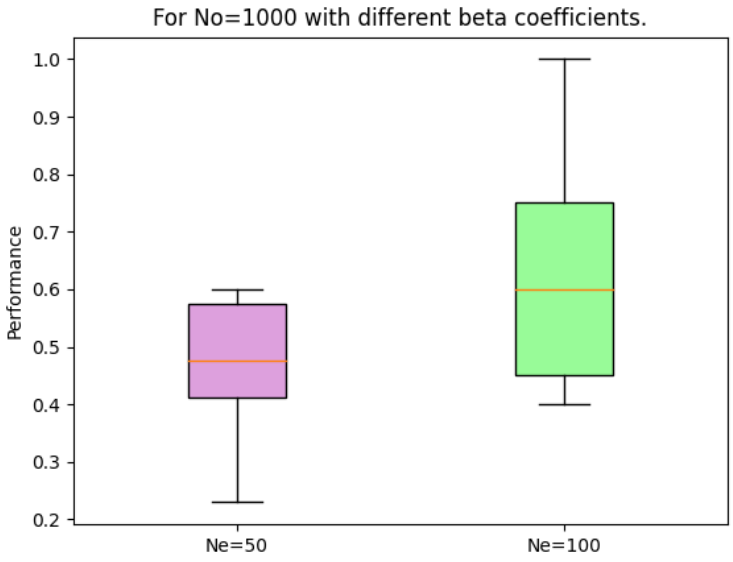


Figure 5.14: Ordinal outcome: Probability of identifying only the correct adjustment set for different "$beta$" coefficients and $N_e = 50, 100$.

In the context of ordinal data, our method exhibits limited performance when confronted with very small experimental sample sizes. This outcome was anticipated, considering that ordinal data poses a more challenging scenario compared to binary data. However, as the effective sample size (N_e) increases, our method demonstrates notably improved performance. We present our results for $N_e = 200, 500$ in Figure 5.15.

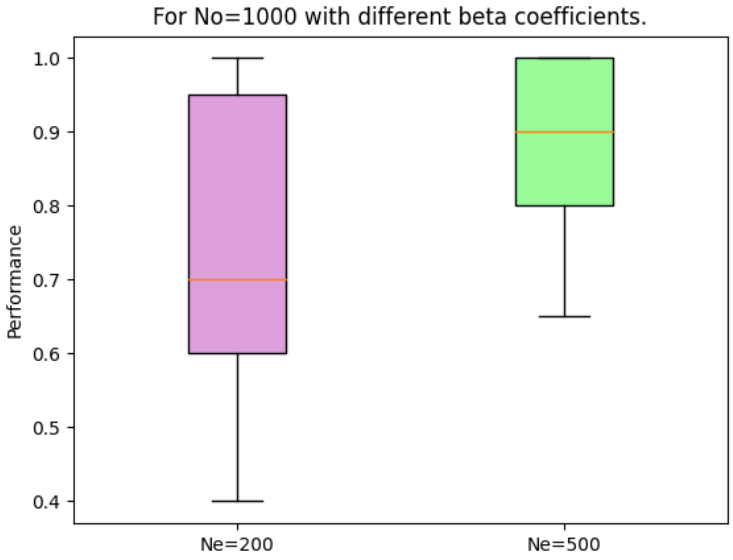


Figure 5.15: Ordinal outcome: Probability of identifying only the correct adjustment set for different "$beta$" coefficients and $N_e = 200, 500$.

5.5.1.3 A more complex example illustrating Feature Selection and Causal Feature selection

Let us consider a scenario wherein three independent variables are present: the treatment variable denoted by X, the lifestyle variable denoted by Z1, and the blood pressure variable denoted by Z2. However, in addition to these variables, there is another variable, Z3, which we are aware is not included in our actual model. Our objective is to determine whether this variable is incorporated into our model or not. Furthermore, it is pertinent to note that the dependent variable has expanded to include seven categories instead of the initial three categories.

