Bayesian Causal Inference: Nonparametric and Parametric Methods

Ahmet Emre Belge

Bocconi University

June 2024

Supervisor: Prof. Dr. Antonio Lijoi

Contents

1 Introduction

Bayesian causal inference is an area of research that combines the concepts of causal inference and Bayesian statistics. Causal inference concentrates on studying and exploring cause-and-effect relationships between different variables, which distinguishes it from more traditional statistical methods that primarily focus on correlation and association between the variables. The fundamental teaching of *correlation does not imply causation* gives us a hint about the fact that distinguishing a mere association between variables from a cause-and-effect relationship has always been a crucial task in statistical studies.

One of the most important frameworks in causal inference is the Potential Outcomes Framework [Rubin, 1974, Splawa-Neyman et al., 1990] where, given a treatment, such as a drug, we estimate its causal effects on an individual by evaluating the difference between the outcome in the case where the individual receives the treatment and the outcome in the case where the treatment is not administered. It is easy to notice the problem here: There is not a way through which we can have information about both cases for the same individual. This is called the *Fundamental Problem of Causal Inference*, leading to the conclusion that it is impossible to directly observe causal effects. Nevertheless, there are various methods through which we are able to overcome this problem and effectively estimate causal effects, some of which will be studied in the following sections. The main causes of problems in evaluating causal effects are confounders, i.e., variables that influence both the treatment assignment and the outcome of interest. For example, if we are trying to estimate the effects of a voluntary after-school program on students' test scores, a confounder would be the inherent motivation and study performance of the students. When comparing the treatment and control groups, if we do not account for the inherent motivation of the students, we might get biased results. This is because motivation (confounder) affects positively both participation in the program (treatment) and the test scores (outcome). In other words, students that are in the treatment group might perform better regardless of the effects of the program, as they are already more motivated compared to the students that are not included in the treatment group. Methods to overcome the problem of confounders will be discussed in this paper.

Although the incorporation of Bayesian methods into causal inference is not a recent phenomenon, it is indisputable that these methods have been becoming more popular thanks to the increase in the amount of data available and to the advancements in computational power. By using Bayesian methods in causal inference, we are able to incorporate our existing knowledge and beliefs about the problem in the form of prior distributions; these beliefs are then updated as new data become available, providing us with the posterior distribution. This iterative process of updating provides a natural way to incorporate uncertainty into causal estimates, offering a more nuanced view than traditional frequentist methods. For example, in the context of evaluating the effects of a voluntary after-school program on students' test scores, Bayesian methods can use prior information about the relationship between motivation, participation in similar programs, and academic achievement to better adjust for the confounding effect of motivation.

In this paper, we will study nonparametric and parametric Bayesian models in the context of causal inference, exploring various theoretical and computational obstacles to overcome. We will then touch on the subject of mediation analysis, a statistical approach used to understand how an independent variable influences an outcome variable through one or more intermediary variables called mediators, along with an application using Bayesian nonparametric causal inference methods.

2 Nonparametric Bayesian Methods in Causal Inference

2.1 Parametric vs Nonparametric Models

In the context of statistical modeling, when we discuss parametric models, we refer to a family of models that can be described using a finite number of parameters, which does not increase as the number of data points increases. Using formal notation, we can define the distribution/model as $\phi = \{\phi_\theta : \theta \in \Theta \subset R^p\}$. A basic example of a parametric model is the multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ with $\{\boldsymbol{\mu} \in R^p, \Sigma \in R^{p \times p}\}$. The number of parameters, in this case equal to $(p + \frac{p(p+1)}{2})$, with p being the dimension of the data points and $\frac{p(p+1)}{2}$ being the number of unique elements in a $p \times p$ symmetric matrix, remains the same as we gain more data. Parametric models have their advantages, such as the fact that they are easier to implement, may require less data, and require less computational power. However, they make lots of assumptions about the form of the underlying distribution, leading to relatively inflexible models that sometimes might not be enough to model the data at hand.

On the other hand, nonparametric models make little-to-no assumptions about the underlying functional form, leading to more flexibility in the modeling and robustness against misspecification of a parametric statistical model [Müller et al., 2015]. An intuitive way to see nonparametric models is that instead of trying to estimate a fixed number of parameters given a fixed functional form, they try to estimate the function itself, with the infinite-dimensionality of their parameter space providing greater structural flexibility to better capture the underlying form of the data. In these models, the infinite-dimensional parameters are usually functions such as distributions in the case of the Dirichlet process or conditional trends such as mean regression functions in the case of Gaussian processes.

The particularity in the usage of nonparametric models in Bayesian causal inference is that, opposed to classical nonparametrics where infinite-dimensional parameters are left as a nuisance parameter while the focus is on estimating some finite-dimensional parameter of interest, the Bayesian approach completes the model with a prior on such an infinite-dimensional parameter, including a complete probabilistic description of all relevant uncertainties [Müller et al., 2015]. The inherent flexibility of nonparametric models, combined with Bayesian updating, results in powerful models that play an important role in the world of causal inference. In the rest of this section, we discuss some of the most relevant Bayesian nonparametric models.

2.2 Gaussian Processes

In statistical analysis, Gaussian processes are usually implemented for regression. In particular, we define a Gaussian process as a prior over possible regression functions that output a value $g(\boldsymbol{x}_i)$ given covariates $x_{i1}, x_{i2}, ..., x_{ip}$.

The following definition is largely based on Linero and Antonelli (2022): Consider the following semiparametric regression model (combining a nonparametric model $g(X_i)$ and a parametric error term):

$$Y_i = g(X_i) + \epsilon_i \tag{1}$$

with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. In this regression model, the parameter is the function g. We say that an arbitrary function $g : R^p \to R$ is a Gaussian process if, for any finite number of points $\boldsymbol{x}_i \in R^p$, we have:

$$(g(\boldsymbol{x}_1), g(\boldsymbol{x}_2), ..., g(\boldsymbol{x}_n)) \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \Sigma) \tag{2}$$

with $\boldsymbol{\mu} = (m(\boldsymbol{x}_1), m(\boldsymbol{x}_2), ..., m(\boldsymbol{x}_n))^T$ and $\Sigma_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where $K : R^p \times R^p \to R$, is a positive definite kernel function. In other words, g is a Gaussian process with mean function $m(\boldsymbol{x})$ and covariance function $K(\boldsymbol{x}, \boldsymbol{x}')$ if the vector constructed with values of g evaluated at the points $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, and it is denoted as $g \sim GP(m, K)$ [Linero and Antonelli, 2022, Müller et al., 2015].

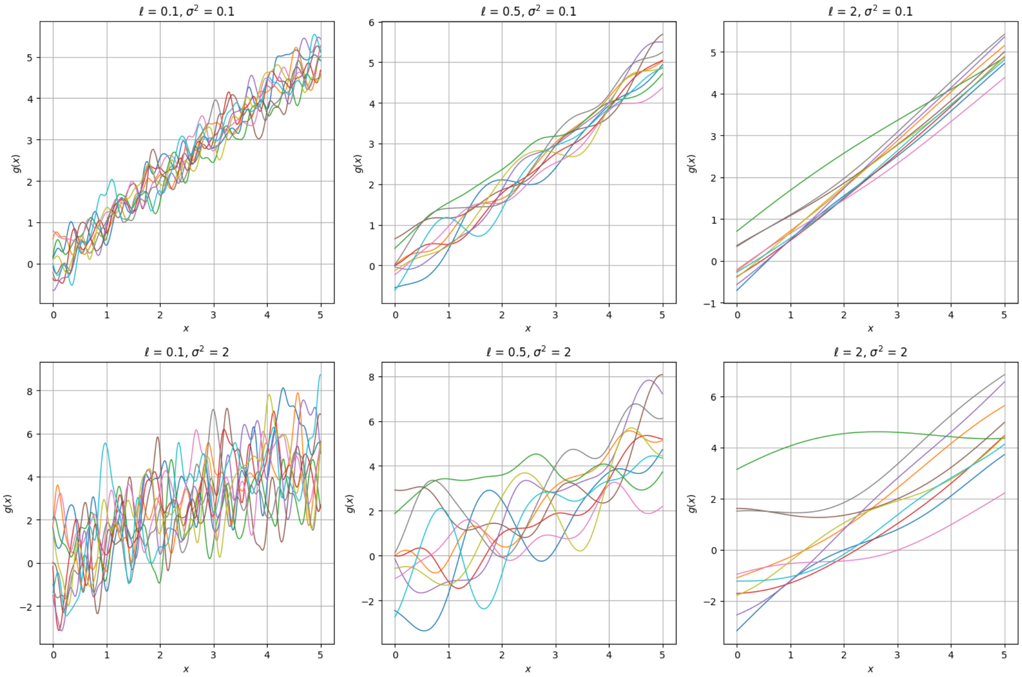
The mean function represents our prior information about the expected value of the output at point $\boldsymbol{x}$. If we have prior observations about the relationship between the covariates and the output, we can, for instance, define the mean function as a linear model of the form $m(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$, where $\boldsymbol{\beta}$ might be a vector of coefficients encoding our knowledge of how the different covariates are weighted in the output.

The covariance function, instead, represents our prior knowledge about how strongly values of the function g at different input values $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ should be correlated. It directly shapes the smoothness and the overall form of g; the more correlated $g(\boldsymbol{x})$ and $g(\boldsymbol{x} + \boldsymbol{\epsilon})$

are, the smoother the function will be. An important thing to note about the covariance function is that we cannot accept an arbitrary function as a covariance function [Rasmussen and Williams, 2006]. As it is the rule for the covariance matrix Σ of a multivariate normal distribution, the covariance matrix that we obtain using this function has to be positive semi-definite for the covariance function to be valid. Functions with this property are called *kernel* functions [Linero and Antonelli, 2022]. One of the most common examples of covariance (or kernel) functions is the squared exponential kernel, defined as:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\ell^2}\right) \tag{3}$$

In this case, the parameter ℓ represents the distance required between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for them to be nearly uncorrelated, with higher values of ℓ leading to smoother functions. On the other hand, $\sigma^2 = K(\boldsymbol{x}, \boldsymbol{x})$ determines how concentrated the samples are around $m(\boldsymbol{x})$, with higher values leading to more variation and to values possibly farther away from $m(\boldsymbol{x})$. Such hyperparameters can be estimated from data using various methods, one of which is empirical Bayes [Linero and Antonelli, 2022]. In the figure below, samples from a GP with mean function $m(x) = x$ and the squared exponential kernel with different values of σ^2 and ℓ are shown:



Through these mean and covariance functions, we then calculate the mean vector and covariance matrix using $\boldsymbol{x}_i$, which is essentially how we are able to model quite flexibly the relationship between the covariates and the output using a Gaussian process.