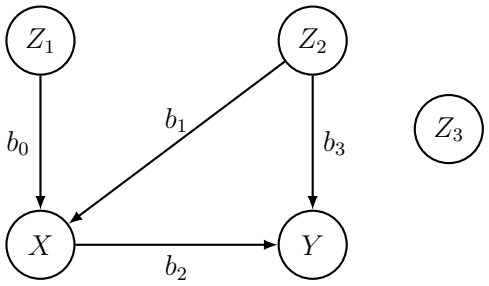


Figure 5.16: Observational distribution illustrates the D_o.

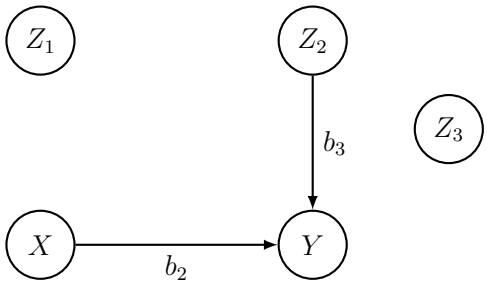


Figure 5.17: Experimental distribution illustrates the D_e.

We run this experiment for two different set of "*alpha*" and "*beta*" coefficients:

- $alpha = [-15, -10, -5, 0, 4, 8]$ and $beta = [1, 2.1, 1.7, 1.2]$ and for $N_o = 1000$ and $N_e = 400$ we have 100% performance in feature selection (finding the correct MB(Y)) and the covariate Z_3 is not in a set of MB. For Causal feature selection, our performance is about 74%. Thus our method finds correct only the true adjustment set in 74% of the cases.

- $alpha = [-6, -3, 5, 15, 17, 20]$ and $beta = [1.2, 1.3, 1.4, 1.7]$ and for $N_o = 1000$ and $N_e = 400$ we have 100% performance in feature selection (finding the correct MB(Y)) and the covariate Z_3 is not in a set of MB. For Causal feature selection, our performance is about 90%. Thus our method finds correct only the true adjustment set in 87% of the cases.

Chapter 6

Conclusions and future extensions

Estimating causal effects from observational data presents challenges due to the presence of confounding factors. While adjusting for an appropriate set of covariates can help mitigate confounding bias, determining this adjustment set is often not feasible using observational data alone. Experimental data offer a solution by enabling unbiased estimation of causal effects. However, the limited sample size of experimental data can result in estimates with high variance. Therefore, a combination of observational and experimental data, when available, can provide more robust and reliable estimates of causal effects.

This thesis aimed to address the challenge of identifying the Markov boundary of a variable Y in the absence of knowledge about the underlying causal graph. Using Bayesian regression methods, we developed a framework that enables the identification of the Markov boundary solely based on the observational distribution. Additionally, we extended the concept of Markov boundaries to predict post-intervention distributions by employing a Bayesian method capable of incorporating both observational and experimental data.

To simplify the problem, we focused on finding the optimal set Z* that enhances the prediction of Y given the intervention do(X). Furthermore, if Z* also serves as an adjustment set, we can leverage both observational (D_o) and experimental (D_e) data to estimate the conditional probability P(Y|do(X), Z*).

We extended the proposed method for estimating causal effects [40] to handle mixed data distributions comprising both continuous and discrete independent variables. While the initial framework was designed for discrete variables and multinomial distributions, we recognized the need to address the challenges posed by mixed data scenarios.

Of particular importance was the inclusion of ordinal data, as this type of data presents unique challenges in predicting post-interventional outcomes. Ordinal variables possess a natural ordering that must be considered when assessing the impact of interventions. However, existing methods often fail to effectively capture the nuances of ordinal data, leading to gaps in the accurate prediction and estimation of causal effects.

Ordinal data are commonly encountered in healthcare, such as patient satisfaction ratings, pain severity scales, or functional disability levels. These ordinal variables provide valuable insights into patient experiences, disease progression, and treatment effectiveness. Therefore, accurately identifying the causal features that influence these ordinal outcomes is critical for improving healthcare decision-making.

By selecting the relevant causal features, healthcare professionals can design targeted interventions and treatments that have the greatest impact on patient outcomes. This enables personalized care and enhances patient satisfaction, quality of life, and overall healthcare outcomes.

Furthermore, accurately predicting post-intervention outcomes in healthcare allows for assessing the effectiveness of different treatment strategies or interventions. It aids in evaluating the benefits and potential risks associated with specific interventions, helping healthcare providers make evidence-based decisions.

Our vision is to extend our method and experiments to encompass complex data structures with a larger number of variables compared to the examples presented in this thesis. Additionally, we aim to broaden the application of our method to scenarios where the experimental data have fewer covariates available, but both observational and experimental data share common variables.

Moreover, in the recent work of Triantafillou, Jabbari, and Cooper [41], they use discrete variables and multinomial Dirichlet distributions and they show that the probability computing using marginal likelihoods is consistent. It will convert to the true data-generating model:

Theorem 6.0.1. *Let D_o, D_e be an observational data set and an experimental data set, respectively, both measuring treatment X, outcome Y, and pre-treatment covariates V, all discrete. Let D_o, D_e contain N_o, N_e cases respectively, sampled from distributions P,$P_{\overline{X}}$ respectively, both strictly positive in the sample limit. Also, let P be a perfect map for an ADMG G. We assume N_o and N_e increase equally without limit. Then the proposed method converges to the data-generating model in the large sample limit:*

$$\lim_{N\to\infty} P(\mathcal{H}^c_{\boldsymbol{Z}}|D_o, D_e) = \begin{cases} 1 & if\ \boldsymbol{Z}\ is\ an\ adjustment\ set \\ & for\ X\ and\ Y \\ 0 & ,otherwise \end{cases} \tag{6.1}$$

It would be highly desirable to provide formal proof that is not limited to discrete pre-treatment covariates.

Furthermore, because of the log odds transformation in regression models, it will be meaningful to simulate prior predictive simulations to check if our uninformative prior maintain to be uninformative after the logit transformation and consider the hidden dangers of specifying Non-informative priors.

Finally, our proposed method holds great potential for application in real-world data sets, and we eagerly anticipate its implementation in practical scenarios.

Bibliography

[1] Statnikov A, Lytkin NIand Lemeire J, and Aliferis CF. Algorithms for Discovery of Multiple Markov Boundaries. In: *Journal of machine learning research: JMLR, 14, 499–566.* (2013).

[2] Camil Băncioiu and Remus Brad. Analyzing Markov Boundary Discovery Algorithms in Ideal Conditions Using the d-Separation Criterion. In: *Algorithms* 15.4 (2022). ISSN: 1999-4893. DOI: 10.3390/a15040105. URL: https://www.mdpi.com/1999-4893/15/4/105.

[3] Brady Neal - Causal Inference. *4.6 - The Backdoor Adjustment*. 2020. URL: https://www.youtube.com/watch?v=U1S8Rq8IcrY (visited on 05/08/2023).

[4] Max Chickering. "Statistically Efficient Greedy Equivalence Search." en. In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. ISSN: 2640-3498. PMLR, Aug. 2020, pp. 241–249. URL: https://proceedings.mlr.press/v124/chickering20a.html.

[5] D. R. Cox. The Regression Analysis of Binary Sequences. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 20.2 (1958). Publisher: [Royal Statistical Society, Wiley], pp. 215–242. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2983890.

[6] Frederick Eberhardt, Patrik Hoyer, and Richard Scheines. "Combining Experiments to Discover Linear Cyclic Models with Latent Variables." en. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. JMLR Workshop and Conference Proceedings, Mar. 2010, pp. 185–192. URL: https://proceedings.mlr.press/v9/eberhardt10a.html.

[7] Tiago M. Fragoso and Francisco Louzada Neto. Bayesian model averaging: A systematic review and conceptual classification. In: *International Statistical Review* 86.1 (Apr. 2018). arXiv:1509.08864 [stat], pp. 1–28. ISSN: 03067734. DOI: 10.1111/insr.12243. URL: http://arxiv.org/abs/1509.08864.