Apart from the intuitive incorporation of prior knowledge, Gaussian processes also have conjugacy properties that play a role in their popularity in Bayesian causal inference. For the semiparametric regression model (1), the posterior that we obtain after we update our GP model with the data is also a Gaussian process. In notation, if we have $g \sim GP(m, K)$, then $[g|X, Y] \sim GP(m^*, K^*)$. m^* and K^* are calculated using the formulae below [Linero and Antonelli, 2022]:

$$K^*(\boldsymbol{x}, \boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}) - K(\boldsymbol{x}, X)(K(X, X) + \sigma^2 I)^{-1} K(X, \boldsymbol{x})$$
$$m^*(\boldsymbol{x}) = m(\boldsymbol{x}) + K(\boldsymbol{x}, X)(K(X, X) + \sigma^2 I)^{-1}(Y - m(X))$$

Here, Y is an N-dimensional vector representing the observed responses, $\boldsymbol{x}$ is a p-dimensional vector of covariates and X is an $N \times p$ matrix where each row is an observation of the p covariates. Each $K(X, X)$ is a matrix with $(j, k)^{\text{th}}$ entry equal to $K(\boldsymbol{x}_j, \boldsymbol{x}'_k)$. How $K(\boldsymbol{x}, X)$ and $K(X, X)$ can be visualized as a vector and a matrix, respectively, is shown below:

$$K(\boldsymbol{x}, X) = \begin{bmatrix} K(\boldsymbol{x}, X_1) \\ K(\boldsymbol{x}, X_2) \\ \vdots \\ K(\boldsymbol{x}, X_N) \end{bmatrix} \quad \text{and} \quad K(X, X) = \begin{bmatrix} K(X_1, X_1) & \cdots & K(X_1, X_N) \\ K(X_2, X_1) & \cdots & K(X_2, X_N) \\ \vdots & \ddots & \vdots \\ K(X_N, X_1) & \cdots & K(X_N, X_N) \end{bmatrix}$$

It is important to note that this particular conjugacy property of Gaussian processes is only valid for the semiparametric regression model (1), although Gaussian processes can be modified to be used with other models.

The main practical difficulties of Bayesian inference through Gaussian processes lie in their high dimensionality and computational complexity. In particular, inverting the $N \times N$

matrix $K(X, X) + \sigma^2 I$ requires $O(N^3)$ computations, which is a challenging complexity even for $N = 10000$ [Linero and Antonelli, 2022]. As for some of the existing solutions to this problem, we can mention low-rank approximation techniques [Banerjee et al., 2007, Cressie and Johannesson, 2008] and low-rank and sparsity inducing kernels [Datta et al., 2016, Katzfuss and Guinness, 2021, Zhang et al., 2015].

2.3 Bayesian Additive Regression Trees (BART)

Before defining the BART model [Chipman et al., 2010], it might be beneficial to discuss the Bayesian Classification and Regression Tree (CART) model [Chipman et al., 1998], which is used to build a BART model by ensembling. In the context of a regression model with covariates $x_{i1}, x_{i2}, ..., x_{in}$, a tree T is a partition of the input (covariate) space into rectangular regions R_i through thresholds on the covariates. Each of these regions corresponds to one of the terminal nodes of the tree, T. Associated with each terminal node is a value μ_i, which represents the mean response for the observations falling within that region. Given T and M, we define the single regression tree model:

$$Y = g(\boldsymbol{x}|T, M) + \epsilon, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{4}$$

with g being the model that assigns a mean response μ_i to $\boldsymbol{x}$ based on the region R_i that $\boldsymbol{x}$ falls into. Under this model, the conditional expectation of the output Y is equal to μ_i, i.e.:

$$E[Y|\boldsymbol{x}] = g(\boldsymbol{x}|T, M) = \mu_i$$

Having defined what a single regression tree is, we can now define BART. As mentioned above, a BART model is built as the sum of regression trees g in (4) and is an ensemble model. In notation:

$$Y = \sum_{j=1}^{m} g(\boldsymbol{x}|T_j, M_j) + \epsilon, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{5}$$

An important property of BART is that it sets regularizing priors on each of the trees T_j, their terminal node values M_j and σ. This is done to prevent some trees from having an overwhelming effect on the outcome, thus limiting the advantages of additive representations in modeling. There are various priors that we can use to achieve this property, one of which is the T_j prior described in Chipman et al. (2010) which models the probability of a node to be non-terminal at depth d as $\alpha(1+d)^{-\beta}$ with $a \in (0,1), \beta \in [0, \infty)$. Furthermore, it specifies the distribution to pick the splitting variable at each interior node and the distribution of the splitting rule assignment for each node conditional on the splitting variable, both of which are uniform priors. In an example given by Chipman et al. (2010), with $\alpha = 0.95$ and $\beta = 2$, this prior assigns trees with 1, 2, 3, 4, and 5 or more terminal nodes with prior probabilities of 0.05, 0.55, 0.28, 0.09, and 0.03, effectively encouraging shallow threes for an additive model, although it can be flexibly modified to allow for more complex trees when working with a small amount of trees.

As a shallow tree of depth 3 will not be complex enough to effectively build a model, we can think of each tree as a weak learner that gets ensembled to build a powerful and flexible model. To fit the model, BART employs a specialized Bayesian backfitting Markov Chain Monte Carlo (MCMC) method based on Friedman et al. (2000). This method is similar to another kind of ensemble model, gradient boosting, as it iteratively adjusts successive residuals in a similar way. The key difference is that while gradient boosting strengthens weak trees, BART weakens trees using a prior and then utilizes Bayesian backfitting on a set number of trees. Essentially, we can see BART as a Bayesian nonparametric model with relatively restrictive priors that fits a parameter-rich model [Müller et al., 2015].

Another important aspect of BART is its relationship with Gaussian processes. In fact, a BART model conditioned on the tree structure is a Gaussian process with $m(\boldsymbol{x}) = 0$ and $K(\boldsymbol{x}, \boldsymbol{x}') = \sigma_g^2 N(\boldsymbol{x}, \boldsymbol{x}')$ where $N(\boldsymbol{x}, \boldsymbol{x}')$ represents the fraction of trees where the inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ are assigned to the same terminal node and σ_g^2 denotes the prior variance of the parameters associated with these leaf nodes. This property may allow us to interpret BART as a Gaussian process where the covariance function is learned nonparametrically [Linero and Antonelli, 2022].

The main difference between these two models is the fact that while a Gaussian process can have interactions of all orders (depending on the kernel), a BART model will be conditioned to have low-order interactions (at most third order for a tree of depth 3), as a result of the regularizing prior limiting the depth and therefore the order of the interactions. Here we can see that, compared to Gaussian processes, BART models impose more restrictions on the prior. However, the restriction to low-order interactions plays an important part in the success of BART in Bayesian causal inference. This is because in real-world scenarios, the interactions that are of interest usually do not go beyond third-order interactions, allowing us to effectively apply BART in a variety of cases without losing modeling power [Linero and Antonelli, 2022]. Furthermore, this property also makes BART more feasible from a computational point of view, qualifying it as a more preferable solution with respect to Gaussian processes when computational resources might be scarce.

One downside of this property and another difference between BART and Gaussian processes is that as BART is a sum of decision trees, its response surface, might not be smooth enough for some applications, such as regression discontinuity design, where the smoothness of g is a desirable property [Linero and Antonelli, 2022].

2.4 Dirichlet Process and Dirichlet Process Mixture

2.4.1 Dirichlet Process

As the basis for the Dirichlet process mixture model, we introduce the Dirichlet process [Ferguson, 1973]. One key difference between the Dirichlet process and the previous methods that we discussed is that Gaussian processes and BART are primarily used for regression problems or for function estimation, whereas the Dirichlet process is mainly utilized for estimating probability distributions or density functions. In a regression problem, our aim is to determine the function that regulates the relationship between an input x and the output y, and we fix priors over the regression function, such as GPs and BART, which are priors over functions. In a density estimation scenario, instead, we try to infer the underlying distribution of an observed i.i.d. sample and fix priors over the distribution. If we intend to apply Bayesian inference to the density estimation problem, then we will need to have a prior model of the underlying distribution, and that is exactly where the Dirichlet process comes into play.

Before defining the Dirichlet process, we define the notion of σ-algebra and the Dirichlet distribution: Let $M > 0$ be a constant and G_0 a probability measure on a σ-algebra $\mathscr{S}$. A σ-algebra $\mathscr{S}$ on any set S is a collection of measurable subsets of S that satisfy the following properties:

1. $\emptyset, S \in \mathscr{S}$
2. $A \in \mathscr{S} \Rightarrow A^{\mathsf{c}} \in \mathscr{S}$
3. $A_i \in \mathscr{S},\ i \in N \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathscr{S}$

The Dirichlet distribution is a multivariate probability distribution that can be considered as a multivariate generalization of the beta distribution. The Dirichlet distribution of order (dimension) $K \leq 2$ is parametrized by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$ with $\alpha_i > 0$ and has the following probability density function:

$$f(x_1, x_2, ..., x_K|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\Gamma(\alpha_i)$ is the Gamma function defined as $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$. Furthermore, the support of the Dirichlet distribution is the set of all K-dimensional vectors $\boldsymbol{x} = (x_1, x_2, ..., x_K)$ such that for all $i \in \{1, 2, ..., K\}$ we have $x_i \in [0, 1]$ and $\sum_{i=1}^{K} x_i = 1$.