[8] Marco Edward Gorelli. *MarcoGorelli/pydataglobal-21*. original-date: 2021-10-23T21:35:35Z. July 2022. URL: https://github.com/MarcoGorelli/pydataglobal-21 (visited on 07/03/2023).

[9] Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical Criteria for Efficient Total Effect Estimation via Adjustment in Causal Linear Models. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.2 (Apr. 2022). arXiv:1907.02435 [math, stat], pp. 579–599. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/rssb.12451. URL: http://arxiv.org/abs/1907.02435 (visited on 05/24/2023).

[10] *How do I interpret the coefficients in an ordinal logistic regression in R? | R FAQ*. URL: https://stats.oarc.ucla.edu/r/faq/ologit-coefficients/.

[11] Yu K et al. Mining Markov Blankets Without Causal Sufficiency. In: *IEEE Trans Neural Netw Learn Syst.* (2018). ISSN: 6333-6347. DOI: 10.1109/TNNLS.2018.2828982. URL: https://www.mdpi.com/1999-4893/15/4/105.

[12] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. "Removing Hidden Confounding by Experimental Grounding." In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://papers.nips.cc/paper_files/paper/2018/hash/566f0ea4f6c2e947f36795c8f58ba901-Abstract.html.

[13] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Google-Books-ID: 7dzpHCHzNQ4C. MIT Press, July 2009. ISBN: 978-0-262-01319-2.

[14] John Lafferty, Han Liu, and Larry Wasserman. *Statistical Machine Learning*. CRC Press, 2009.

[15] *Learn PyMC & Bayesian modeling — PyMC v5.4.0 documentation*. URL: https://www.pymc.io/projects/docs/en/v5.4.0/learn.html.

[16] Osvaldo A. Martin. *Bayesian Analysis with Python (Second edition)*. original-date: 2016-11-24T16:56:42Z. June 2023. URL: https://github.com/aloctavodia/BAP (visited on 07/03/2023).

[17] Peter McCullagh. Regression Models for Ordinal Data. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 42.2 (1980), pp. 109–142. ISSN: 00359246. URL: http://www.jstor.org/stable/2984952.

[18] Arash Mehrjou, Reshad Hosseini, and Babak Nadjar Araabi. Improved Bayesian information criterion for mixture model selection. en. In: *Pattern Recognition Letters* 69 (Jan. 2016), pp. 22–27. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2015.10.004. URL: https://www.sciencedirect.com/science/article/pii/S016786551500344X.

[19] Joris M. Mooij, Sara Magliacane, and Tom Claassen. *Joint Causal Inference from Multiple Contexts*. arXiv:1611.10351 [cs, stat]. Aug. 2020. DOI: 10.48550/arXiv.1611.10351. URL: http://arxiv.org/abs/1611.10351.

[20] Yi Mu, Isaac See, and Jonathan R. Edwards. Bayesian model averaging: improved variable selection for matched case-control studies. In: *Epidemiology, biostatistics and public health* 16.2 (2019), e13048. ISSN: 2282-2305. DOI: 10.2427/13048. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6879006/ (visited on 05/25/2023).

[21] Brady Neal. Introduction to Causal Inference. en. In: ().

[22] Judea Pearl. *Causality: models, reasoning and inference*. eng. 2nd ed. OCLC: 800430518. Cambridge: Cambridge University Press, 2009. ISBN: 978-0-521-89560-6.

[23] Judea Pearl. *Causality: Models, reasoning, and inference*. Causality: Models, reasoning, and inference. Pages: xvi, 384. New York, NY, US: Cambridge University Press, 2000. ISBN: 978-0-521-77362-1.

[24] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988. ISBN: 1558604790.

[25] Emilija Perković et al. *Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs*. arXiv:1606.06903 [math, stat]. June 2018. DOI: 10.48550/arXiv.1606.06903. URL: http://arxiv.org/abs/1606.06903 (visited on 05/24/2023).

[26] Evan Rosenman et al. *Propensity Score Methods for Merging Observational and Experimental Datasets*. arXiv:1804.07863 [stat]. Oct. 2018. DOI: 10.48550/arXiv.1804.07863. URL: http://arxiv.org/abs/1804.07863 (visited on 05/24/2023).

[27] Andrea Rotnitzky and Ezequiel Smucler. *Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models*. arXiv:1912.00306 [cs, math, stat]. Dec. 2019. DOI: 10.48550/arXiv.1912.00306. URL: http://arxiv.org/abs/1912.00306 (visited on 05/24/2023).

[28] John W. Seaman and James D. Stamey. Hidden Dangers of Specifying Noninformative Priors. In: *The American Statistician* 66.2 (2012). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 77–84. ISSN: 0003-1305. URL: https://www.jstor.org/stable/23339464 (visited on 07/03/2023).

[29] Ilya Shpitser and Judea Pearl. Identification of Conditional Interventional Distributions. en. In: ().

[30] Ilya Shpitser and Judea Pearl. Identification of Conditional Interventional Distributions. en. In: ().

[31] Ilya Shpitser and Judea Pearl. Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models. en. In: ().

[32] Ilya Shpitser and Jin Tian. "On identifying causal effects." en. In: ed. by Rina Dechter, Hector Geffner, and Joseph Y. Halpern. College Publications, Feb. 2010. ISBN: 978-1-904987-65-9. URL: https://eprints.soton.ac.uk/350591/.

[33] Ilya Shpitser, Tyler VanderWeele, and James M. Robins. *On the Validity of Covariate Adjustment for Estimating Causal Effects*. arXiv:1203.3515 [cs, stat]. Mar. 2012. DOI: 10.48550/arXiv.1203.3515. URL: http://arxiv.org/abs/1203.3515.

[34] Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. *Efficient adjustment sets in causal graphical models with hidden variables*. arXiv:2004.10521 [cs, math, stat]. May 2020. DOI: 10.48550/arXiv.2004.10521. URL: http://arxiv.org/abs/2004.10521 (visited on 05/24/2023).

[35] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Vol. 81. Journal Abbreviation: Causation, Prediction, and Search Publication Title: Causation, Prediction, and Search. Jan. 1993. ISBN: 978-1-4612-7650-0. DOI: 10.1007/978-1-4612-2748-9.

[36] Jin Tian and Judea Pearl. On the Identification of Causal Effects. en. In: ().

[37] Jin Tian and Ilya Shpitser. On Identifying Causal Effects. en. In: ().

[38] Sofia Triantafillou. "Integrative causal analysis of heterogeneous data sets." en. Accepted: 2015-09-03T07:31:22Z Artwork Medium: LEATHER Interview Medium: LEATHER Journal Abbreviation: Ολοκληρωνένη αιτιακή ανάλυση ετερογενών συνόλων δεδομένων. Διδακτορική Διατριβή. Πανεπιστήμιο Κρήτης. Σχολή Θετικών και Τεχνολογικών Επιστημών. Τμήμα Επιστήμης Υπολογιστών, 2015. DOI: 10.12681/eadd/36134. URL: http://hdl.handle.net/10442/hedi/36134.

[39] Sofia Triantafillou and Greg Cooper. Learning Adjustment Sets from Observational and Limited Experimental Data. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.11 (May 2021). Number: 11, pp. 9940–9948. ISSN: 2374-3468. DOI: 10.1609/aaai.v35i11.17194. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17194.

[40] Sofia Triantafillou, Fattaneh Jabbari, and Gregory F. Cooper. "Causal and interventional Markov boundaries." In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Uncertainty in Artificial Intelligence. ISSN: 2640-3498. PMLR, Dec. 1, 2021, pp. 1434–1443. URL: https://proceedings.mlr.press/v161/triantafillou21a.html.

[41] Sofia Triantafillou, Fattaneh Jabbari, and Gregory F. Cooper. "Learning Treatment Effects from Observational and Experimental Data." en. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, Apr. 2023, pp. 7126–7146. URL: https://proceedings.mlr.press/v206/triantafillou23a.html.