We can now formally define the Dirichlet process, largely following the definition from Müller et al. (2015): A Dirichlet process with parameters (M, G_0) is a stochastic process whose realization G is a probability measure over $\mathscr{S}$ which assigns probability $G(B)$ to every set $B \in \mathscr{S}$ such that for any measurable finite partition $\{B\}_{i=1}^{n}$ we have:

$$[G(B_1), G(B_2), ..., G(B_n)] \sim Dirichlet(MG_0(B_1), MG_0(B_2), ..., MG_0(B_n))$$

that is, the joint distribution of the vector $[G(B_1), G(B_2), ..., G(B_n)]$ is the Dirichlet distribution with parameters $(MG_0(B_1), MG_0(B_2), ..., MG_0(B_n))$. The distribution G_0 is called the *base distribution* of the process and can be thought of as the distribution around which the process samples distributions, i.e., $E[G(B)] = G_0(B)$. The parameter M can be called the *concentration*, *total mass*, or *precision* parameter, and it affects how concentrated G will be around G_0 with higher values leading to higher concentrations. As $M \to \infty$ the process essentially becomes G_0 [Müller et al., 2015].

A particular property of the Dirichlet process is that even if G_0 is continuous, G will be discrete almost surely. This property allows us to express the sampled distribution as $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot)$, i.e., as a weighted sum of point masses represented by the Dirac measure at that point with probability weights $w_h, h \in \mathbb{N}$.

Another important property of the DP is that it has what is called a *large weak support*. Meaning that under mild regularity conditions, a random distribution G sampled from a DP with base distribution G_0 can approximate weakly any distribution H with the same support as G_0 (i.e., G converges to H in distribution).

Given the discrete nature of the DP, we can define one of the methods through which we can construct the DP and realizations of it, called *Stick Breaking Construction*. With this method, we construct G by generating i.i.d. samples from G_0, which will be the locations m_h that we can observe below in (6), and by sampling w_h as a fraction of $\{1 - \sum_{\ell<h} w_\ell\}$. That is, we sample w_h as a fraction of what is left after we distribute probabilities to the preceding $h - 1$ locations, as if we were breaking a stick of length 1 that represents the total probability to be distributed. At $h = 1$, $w_{h=1}$ is simply sampled from the entire "stick" of length 1.

Formally, we define each probability weight as $w_h = v_h \prod_{\ell<h}(1-v_\ell)$ with $v_i \overset{\text{iid}}{\sim} Beta(1, M)$ independent from $m_h \overset{\text{iid}}{\sim} G_0$, leading to:

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot) \tag{6}$$

which is a random probability distribution sampled from a $DP(M, G_0)$. A result of this construction is that if we have $G \sim DP(M, G_0)$ and we sample a point m and a weight W with $m \sim G_0$ and $W \sim Beta(1, M)$ (with all of them independent) then:

$$W\delta_m(\cdot) + (1 - W)G(\cdot) \sim DP(M, G_0)$$

that is, $W\delta_m(\cdot) + (1 - W)G(\cdot)$ also follows a $DP(M, G_0)$.

A final important property of the Dirichlet process is its conditioning property. Given a measurable set A with $G_0(A) > 0$, then the restriction of G to A, $G|A(B) = G(B|A)$, is also a Dirichlet process with parameters M and $G_0|A$ and is independent of $G(A)$. This property can be extended to multiple sets, locally splitting the DP into multiple independent DPs [Müller et al., 2015].

The posterior distribution for the prior model $y_1, y_2, ..., y_n|G \overset{\text{iid}}{\sim} G,\ G \sim DP(M, G_0)$ is still a Dirichlet process. The intuition behind the posterior distribution of a DP prior is that at each observed data point y_i, a point mass is added to the prior base measure G_0 in the posterior. Intuitively, the posterior base measure becomes the weighted average of the prior base measure G_0 and $\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}(\cdot)$, and we increment by n the precision parameter to $M + n$. Formally, given the above prior model, the posterior distribution of G is [Ferguson, 1973]:

$$G|y_1, y_2, ..., y_n \sim DP(M + n, \frac{M}{M + n}G_0 + \frac{1}{M + n}\sum_{i=1}^{n} \delta_{y_i}) \tag{7}$$

with $DP(M + n, \frac{M}{M+n}G_0 + \frac{1}{M+n}\sum_{i=1}^{n} \delta_{y_i})$ being the posterior Dirichlet process.

2.4.2 Dirichlet Process Mixture

Having laid the fundamentals of the Dirichlet process, we introduce the Dirichlet process mixture model, which has been so far the main application of the Dirichlet process in Bayesian causal inference.

A main limitation of the DP is that it generates distributions that are almost surely discrete, which makes it problematic for continuous density estimation. The DPMM aims to overcome this problem by using a DP prior as the mixing distribution in a mixture model of simple parametric distributions [Escobar and West, 1994, Ferguson, 1983]. We define the following hierarchical structure of the DPMM:

$$\begin{aligned} y_i|\theta_i &\sim f_{\theta_i} \\ \theta_i|G &\sim G \\ G &\sim DP(M, G_0) \\ f_G(y) &= \int f_\theta(y) dG(\theta) \end{aligned}$$

The observed data points y_i are distributed according to f_{θ_i}, a continuous distribution parametrized by $\theta_i \in \Theta$. The parameters θ_i are distributed according to the random probability distribution G sampled from $DP(M, G_0)$. Given G and f_θ, we define the mixture of f_θ with respect to G with p.d.f. $f_G(y) = \int f_\theta(y) dG(\theta)$. A common choice for f_{θ_i} is $N(\mu_i, \sigma_i^2)$ with $\theta_i = (\mu_i, \sigma_i^2)$.

An implication of the hierarchical structure defined above is that the Dirichlet process mixture model generates a probability model on clusters [Müller et al., 2015]. Clusters are formed when multiple data points share the same parameters. Given the discrete nature of the Dirichlet process leading to the almost surely discreteness of G, there is a positive probability that multiple θ_i values sampled from G will be the same. This means that there is a positive probability that multiple data points will share the same θ_i value, forming a cluster where f_{θ_i} is the distribution governing the data points in that cluster. Through the parameters of the DP, we can probabilistically adjust how many clusters are created and whether a data point is assigned to an existing cluster or a new cluster is created. An important property of the Dirichlet process mixture model is the *rich-get-richer* dynamic that stems from the predictive structure induced by the DP. This property ensures that the probability of assigning a data point to a cluster is proportional to the current size of that cluster. On the other hand, the probability that a new cluster is created is positively influenced by the parameter M and the number of data points already observed, with the possibility of infinite clusters.

The ability to adaptively model the (possibly infinite) number of clusters is an important property of the DPMM that is crucial in cases where we might not have information about the number of latent classes in the data. This can be adjusted according to the complexity of the data and potentially grow as the number of data points grows, reflecting the nonparametric nature of this method.

3 Parametric Bayesian Methods in Causal Inference

In this section, we study various parametric regression methods applied in causal inference contexts. We dive deeper into examples of priors for Bayesian linear and logistic models for estimating causal effects, followed by a theoretical overview of the Bayesian Lasso.

3.1 Bayesian Linear Model with AR1 Priors

In the context of Bayesian causal inference within point-treatment settings (settings where the treatment is administered once) involving multiple dose levels, the use of autoregressive (AR1) priors presents a refined approach for modeling causal dose effects as presented in Oganisian and Roy (2021).

Consider a specific causal inference scenario where we have a treatment consisting of K dose levels $A_i \in \{0, 1, 2, ..., K\}$ with $A_i = 0$ indicating no treatment level. We define an indicator function $A_{ik} = I(A_i = k)$ for the assignment of individual i to dose $k = \{1, ..., K\}$ and we denote by $Y_{A=k}$ the potential outcome for an individual receiving the dose $A = k$. Also note that the doses are ordered, with $k-1$ being a lower dose than k. Considering also the pre-treatment variables L_i, which represent characteristics of the study participants that are measured before the administration of the treatment, this structure leads to the below linear outcome model:

$$E[Y_i|A_i, L_i] = \theta_0 + L_i'\beta + \sum_{k=1}^{K} \theta_k A_{ik} \tag{8}$$

where the aim is to estimate the causal incremental dose effect curve $\Psi(k) = E[Y_{A=k}] - E[Y_{A=k-1}]$ as a function of the dose k. Under several assumptions that are essential for attributing observed outcome differences to the effects of changing treatment levels rather than to confounding factors, which will be studied in the following section, we can state the following identity: $\Psi(i) = \theta_i - \theta_{i-1}, \quad i = \{2, ..., K\}$ with $\Psi(1) = \theta_1$.

Our aim is to find a suitable prior for the distribution of θ_i, $i \in 1, ..., K$ that allows us to easily integrate our prior knowledge into the model. A first naive approach for the joint prior of $\theta_{1:K}$ might be modeling each θ_i independently, leading to the joint distribution $p(\theta_{1:K}) = \prod_{k=1}^{K} p(\theta_k)$. A common choice for $p(\theta_k)$ is $N(\mu_k, \sigma_k^2)$. However, in this context, we might have prior information leading us to believe that the incremental dose effect of neighboring levels might be correlated. With this expertise at hand, we can implement a more useful dependent prior that can be factorized as follows:

$$p(\theta_{1:K}) = p(\theta_1)p(\theta_2|\theta_1) \prod_{k=3}^{K} p(\theta_k|\theta_{k-1}, \theta_{k-2})$$

We specify each term as:

$$\begin{aligned}\theta_1 &\sim N(\mu_1, \tau_1)\\ \theta_2|\theta_1 &\sim N(2\theta_1, \tau_2)\\ \theta_k|\theta_{k-1}, \theta_{k-2} &\sim N(2\theta_{k-1} - \theta_{k-2}, \tau_k), \quad k > 2\end{aligned}$$

The above is a second-order autoregressive model for θ_k, $k > 2$ as the parameters of the distribution of θ_k depend on the preceding 2 levels, θ_{k-1} and θ_{k-2}. Furthermore, Oganisian and Roy (2021) state that one can either directly specify the hyperparameters μ_1 and $\tau_{1:K}$ or specify hyperpriors (prior distributions on the hyperparameters of another prior model) for them. In the following paragraphs, we show that the second-order autoregressive model for θ_k induces a first-order autoregressive (AR1) prior on the incremental dose effect curve $\Psi(k)$.

As we condition on θ_{k-1} In the last line of the above AR2 model, we can treat it as a constant and subtract it from θ_k. Through this process, we also modify the conditional distribution of θ_k from $N(2\theta_{k-1} - \theta_{k-2}, \tau_k)$ to $N(\theta_{k-1} - \theta_{k-2}, \tau_k)$ as below:

$$\theta_k - \theta_{k-1}|\theta_{k-1}, \theta_{k-2} \sim N(\theta_{k-1} - \theta_{k-2}, \tau_k)$$

Using the definition of $\Psi(k)$, we can then substitute $\theta_k - \theta_{k-1}$ with $\Psi(k)$ to obtain the induced prior on $\Psi(k)$:

$$\begin{aligned}\Psi(k)|\Psi(k-1) &\sim N(\Psi(k-1), \tau_k), \quad k > 1 \qquad (9)\\ \Psi(1) &\sim N(\mu_1, \tau_1)\end{aligned}$$

We see that the prior distribution of $\Psi(k)$ for $k > 1$ (9) only depends on $\Psi(k-1)$, demonstrating that a first-order autoregressive prior is induced on $\Psi(k)$ through the AR2 model for θ_k [Oganisian and Roy, 2021].

The prior in (9) incorporates into the model the prior belief that the effect of increasing the dose level should not be too different from the effect of the previous increase. The AR1 prior is particularly effective in cases where we would like to regularize the MLE estimates, which are volatile when data is scarce. It is both flexible enough to follow the posterior inference with enough data and to provide shrinkage towards the prior in less optimal cases.

A common heuristic approach to decreasing sample sizes in higher doses, as mentioned in Oganisian and Roy (2021), is pooling patients together at higher doses, such as K and $K-1$. Although this approach may increase stability, it begs the assumption that there are no real differences between $\Psi(K)$ and $\Psi(K-1)$, limiting the ability to distinguish between the two cases. The AR1 prior can therefore be seen as a middle ground between assuming independence between priors and pooling dose groups. It maintains the distinction between $\Psi(K)$ and $\Psi(K-1)$ while stabilizing the estimates through the autoregressive structure that induces correlation between dose levels, avoiding the strong assumptions of pooling and the instability of complete independence, such as with the sole use of MLE, in insufficient sample sizes.

3.2 Bayesian Logistic Model with Partial Pooling Priors

In another example covered by Oganisian and Roy (2021), we consider a logistic regression model in a study with binary treatment $A_i \in \{0,1\}$ and binary outcome $Y_i \in \{0,1\}$. In this model, we relate the covariates L to the mean $E[Y|A,L]$ through the sigmoid function, $\sigma(z) = \frac{1}{1+e^{-x}}$, which maps a real-valued number to the (0, 1) interval. As we are in a binary case, we can identify the mean with the probability of obtaining the outcome $Y=1$.

An estimand of interest in this setting might be the casual odds ratio at a given level v of a q-dimensional subset V of the pre-treatment covariates L. This ratio is defined as:

$$\Psi(v) = \frac{E[Y|A=1,V=v]/(1-E[Y|A=1,V=v])}{E[Y|A=0,V=v]/(1-E[Y|A=0,V=v])} \tag{10}$$

The odds, defined as $\frac{p}{1-p}$ for an outcome that happens with probability p, provides a measure between 0 and ∞ of the likelihood of that particular outcome. The causal odds ratio compares the odds of an outcome occurring under two different conditions, typically between a treatment group $A = 1$ and a control group $A = 0$. In our case, it can be interpreted as follows:

- If $\Psi(v) > 1$: the treatment increases the odds of the outcome at level v of V.
- If $\Psi(v) = 1$: the treatment does not affect the odds of the outcome at level v of V.
- If $\Psi(v) < 1$: the treatment decreases the odds of the outcome at level v of V.

Denoting the remaining pre-treatment covariates as $W = L \backslash V$, we define the following logistic regression model:

$$E[Y|A = a, V, W] = \sigma \left\{ \beta_w' W + \beta_v' V + (\theta_0 + \theta_{1:q}' V) A \right\} \tag{11}$$

In the above model, we include intercepts for W and V to account for the effects of the covariates on the outcome that do not stem from the treatment. In Oganisian and Roy (2021), a case with $V \in \{0, 1\}^4$ is evaluated, V being a vector of indicators for $q = 4$ ethnicities: Black, Asian, Hispanic, Native American; with White taken as the reference ethnicity against which the treatment effects for the other ethnicities are compared.

With this information at hand, we can see that θ_0 is the baseline treatment effect and $\theta_{1:q}' V$, which will result in θ_j, $j \in \{1, 2, 3, 4\}$, represents the change in treatment effect given that an individual is of ethnicity j. A possible problem in these settings is the sparsity of data for certain ethnicities such as Hispanic and Native American. As mentioned in Oganisian and Roy (2021), some common approaches to this problem include the pooling of these categories to estimate a single odds ratio or the exclusion of them. These two extremes might not be desirable, and as it was the case with the AR1 prior in the linear model, we can build a useful prior that could act as a middle ground between these approaches.

We consider the prior assumption that the treatment effects across the different ethnicities, $\{\theta_0, \theta_0 + \theta_1, ..., \theta_0 + \theta_4\}$, are distributed normally according to a common overall effect μ and variance τ. We can denote this as $\theta_0 \sim N(\mu, \tau)$ for the reference group and $\theta_0 + \theta_j | \theta_0 \sim N(\mu, \tau)$ for the other ethnicities, which is equivalent to $\theta_j | \theta_0 \sim N(\mu - \theta_0, \tau)$. We integrate our prior assumptions into the model by defining the below joint prior [Oganisian and Roy, 2021]:

$$p(\theta_{0:4}|\mu, \tau) = N(\theta_0|\mu, \tau) \prod_{j=1}^{4} N(\theta_j|\mu - \theta_0, \tau) \tag{12}$$

With this prior, the posterior inference of the treatment effects for categories with sufficient data will not be too conditioned by our prior assumptions, allowing us to smoothly incorporate new information into the model. On the other hand, for categories with a smaller amount of data, which would otherwise be pooled together or left out of the inference, the treatment effect is pushed towards the common overall effect μ in order to limit the variance stemming from the absence of data, with lower values of τ leading to stronger shrinkage towards μ [Oganisian and Roy, 2021]. The structure that we obtain with this prior is called partial pooling, as we allow each category to maintain its own estimates for the treatment effects while still sharing information across subgroups by allowing the estimates to be influenced by the overall trends observed across all subgroups. It is important to note that, to proceed with inference using this model, priors for μ, τ, and the other regression coefficients must be specified as outlined in Oganisian and Roy (2021).

3.3 The Bayesian Lasso

For the last parametric method of this paper, we switch to a more abstract setting and study the theory behind the Bayesian Lasso [Park and Casella, 2008]. In a more classical setting, the Lasso [Tibshirani, 1996] is used to induce sparsity in the parameters of a linear regression problem (13) by estimating them through an L1-norm-constrained least squares method (14):

$$\boldsymbol{y} = \mu\mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{13}$$

$$\min_{\boldsymbol{\beta}}(\tilde{\boldsymbol{y}} - X\boldsymbol{\beta})^T(\tilde{\boldsymbol{y}} - X\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j| \tag{14}$$

with μ being the intercept, X matrix of standardized regressors, $\boldsymbol{\epsilon}$ vector of i.i.d. normal errors with mean 0, $\boldsymbol{y}$ vector of responses, $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$ vector of centered responses with $\bar{y}$ denoting the mean of the observed responses and $\lambda \geq 0$ tuning parameter of the penalty term $\lambda\sum_{j=1}^{p}|\beta_j|$.

The main motivation for the Bayesian lasso comes from the observation of Tibshirani (1996), which suggested that the lasso estimates can be interpreted as posterior mode estimates when the parameters have i.i.d. Laplace priors [Park and Casella, 2008] such as the one shown below:

$$p(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p}\frac{\lambda}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right) \tag{15}$$

We can see that each prior is a Laplace distribution with location parameter $\mu = 0$ and scale parameter $b = \sqrt{\frac{\sigma^2}{\lambda}}$. We also specify a noninformative and scale-invariant prior on σ^2 as $p(\sigma^2) = \frac{1}{\sigma^2}$, called the Jeffreys prior. It is pointed out in Park and Casella (2008) that conditioning on σ^2 is important in order to guarantee a unimodal full posterior.

In order to sample from the joint posterior distribution of the parameters, we implement the Gibbs Sampler, a MCMC sampling method. Using a Laplace prior for the regression parameters β_j directly would make the posterior sampling computationally challenging due to the non-conjugacy of the Laplace distribution with the normal likelihood of these parameters. The Gibbs Sampler for the Bayesian Lasso overcomes this problem by representing the Laplace distribution as a scale distribution of normals with an exponential mixing density [Park and Casella, 2008]:

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}}exp(-\frac{z^2}{2s})\frac{a^2}{2}exp(-\frac{a^2s}{2})ds \tag{16}$$

Following this, in Park and Casella (2008), the hierarchical representation of the Bayesian Lasso is suggested as below:

$$\begin{aligned}
\boldsymbol{y} \mid \mu, X, \boldsymbol{\beta}, \sigma^2 &\sim N_n(\mu \mathbf{1}_n + X\beta, \sigma^2 I_n) \\
\boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim N_p(\mathbf{0}_{\boldsymbol{p}}, \sigma^2 D_\tau) \\
D_\tau &= \text{diag}(\tau_1^2, \ldots, \tau_p^2) \\
\sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim p(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} exp(-\frac{\lambda^2 \tau_j^2}{2}) d\tau_j^2 \\
\sigma^2, \tau_1^2, \ldots, \tau_p^2 &> 0
\end{aligned} \tag{17}$$

After integrating out $\tau_{1:p}^2$ in the conditional prior for $\boldsymbol{\beta}$, we achieve the form specified in (15), but with less computational burden. It is suggested that an independent and flat prior might be given to μ and the Jeffreys prior for σ^2, although any inverse-Gamma prior would also work for conjugacy purposes.