[42] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based Causal Discovery from Multiple Interventions over Overlapping Variable Sets. In: *Journal of Machine Learning Research* 16.66 (2015), pp. 2147–2205. ISSN: 1533-7928. URL: http://jmlr.org/papers/v16/triantafillou15a.html.

[43] Ioannis Tsamardinos, Constantin Aliferis, and Alexander Statnikov. "Algorithms for Large Scale Markov Blanket Discovery." In: Jan. 2003, pp. 376–381.

[44] Tyler J. VanderWeele and Ilya Shpitser. A new criterion for confounder selection. eng. In: *Biometrics* 67.4 (Dec. 2011), pp. 1406–1413. ISSN: 1541-0420. DOI: 10.1111/j.1541-0420.2011.01619.x.

[45] Tian-Zuo Wang et al. "Cost-effectively Identifying Causal Effects When Only Response Variable is Observable." en. In: *Proceedings of the 37th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, Nov. 2020, pp. 10060–10069. URL: https://proceedings.mlr.press/v119/wang20w.html.

[46] Eric P Xing and Fan Guo. Bayesian Feature Selection. en. In: (). URL: https://www.cs.cmu.edu/~epxing/Class/10801-07/lectures/note3.pdf.

[47] Kui Yu et al. Mining Markov Blankets Without Causal Sufficiency. eng. In: *IEEE transactions on neural networks and learning systems* 29.12 (Dec. 2018), pp. 6333–6347. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2018.2828982.

[48] Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. "Constructing Separators and Adjustment Sets in Ancestral Graphs." English. In: *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction - Volume 1274*. CEUR-WS.org, July 2014, pp. 11–24. URL: https://research.uni-luebeck.de/en/publications/constructing-separators-and-adjustment-sets-in-ancestral-graphs.

[49] Jiji Zhang. "Causal Inference and Reasoning in Causally Insu-cient Systems." In: 2006. URL: https://www.semanticscholar.org/paper/Causal-Inference-and-Reasoning-in-Causally-Systems-Zhang/9c54c1c3d5a0225bda1bd332b52ece0547b1e732.

Appendix A

First appendix

A.1 Proofs

Assumptions:

- X (treatment) causes Y (outcome)
- All variables **V** are pre-treatment.

Definition A.1.1. *(Backdoor Set). **Z** is a backdoor set for X, Y if and only if **Z** m-separates X and Y in $G_{\underline{X}}$.*

Lemma 1 *Let Z be a set for which $P(Y \mid do(X), Z)$ is identifiable from $P(Y \mid X, Z)$, then Z is a subset of a backdoor set. [proof provided in the supplementary of [40]]*

Lemma 2 *Every causal Markov boundary is a backdoor set.*

Proof. Assume that **Z** is a causal Markov boundary, but **W** is not a backdoor set. Since $P(Y|do(X), \mathbf{W})$ is identifiable, by *Lemma 1* **W** is a subset of backdoor set $\mathbf{W} \cup \mathbf{Q}$, where $\mathbf{Q} \subseteq (\mathbf{V} \setminus \mathbf{W})$. Since by assumption **W** is not a backdoor set, **Q** is not the empty set (i.e., **W** is a proper subset of a backdoor set). We will show that $P(Y|do(X, \mathbf{W}, \mathbf{Q})) \neq P(Y|do(X, \mathbf{W}))$. To show that, we only need to show that **Q** is not independent of **W** in $G_{\overline{X}}$. Since **W** is not a backdoor set, there exists a backdoor path from X to Y that is m–connecting given **W**, but blocked given $\mathbf{W} \cup \mathbf{Q}$. Thus, some $Q \in \mathbf{Q}$ is a non-collider on that path, therefore **Q** are not independent with Y given **W**. Hence, $P(Y|do(X, \mathbf{W}, \mathbf{Q})) \neq P(Y|do(X, \mathbf{W}))$ and therefore **Z** does not satisfy Condition (2), and **Z** is not a causal Markov boundary (Contradiction).

Theorem A.1.1. *We assume that P_x and $\mathcal{G}_{\overline{X}}$ are faithful to each other. Every causal Markov boundary* **Z** *of an outcome variable Y w.r.t a treatment variable X is a subset of the Markov boundary* $\mathrm{MB}(Y)$.