The Bayesian Lasso also provides additional methods to choose the regularization parameter λ, on top of the existing methods implemented in the classical Lasso, such as cross-validation [Hastie et al., 2001]. Two of the most important methods, as mentioned in Park and Casella (2008), are *Empirical Bayes by Marginal Maximum Likelihood* and *Hyperpriors for the Lasso Parameter*. With empirical Bayes [Casella, 2001], we iteratively estimate λ using a Monte Carlo Expectation-Maximization algorithm, adjusting it with each iteration based on previous Gibbs sampler runs. We derive the initial λ estimate from the least squares procedure, and while successive estimates may not converge fully, they tend to approximate the true maximum likelihood estimate. On the other hand, hyperpriors introduce a diffuse prior for λ^2 (a prior that assigns similar probabilities to a wide range of values), allowing easy extension of the Gibbs sampler thanks to the resulting conjugacy. With the proper specification of the hyperprior, λ can join the other parameters in the Gibbs sampler, leading to what we can call an *augmented* Gibbs sampler.

The Bayesian Lasso offers a powerful alternative to classical Lasso in cases where the modeling may benefit from prior information and uncertainty quantification for the parameters, as it automatically provides interval estimates for all point estimates, including the error variance [Park and Casella, 2008]. By treating coefficients as random variables with prior distributions, we are able to directly obtain credible intervals from their posterior distributions. On the other hand, obtaining interval estimates with the classical Lasso involves complex and often unreliable methods to correct for bias and sparsity-induced issues in standard error calculation. Furthermore, the Bayesian Lasso benefits from its ease of extension into generalized linear models, where it may be computationally competitive relative to its classical counterpart, and the availability of a more robust version with a Huber-type loss, the *Huberized* Lasso [Park and Casella, 2008].

4 Estimands, Assumptions, and Challenges

Having completed the overview of some of the most important Bayesian modeling tools, we now move on to discussing some fundamental concepts, goals, and challenges in causal inference. We will start this section by defining the causal estimands that are the most relevant in causal inference problems. We will then go over the identifying assumptions under which our causal estimands are well-defined and the causal inference methods to estimate them are valid. Finally, we will end this section by talking about some practical challenges in estimating causal estimands that stem from our choice of priors and the relationship between the assignment mechanism and outcome model.

4.1 Causal Estimands

Consider a binary treatment scenario with treatment levels $a \in \{0, 1\}$ and a sample of size N taken from a target population. Given an individual i, we denote by $Y_i(a)$ the outcome that the individual would have if treatment a is administered. In our binary case, we are restricted to $Y_i(0)$ and $Y_i(1)$ for a certain individual. The data we observe is composed of the treatment exposures $A = (A_1, A_2, ..., A_N)$, outcomes $Y = (Y_1(A_1), Y_2(A_2), ..., Y_N(A_N))$

and covariates (or confounders) $X = (X_1, X_2, ..., X_N)$ with $X_i \in R^p$, p being the number of covariates. The fundamental challenge in causal inference, which we also mentioned in the introduction, is to make inferences about the distribution of $\{Y_i(0), Y_i(1)\}$ having only access to $A_i, Y_i(A_i), X_i$ for an individual i [Linero and Antonelli, 2022].

We can start by defining the simplest causal estimand. The *individual treatment effects*, ITE, is just the difference between the potential outcomes of an individual, which is: $\tau_i = Y_i(1) - Y_i(0)$. The sample average of the different individual treatment effects is called the *sample average treatment effect*, SATE, and is denoted as: $\tau_s = \frac{1}{N}\sum_{i=1}^{N} \tau_i$. If we extend the average treatment effect to the whole population of interest, we obtain the *population average treatment effect*, PATE, denoted as: $\tau_p = E[Y_i(1) - Y_i(0)] = E[\tau_i]$. We can also have variations of the SATE and PATE that focus on certain groups of individuals grouped by their exposure A_i, such as the *average treatment effect on treated*, ATT, defined as: $E[Y_i(1) - Y_i(0)|A_i = 1]$, with analogous definitions for the control population [Linero and Antonelli, 2022]. The last causal estimand that we mention is the *conditional average treatment effects*, CATE, defined as: $\tau_x = E[Y_i(1) - Y_i(0)|X_i = x]$. CATE can also be extended to the ATE conditioned on the individual belonging to a certain subgroup C based on the covariates, $\tau_C = E[Y_i(1) - Y_i(0)|X_i \in C]$ [Linero and Antonelli, 2022].

The SATE is typically of interest in randomized experiments with the specific sample as the target population, while the PATE is used in observational studies where the target population is the population from which the sample was drawn. The CATE, in general, can help us study how the treatment effect differs between certain subgroups, characterizing treatment effect heterogeneity [Fan et al., 2023].

As we do not observe both $Y_i(1)$ and $Y_i(0)$, these estimands are obtainable from our observed data (i.e., identified) only under certain assumptions that we call the *identifying assumptions*. The estimands that we mentioned above are all identified under the same set of assumptions [Linero and Antonelli, 2022], which we study in the following subsection.

4.2 Identifying Assumptions

The identifying assumptions for the above causal estimands are:

1. *Stable unit treatment value*
2. *Positivity*
3. *Strong ignorability*

The stable unit treatment value assumption (SUTVA) comprises two assumptions: *no interference* and *consistency*. The no interference assumption states that for a treatment assignment $A = (A_1 = a_1, A_2 = a_2, ..., A_N = a_N)$, the potential outcome of an individual i, Y_i only depends on a_i. That is, $Y_i = Y_i(a_i)$. This assumption makes sure that the outcome of individual i is not affected by the treatment assignment of other individuals, whether or not they receive the treatment. The consistency assumption then states that there are not various versions of the treatment that is administered.

Before passing to the positivity assumption, we define what the *propensity score* is. The propensity score is the probability of receiving the treatment given the observed covariates for each individual, that is, $e(x) = P(A_i = 1|X = x)$. In randomized controlled trials with binary treatment, the expectation of the propensity score is 0.5. However, in non-randomized circumstances, such as in the case of observational studies where treatments are assigned based on external factors not controlled by the researcher, the propensity score might be, and usually is, heterogeneous between different groups. For example, in an observational medical study, a young patient with milder symptoms might have a lower probability of receiving aggressive treatment compared to an older patient with severe symptoms.

The positivity assumption puts the following bounds on the propensity score: $e(x) = P(A_i = 1|X = x)$: $0 < e(x) < 1, \forall x \in \mathcal{X}$, $\mathcal{X}$ being the covariate space. This makes sure that every individual has a chance to get treated, and no individual is certain to be treated.

The last, and perhaps the most important, assumption is strong ignorability. This assumption states that, conditioning on $X_i = x$, the potential outcomes $Y_i(0), Y_i(1)$ are (conditionally) independent of the treatment assignment A_i, in notation: $\{Y_i(0), Y_i(1) \perp A_i | X_i = x\}, \forall x \in \mathcal{X}$. Intuitively, this assumption means that given the covariates X_i, we cannot obtain more information about the characteristics of individual i by observing the treatment assignment A_i. That is, we have measured all common causes of the treatment and outcome in the covariates X_i and we can treat the treatment assignment as if it were random among a group of individuals with the same covariates. The importance of this assumption is that it allows us to use information from the treated or control sample to make inferences about the entire population of interest in non-randomized studies [Linero and Antonelli, 2022]. This is because we can now be confident that, having accounted for all possible confounders in X_i, the difference in observed outcomes of treated and untreated individuals comes from the treatment itself and not from some other confounder, i.e., we remove the confounding bias. As an example, we can go back to the after-school program example that we mentioned in the introduction. If we condition on students' motivation by including it among the covariates and measure the difference in outcomes between students of the same motivation level, we can be fairly confident that the results come from the program (treatment) and not from the inherent motivation of the students. Of course, this is a very simplified example, and in real-world scenarios there might be confounders other than motivation. In real-world observational studies, where violations of the ignorability assumption are common, sensitivity analysis is essential. This analysis evaluates how robust our inferences are against violations of this assumption, helping us assess the reliability of our results.

Thanks to these assumptions, we can identify the above estimands in terms of quantities calculated using the observed data. We identify the conditional density of the potential outcomes as:

$$f_\theta\{Y_i(a) = y | X_i = x\} = f_\theta\{Y_i(a) = y | A_i = a, X_i = x\} \tag{18}$$

for $\forall x \in \mathcal{X}$ and with θ representing the parameters of this conditional distribution. The above equation can be interpreted as follows: As we do not obtain additional information from observing treatment assignment A_i having conditioned on the covariates X_i, conditioning on A_i does not affect the distribution of the outcomes conditioned on X_i. Starting from (18), we can then identify CATE as [Linero and Antonelli, 2022]:

$$\tau_x = E_\theta\{Y_i(1) - Y_i(0)|X = x\} = E_\theta\{Y_i(1)|X = x, A_i = 1\} - E_\theta\{Y_i(0)|X = x, A_i = 0\}$$

By marginalizing CATE over the covariate space $\mathcal{X}$, we can then obtain the PATE:

$$\tau_p = E_\theta\{Y_i(1)-Y_i(0)\} = \int_{\mathcal{X}} (E_\theta\{Y_i(1)|X = x, A_i = 1\}-E_\theta\{Y_i(0)|X = x, A_i = 0\}) f_\theta(x) dx$$

We have studied the assumptions that allow us to compute estimands such as CATE and PATE from the observed data. It is important to note that there exist other estimands that may be of interest in more specific circumstances, such as the estimands in mediation analysis that we will talk about in the next section, that might have different identifying assumptions.

4.3 Practical Challenges in Bayesian Causal Inference

In this subsection, we will talk about some common challenges in applications of Bayesian causal inference models, stemming from the construction of the prior distribution or from the relationship between the outcome and the propensity score model.

4.3.1 Regularization Induced Confounding

As we have mentioned in the previous sections, an important feature of Bayesian methods that led to their popularity in causal inference is their ability to integrate prior information into the model through informative priors. These informative priors help us regularize the nuisance parameters, the parameters on the covariates that we must account for to avoid confounding, aiming to prevent overfitting and decreasing the complexity of the model. An example of this is the function $g(X_i)$ (1) in the context of Gaussian processes. However, in some cases, this regularization might become counterproductive by regularizing the parameters on covariates so much that the causal parameters that we are trying to estimate are also indirectly regularized. We can think about this as follows: By regularizing the covariates, we risk not accounting enough for their effect on our model, which leads to biased estimates of causal parameters as we do not account for confounders. This phenomenon is called *regularization induced confounding* (RIC) [Hahn et al., 2018].

As another form of regularization, we can mention the assignment of independent priors on the outcome and propensity score models. This assumption of prior independence is very common in the context of causal inference and is adopted in order to simplify the model and decrease the computational complexity. However, just as regularization in the outcome model may lead to bias through the oversimplification of the influence of the covariates, this assumption of independence may cause us to fail to capture important relationships between the treatment mechanism and the outcomes, causing another source of bias in the estimates.

To demonstrate the complications of RIC, we take a look at the example provided in Linero and Antonelli (2022) that focuses on RIC in nonparametric Bayesian inference. We consider a semiparametric model with a binary treatment assignment model and outcome model:

$$A_i \sim Bernoulli(e_\theta(x))$$
$$Y_i(a) = g(X_i, a) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

In this example, the causal parameter that gets impacted by RIC is the PATE through the unusual regularization of its selection bias. The selection bias gives us an insight into the presence of confounders that affect both the treatment assignment and the outcome and is defined as follows: $\Delta(a) = E_\theta[Y_i(a)|A_i = a] - E_\theta[Y_i(a)]$. This is the difference between the expected outcome of people who actually receive/not receive the treatment and the expected outcome in a hypothetical scenario where everyone receives/not receives the treatment. In other words, $\Delta(a)$ represents the bias in estimating $E_\theta[Y_i(a)]$ with a naive estimator such as $\frac{\sum_i I(A_i=a)Y_i}{\sum_i I(A_i=a)}$ that does not account for confounders. Of course, we expect $\Delta(a) \neq 0$ as we do not account for any confounders. However, as outlined by Linero and Antonelli (2022), in many straightforward applications of nonparametric Bayes, we find $\Delta(a)$ to be strongly shrunk towards 0, interfering with the causal inference process as we cannot robustly estimate the selection bias to adjust for it. To motivate this phenomenon, Linero and Antonelli (2022) use the following factorization of the selection bias:

$$\Delta(a) = \frac{Var_\theta\{e_\theta(X_i)\}^{\frac{1}{2}} Var_\theta\{g(1, X_i)\}^{\frac{1}{2}}}{E_\theta[e_\theta(X_i)]} Cor_\theta\{e_\theta(X_i), g(1, X_i)\}$$

It turns out that when we assign independent priors to $e_\theta(X_i)$ and $g(1, X_i)$, although we would expect some kind of correlation and hence $\Delta(a)$ to be nonzero, these turn out to be nearly uncorrelated in posterior inference, which is counter-intuitive and indicates a problem with the posterior distribution that we obtain. This unexpected independence in the posterior and the incorrect shrinkage of Δ happen when P is of moderate dimension in nonparametric models or when P is large in parametric models [Linero and Antonelli, 2022]. In these cases, we are not able to make correct posterior inferences about some parameters of interest, such as $\Delta(a)$. This is because the prior assumptions are so dominant that the data has very little influence over the posterior, a phenomenon that is also called *prior dogmatism*.

Here, we mention two approaches provided in Linero and Antonelli (2022) that impose a functional relationship between the outcome and propensity score models in order to overcome RIC and prior dogmatism. Zhou et al. (2019) propose to set the relationship $g(1, x) = \mu(1, x) + \beta\{e_\theta(x)\}$ in order to force correlation between the outcome and the propensity score models, with β modeled nonparametrically for more flexibility. On the other hand, Hahn et al. (2020) propose to include the propensity score $e_\theta(x)$ as a predictor in g by setting $Y_i(a) = g\{a, x, e_\theta(x) + \epsilon_i\}$. We can think of this approach as making explicit the specific covariates that influence the propensity score, rendering the model less prone to wrongly attributing to the treatment the effects that are actually due to confounding variables.

4.3.2 Instability of the CATE

Another issue outlined by Linero and Antonelli (2022) is the instability of the CATE, $\tau_x = g(1, x) - g(0, x)$, when we place independent priors on $g(1, x)$ and $g(0, x)$. The motivations behind choosing independent priors for these quantities are similar to assuming prior independence between the outcome and propensity score model, such as decreasing model complexity. The independence between the priors of $g(1, x)$ and $g(0, x)$ might lead us to find heterogeneous effects when there are none, with the heterogeneity of effects referring to variations in the treatment effect across different levels or values of the covariates x.

In Hahn et al. (2020), the following parametrization is proposed to deal with RIC and the instability of the CATE:

$$Y_i(a) = \mu(X_i, \hat{e}_i) + a\tau_x + \epsilon_i(a) \tag{19}$$

with $\hat{e}_i$ being an estimate of $e_i(X_i)$. In the above parametrization, instead of specifying independent priors for $g(1, X_i)$ and $g(0, X_i)$, we specify independent priors for the prognostic effect of the confounders $\mu(x, e)$ and the treatment effect τ_x. In this case, we can interpret $\mu(x, e)$ as representing the expected outcome in the absence of treatment, that is, the part of the outcome that originates solely from the covariates. Because one would expect a larger heterogeneity in the effects of the covariates compared to treatment effects, τ_x might be shrunk towards a constant so that the model is *shrunk towards homogeneity* [Linero and Antonelli, 2022].

We can specify any nonparametric prior for $\mu(x, e)$ and τ_x. In particular, Hahn et al. (2020) propose BART priors for these quantities, calling the final model a *Bayesian Causal Forest*. In a BCF, the prior for τ_x encourages empty trees, trees with only the root, enforcing the assumption that the treatment effects are homogeneous as empty trees will result in constant treatment effects. Linero and Antonelli (2022) mention BCFs as a state-of-the-art method for estimating the SATE, PATE, and CATE, citing the work of Dorie et al. (2018).

4.3.3 Posterior Inference of Causal Estimands

In Bayesian causal inference, one of the main practical challenges is the posterior inference of causal estimands that do not have a closed form. Suppose we have sampled, through MCMC, B posterior samples from the posterior distribution of the model parameters, which we denote as $\theta_1, \theta_2, ..., \theta_B$. We denote our causal estimand of interest as $\tau(\theta)$. If $\tau(\theta)$ is closed-form, we can easily compute it by taking its expectation over the sampled θ, $\theta_{1:B}$. An example of a closed-form estimand is the CATE, which under the identifying assumptions from Section 4.2 is identified as $\tau_x(\theta) = E_\theta\{Y_i(1)|X = x, A_i = 1\} - E_\theta\{Y_i(0)|X = x, A_i = 0\}$. According to Linero and Antonelli (2022), if we use the BCF specification for (19), then the estimates of the CATE are readily available.

An example of a causal estimand that is not a closed-form function of θ is the PATE which we cannot simplify beyond the below expression:

$$\tau_p = \int_{\mathcal{X}} (E_\theta\{Y_i(1)|X = x, A_i = 1\} - E_\theta\{Y_i(0)|X = x, A_i = 0\}) f_\theta(x) dx \tag{20}$$

There is no guarantee that this expression can be integrated in closed form. Linero and Antonelli (2022) propose addressing this problem by using the Bayesian Bootstrap [Rubin, 1981] by expressing the PATE as:

$$\tau_p(\theta) = \sum_{j=1}^{J} \omega_j \{\mathbb{E}_\theta(Y_i \mid A_i = 1, X_i = x_j) - \mathbb{E}_\theta(Y_i \mid A_i = 0, X_i = x_j)\}$$

with $\omega \sim Dirichlet(m_1, ..., m_j)$, $(x_1, ..., x_J)$ being the unique values of the covariates X_i and m_j is the number of times we observe $X_i = x_j$. This method directly provides us with posterior samples of $\tau_p(\theta_{1:B})$ to construct point estimates and intervals.

In settings where the Bayesian Bootstrap is not feasible, such as in cases where some components of X_i are missing, Linero and Antonelli (2022) suggest approximating (20) with another round of Monte Carlo. Given NK samples $X^*_{ik} \sim f_\theta(x)$ we can approximate the PATE as:

$$\tau_p(\theta) \approx \frac{1}{NK}\Sigma_{i,k}\{E_\theta[Y_i|A_i = 1, X_i = X^*_{ik}] - E_\theta[Y_i|A_i = 0, X_i = X^*_{ik}]\}$$

with K being the number of "pseudo-datasets" used for the approximation of the integral. K can be set as large as desired to make the Monte Carlo error negligible. It is noted by Linero and Antonelli (2022) that even $K = 1$ often provides an accurate approximation and that the inferences are conservative even in cases where the Monte Carlo error is large.

5 Mediation Analysis

In causal inference studies, it is not uncommon to come across treatments that affect the outcome through more than one *causal pathway* [Kim et al., 2017]. Going back to the after-school program example, we can argue that participation in the program improves students' performance not only by providing them with extra knowledge during the after-school lessons themselves but also by increasing the time that students spend on homework, which could be a distinct factor influencing students' performance. In this case, we can distinguish between the *direct effects* of the program that directly influence the outcome and the *indirect effects* where the outcome is influenced by changes in a *mediator* that is itself influenced by the treatment. In our example, the mediator is the time spent on homework. Formally, we can denote the treatment A_i, the outcome Y_i, and the covariates X_i as before, and denote the mediator as $M_i(A_i)$.