Proof. We will show this by contradiction. Specifically, we will show that any set **Z** that includes variables **Q** not in the Markov boundary of Y cannot satisfy one of the Conditions (2) or (3) of the causal Markov boundary.

Assume that **Z** is a causal Markov boundary for Y with respect to X and let $\mathbf{W} = \mathbf{Z} \backslash X$. Let $\mathbf{Q} = \mathbf{W} \backslash \mathrm{MB}(Y)$ be the non-empty subset of **W** that is not a part of the Markov boundary of Y.

If there exists no $Q \in \mathbf{Q}$ that has an m-connecting path $Q\pi_{QY}Y$ to Y given $\mathbf{W}\backslash Q$, then $\mathbf{Q} \perp\!\!\!\perp Y \mid (\mathbf{W}\backslash\mathbf{Q})$ in $\mathcal{G}_{\overline{X}}$. Conditioning on X cannot open any paths from X to Y; therefore, $\mathbf{Q} \perp\!\!\!\perp Y \mid X, (\mathbf{W}\backslash\mathbf{Q})$ in $\mathcal{G}_{\overline{X}}$. Then by Rule 1 of the do-calculus [Pearl, 2000], $P(Y \mid \text{do}(X), \mathbf{W}) = P(Y \mid \text{do}(X), \mathbf{W}\backslash Q)$, and $\mathbf{Z}$ does not satisfy Condition (3) of the causal Markov boundary definition (Contradiction).

If there exists a $Q \in (\mathbf{W}\backslash \text{MB}(Y))$ that has an m-connecting path $Q\pi_{QY}Y$ with Y given $\mathbf{Z}\backslash Q$, then by Lemma 1.7, there exists a variable W in $\text{MB}(Y)\backslash\mathbf{Z}$ such that $\mathbf{Z}\cup W$ is also a backdoor set, and $W \not\perp\!\!\!\perp Y \mid X, \mathbf{Z}$ in $\mathcal{G}_{\overline{X}}$. Then $P(Y \mid do(X), \mathbf{Z}, W) \neq P(Y \mid do(X), \mathbf{Z})$. Thus, $\mathbf{Z}$ does not satisfy Condition (2) of the Causal Markov boundary definition (Contradiction).

Thus, $\mathbf{Z}$ cannot include any variables that are not in the Markov boundary of Y.

Theorem A.1.2. *Let $\mathcal{G}$ be a SMCM over X, Y, $\mathbf{V}$ with $\mathbf{V}$ occurring before X and Y. Let $\mathbf{Z} \subseteq \mathbf{V} \cup X$ be the IMB of Y relative to X. If $\mathbf{Z}$ is a causal Markov boundary, then $\text{MB}(Y) = \mathbf{Z}$.*

Proof. $\text{MB}_X(Y) \subseteq \text{MB}(Y)$, so we need to show that $\text{MB}(Y) \subseteq \text{MB}_X(Y)$ when $\text{MB}_X(Y) \in \mathbf{CMB}_X(Y)$. Assume that $\mathbf{Z}$ is both the $\text{MB}_X(Y)$ and a causal Markov boundary, but there exists a variable Q in $\mathbf{Z}$ that is not in $\text{MB}(Y)$. Then Q is reachable from Y through a bidirected path in $\mathcal{G}$ but not in $\mathcal{G}_{\overline{X}}$. Since $\mathcal{G}$ and $\mathcal{G}_{\overline{X}}$ only differ in edges that are into X, this path must be going through an edge that is incoming into X. Thus, $\mathcal{G}$ includes a bidirected path $Y \leftrightarrow \cdots \leftrightarrow X$, and every variable on this path is in $\text{MB}_X(Y) = \mathbf{Z}$. But then $\mathbf{Z}\backslash X$ cannot be a backdoor set, and $\mathbf{Z}$ cannot be a causal Markov boundary. Contradiction. Thus, the Markov boundary of Y cannot include any more variables than $\mathbf{Z}$.