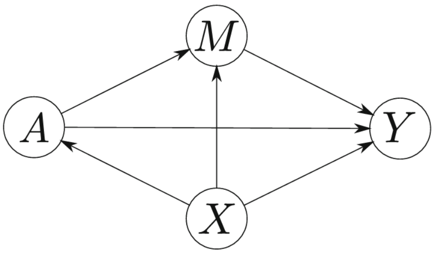


Figure 1: Directed acyclic graph from Linero and Antonelli (2022) representing the causal structure of the simple mediation example described above.

Mediation analysis is concerned with exploring the indirect effects of a treatment through a mediator in the context of causal inference. In the following subsections, we will study the identifying assumptions for causal estimands in mediation and look into some Bayesian models for estimating direct and indirect effects.

5.1 Identification and Assumptions

The most important causal estimands in mediation analysis are the *natural indirect effects* (NIE) and the *natural direct effects* (NDE). Given covariates X, the binary treatment A_i, and the potential mediator $M_i(A_i = a)$, the NIE and NDE conditioned on the covariate values $X = x$ are defined in Kim et al. (2017) as follows:

$$NIE(x) = E[Y_i\{A_i = 1, M_i(1)\} - Y_i\{A_i = 1, M_i(0)\}|X = x]$$
$$NDE(x) = E[Y_i\{A_i = 1, M_i(0)\} - Y_i\{A_i = 0, M_i(0)\}|X = x]$$

Another way of defining NIE and NDE can be seen in Linero and Antonelli (2022), where the conditioning is on the treatment level a instead:

$$NIE(a) = \delta(a) = E[Y_i\{A_i = a, M_i(1)\} - Y_i\{A_i = a, M_i(0)\}]$$
$$NDE(a) = \zeta(a) = E[Y_i\{A_i = 1, M_i(a)\} - Y_i\{A_i = 0, M_i(a)\}]$$

These definitions are closely connected, and one may prefer one to the other depending on the use case. We can easily notice that by marginalizing $NIE(x)$ and $NDE(x)$ we obtain

$NIE = \delta(1)$ and $NDE = \zeta(0)$. Additionally, we define the total effects, TE, as the sum of the natural indirect and direct effects. That is, $TE = NIE + NDE$.

There are several sets of assumptions that we can consider to identify and estimate the causal effects of interest. Below, we will consider the *sequential ignorability* assumption as presented in Linero and Antonelli (2022). Weaker set of assumptions, such as the *Mediator Induction Equivalence* assumptions, can also be considered in specific cases, as mentioned in Kim et al. (2017). The sequential ignorability assumption consists of the following statements:

$$[\{Y_i(a', m), M_i(a)\} \perp A_i | X_i = x] \tag{21}$$

$$[Y_i(a', m) \perp M_i(a) | A_i = a, X_i = x] \tag{22}$$

$$P(A_i = a | X_i = x) > 0 \quad \& \quad P(M_i(a) = m | A_i = a, X_i = x) > 0 \tag{23}$$

for all (a, a', m, x). Condition (21) represents the assumption that the potential outcomes of the response variable Y_i and the mediator M_i are conditionally independent from the treatment assignment given the observed covariates x. Condition (22) represents the assumption that the potential outcomes of the response and the potential outcomes of the mediator are conditionally independent given the assigned treatment a and the observed covariates x. Finally, (23) represents the assumption that any combination of the treatment and the mediator is observable for any value of x.

It is pointed out that (21) and (22) can be thought of as variations of the strong ignorability assumption, and although (21) may be known to hold in randomized studies, these are not directly verifiable from the observed data.

Given the sequential ignorability assumption, Linero and Antonelli (2022) propose the following method for posterior inference of $\zeta(a)$ and $\delta(a)$. We identify the marginal means

$\mu(a, a') = E_\theta[Y_i\{a, M_i(a')\}]$ as:

$$\mu(a, a') = \int E_\theta(Y_i | A_i = a, M_i = m, X_i = x) f_\theta(M_i = m | A_i = a', X_i = x) f_\theta(x)\, dm\, dx \tag{24}$$

The marginal means can be computed after computing the posterior distribution of θ. We then follow with the posterior inference by setting $\zeta(a) = \mu(1, a) - \mu(0, a)$ and $\delta(a) = \mu(a, 1) - \mu(a, 0)$.

5.2 Bayesian Models for Mediation Analysis

5.2.1 Dirichlet Process Mixtures of Multivariate Normals

For the joint distribution of the observations of Y_i, M_i and X, Kim et al. (2017) propose a Dirichlet process mixtures of multivariate normals. Given q-dimensional data, we specify, for each treatment $A_i = a$, the following nonparametric model:

$$\begin{aligned}
(Y_i^a, M_i^a, X_i^a) \mid (\boldsymbol{\mu}_{ai}, \Sigma_{ai}) &\overset{\text{ind}}{\sim} N_q(\boldsymbol{\mu}_{ai}, \Sigma_{ai}), \quad i = 1, \ldots, n_a \\
(\boldsymbol{\mu}_{ai}, \Sigma_{ai}) \mid G_a &\overset{\text{iid}}{\sim} G_a, \quad i = 1, \ldots, n_a \\
G_a &\sim DP(\alpha_a, G_{0a})
\end{aligned}$$

Here, (Y_i^a, M_i^a, X_i^a) denotes the observed value of the outcome, mediator, and covariates for individual i under treatment a, and n_a denotes the number of individuals under treatment a. For the mass parameter α_a of the Dirichlet process, we specify Gamma prior $G(1, 1)$. On the other hand, for the base distribution G_{0a}, a *normal-inverse-Wishart* (NIW) distribution is used. The normal-inverse-Wishart distribution is defined as follows:

$$\begin{aligned}
\boldsymbol{\mu} &\sim N_n(\boldsymbol{\mu}_0, \frac{1}{\lambda}\Sigma) \\
\Sigma &\sim \mathcal{W}^{-1}(\Psi, \nu) \\
(\boldsymbol{\mu}, \Sigma) &\sim NIW(\boldsymbol{\mu}_0, \lambda, \Psi, \nu)
\end{aligned}$$

where N_n represents a multivariate normal distribution and $\mathcal{W}^{-1}$ an inverse-Wishart distribution.

The NIW distribution is a common choice for modeling the joint prior of the mean vector and the covariance matrix of a multivariate normal. This is because, given a multivariate normal likelihood, the NIW distribution is a conjugate prior for the parameters $(\boldsymbol{\mu}, \Sigma)$. Therefore, in cases like the one described above, where the data on which we update the prior is modeled according to a multivariate normal distribution, using NIW priors makes the computation of the posterior distribution more straightforward.

Kim et al. (2017) suggest following the specification proposed by Taddy (2008) and Jara et al. (2011) for the various hyperpriors such as the Gamma prior on the mass parameter of the DP and the parameters of the NIW distribution.

5.2.2 Structural Equation Model

We now have a look at a more applied example presented in Linero (2022). This example considers the Job Search Intervention Study (JOBS II), a randomized trial that explored the effects of a job training program on the participants' mental well-being and ability to find employment. Linero (2022) performs a mediation analysis aiming to decompose the effects of the program into its direct effects on the participants' mental health and its indirect effects arising due to the effect of the program on self-efficacy in job finding.

With $M_i(a)$ (mediator) representing a measure of self-efficacy in job search given treatment $A_i = a$ and $Y_i(a, m)$ (outcome) representing a measure of depression given treatment $A_i = a$ and mediator value m, we define the below structural equation model [Baron and Kenny, 1986]:

$$M_i(a) \mid (\alpha_1, \beta_1, \xi_1, \sigma_m^2) \overset{\text{ind}}{\sim} N(\alpha_1 + \beta_1 a + \xi_1^T X_i, \sigma_m^2) \tag{25}$$

$$Y_i(a, m) \mid (\alpha_2, \beta_2, \gamma, \xi_2, \sigma_y^2) \overset{\text{ind}}{\sim} N(\alpha_2 + \beta_2 a + \gamma m + \xi_2^T X_i, \sigma_y^2) \tag{26}$$

Here, the parameters α_1, β_1, ξ_1 and σ_m^2 represent, respectively, the parameters for the mean and for the variance of the mediator model $M_i(a)$. Likewise, the parameters $\alpha_2, \beta_2, \gamma, \xi_2$ and σ_y^2 represent, respectively, the parameters for the mean and for the variance of the outcome model $Y_i(a, m)$. Furthermore, the observed covariates X_i are said to be distributed independently according to F_x.

The direct and indirect effects can be rewritten as $\zeta(a) = \beta_2$ and $\delta(a) = \gamma\beta_1$. Therefore, $\zeta(a)$ and $\delta(a)$ can be estimated using point estimates of these coefficients. One way to estimate β_1, β_2, γ is by regressing Y_i on (M_i, A_i, X_i) and M_i on (A_i, X_i). On the other hand, we can also proceed with the estimation by sampling from the posterior distribution of $(\beta_1, \beta_2, \gamma)$ under a Jeffreys prior [Linero, 2022].

An alternative approach to estimating the direct and indirect effects is by using Monte Carlo integration. Linero (2022) describes this approach as follows: We first sample from the posterior distribution of $\theta = (\alpha_1, \beta_1, \xi_1, \alpha_2, \beta_2, \xi_2, \gamma, \sigma_m, \sigma_y, F_x)$ with a Bayesian bootstrap prior on F_x. Then, N_W values of $X_i^* \sim F_x$ are simulated, and $[M_i^*(a), Y_i^*\{a, M_i^*(a')\} : a, a' \in \{0, 1\}, i = 1, ..., N_W]$ is sampled according to (25) and (26). In the end, $\delta(a)$ and $\zeta(a)$ are approximated as below:

$$\hat{\delta}(a) = \frac{\sum_{i=1}^{N_W} Y_i^*\{a, M_i^*(1)\} - Y_i^*\{a, M_i^*(0)\}}{N_W}$$
$$\hat{\zeta}(a) = \frac{\sum_{i=1}^{N_W} Y_i^*\{1, M_i^*(a)\} - Y_i^*\{0, M_i^*(a)\}}{N_W}$$

Furthermore, Linero (2022) states that the variance of $\hat{\delta}(a)$ is approximated by:

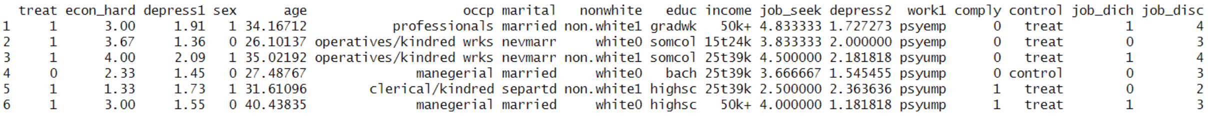
$$\hat{V} = \frac{\widehat{\text{Var}}\{Y_i^*\{a, M_i^*(1)\} - Y_i^*\{a, M_i^*(0)\}\}}{N_W}$$

where $\widehat{\text{Var}}$ represents taking the sample variance. The variance of $\hat{\zeta}(a)$ is also approximated analogously.

Some shortcomings of Monte Carlo integration in this example are highlighted in Linero (2022). While it performs well for the estimates of ζ, the approximated posterior distribution of δ suffers from excessive variance compared to the true posterior. As a solution, Linero (2022) proposes a computational method called *Accelerated g-computation* (AGC), which not only exhibits much better performance but also approximates a posterior that aligns closely with the true posterior, in addition to being computationally faster.

6 Application

We illustrate an example of Bayesian nonparametric mediation analysis by applying the Dirichlet process mixtures of multivariate normals model described in Section 5.2.1 to the JOBS II dataset from Section 5.2.2. As mentioned previously, the JOBS II dataset is based on the results of a randomized field experiment that investigated the efficacy of a job training intervention on unemployed workers. The dataset is readily available as part of the `mediation` package in R, requires minimal preprocessing, and contains no missing values, which makes it a great dataset to be used for illustration purposes. Below, the first few entries of `data(jobs)` are shown:



```
  treat econ_hard depress1 sex      age                    occp marital   nonwhite   educ income job_seek depress2  work1 comply control job_dich job_disc
1     1      3.00     1.91   1 34.16712           professionals married non.white1 gradwk   50k+ 4.833333 1.727273 psyemp      0   treat        1        4
2     1      3.67     1.36   0 26.10137 operatives/kindred wrks nevmarr     white0 somcol 15t24k 3.833333 2.000000 psyemp      0   treat        0        3
3     1      4.00     2.09   1 35.02192 operatives/kindred wrks nevmarr non.white1 somcol 25t39k 4.500000 2.181818 psyump      0   treat        1        4
4     0      2.33     1.45   0 27.48767              manegerial married     white0   bach 25t39k 3.666667 1.545455 psyump      0 control        0        3
5     1      1.33     1.73   1 31.61096        clerical/kindred separtd non.white1 highsc 25t39k 2.500000 2.363636 psyump      1   treat        0        2
6     1      3.00     1.55   0 40.43835              manegerial married     white0 highsc   50k+ 4.000000 1.181818 psyump      1   treat        1        3
```

Figure 2: The Jobs II dataset

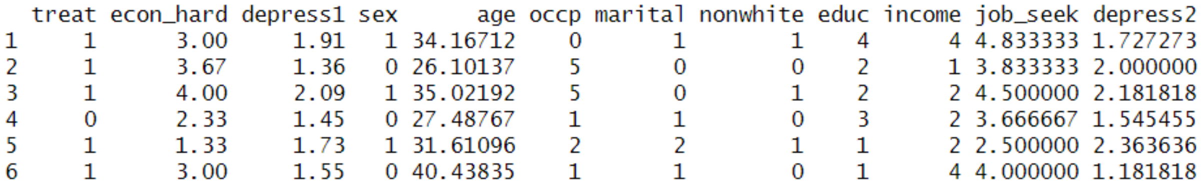
Here, `treat` shows whether the individual has participated in the program; `job_seek`, the mediator variable, is a continuous scale measuring the level of job-search self-efficacy with values from 1 to 5; and `depress2`, the outcome, is a measure of depressive symptoms post-treatment. Moreover, `depress1` is a measure of depressive symptoms pre-treatment, `econ_hard` is a measure of economic hardship pre-treatment with values from 1 to 5, and `occp` is a factor with seven categories for various occupations.

6.1 Data Preprocessing

To prepare the data for modeling, we drop some columns and transform the non-numerical columns into numerical ones. We drop the `control` variable, as it conveys the same information as `treat` and the `job_dich` and `job_disc` variables, as these just record the `job_seek` variable into categories. Moreover, the variables `work1` and `comply` are also dropped. This is because `work1`, which is an indicator for employment after the training, is a post-treatment variable, that is, a variable that we know is influenced by the treatment, and including it may introduce what is called *post-treatment bias*, which occurs when variables that are affected by the treatment in a study are included among the covariates, distorting the true effect of the treatment. On the other hand, `comply` indicates whether the individual selected for participation in the program, that is, included in the treatment group $A_i = 1$, actually participated in the program. The choice between using `treat` and `comply` as the treatment indicator will be discussed at the end of this section. As the final preprocessing step, we then convert to binary or categorical (starting from 0) the variables `occp`, `marital`, `nonwhite`, `educ`, `work1`, and `income`. The code for the preprocessing and the final dataset `jobs2` are presented below:

```r
jobs2 <- jobs[, !names(jobs) %in% c("control", "job_disc",
 ↪  "job_dich", "work1", "comply")]

jobs2$occp <- as.integer(as.factor(jobs$occp)) - 1
jobs2$marital <- as.integer(as.factor(jobs$marital)) - 1
jobs2$nonwhite <- as.integer(as.factor(jobs$nonwhite)) - 1
jobs2$educ <- as.integer(as.factor(jobs$educ)) - 1
jobs2$income <- as.integer(as.factor(jobs$income)) - 1
```



```
  treat econ_hard depress1 sex      age occp marital nonwhite educ income job_seek depress2
1     1      3.00     1.91   1 34.16712    0       1        1    4      4 4.833333 1.727273
2     1      3.67     1.36   0 26.10137    5       0        0    2      1 3.833333 2.000000
3     1      4.00     2.09   1 35.02192    5       0        1    2      2 4.500000 2.181818
4     0      2.33     1.45   0 27.48767    1       1        0    3      2 3.666667 1.545455
5     1      1.33     1.73   1 31.61096    2       2        1    1      2 2.500000 2.363636
6     1      3.00     1.55   0 40.43835    1       1        0    1      4 4.000000 1.181818
```

Figure 3: The preprocessed Jobs II dataset

6.2 Model Specification

As the next step, we specify the Dirichlet Process Mixture (DPM) model. For this goal, we use the `dirichletprocess` package in R, particularly the `DirichletProcessMvnormal` function. This function specifies a DPM of multivariate normals similarly to how we have described it in Section 5.2.1. The only difference between our specification and that of `DirichletProcessMvnormal` is the use of the Normal-Wishart distribution to specify a prior for the precision matrix (the inverse of the covariance matrix), in contrast to our approach, where we use a normal-inverse-Wishart prior directly for the covariance matrix. This is not a problem, as, in the end, both approaches result in the same model specification.

The abbreviated code below shows how the covariates, outcome, and mediator are grouped to achieve the following structure: (Y_i^a, M_i^a, X_i^a). We first define the covariates $x_{1:9}$ and the treatment, outcome, and mediator variables. The covariates, outcome, and mediator are then split according to their treatment group.

```
x1 <- jobs2$age; x2 <- jobs2$econ_hard; ...

out <- jobs2$depress2; trt <- jobs2$treat
y1 <- out[which(trt == 1)]; m1 <- jobs2$job_seek[which(trt == 1)]
y0 <- out[which(trt == 0)]; m0 <- jobs2$job_seek[which(trt == 0)]

x1_1 <- x1[which(trt == 1)]; x2_1 <- x2[which(trt == 1)];...
x1_0 <- x1[which(trt == 0)]; x2_0 <- x2[which(trt == 0)];...

w1 <- cbind(y1, m1, x1_1, x2_1, x3_1, x4_1, x5_1, x6_1, x7_1,
 ↪  x8_1, x9_1)
w0 <- cbind(y0, m0, x1_0, x2_0, x3_0, x4_0, x5_0, x6_0, x7_0,
 ↪  x8_0, x9_0)
```

The DPM model for the distribution of (Y_i^1, M_i^1, X_i^1) is then defined as below:

```
RANGE <- function(x){result <- (range(x)[2]-range(x)[1])^2/16}
```

```
wbar1 <- c(median(y1), median(m1), median(x1_1),... ,
 ↪  median(x6_1), 0.5, 0.5, 0.5)

wcov1 <- diag(c(RANGE(y1), RANGE(m1), RANGE(x1_1),...,
 ↪  RANGE(x6_1), 1, 1, 1), 11, 11)

g0priors1 <- list(mu0 = wbar1, kappa0 = 0.01, Lambda = wcov1, nu
 ↪  = 11)

dp1 <- DirichletProcessMvnormal(y = w1, g0Priors = g0priors1,
 ↪  alphaPriors = c(1, 1), numInitialClusters = 1)
```

The DPM model for the distribution of (Y_i^0, M_i^0, X_i^0) is specified in the same way with analogous inputs for `wbar0` and `wcov0`.

The `DirichletProcessMvnormal` function creates a Dirichlet process object and admits the following parameters: `y`: data, `g0priors`: the prior parameters for the base distribution of the DP, `alphaPriors`: parameters for the Gamma prior of the mass parameter of the DP, `numInitialClusters`: the number of clusters we initialize with.

The elements of `g0Prior` are the terms on which G_0 is conditioned in the following expression: $G_0(\mu, \Lambda | \mu_0, \kappa_0, \nu_0, T_0) = N(\mu | \mu_0, (\kappa_0 \Lambda)^{-1}) \mathcal{W}_{\nu_0}(\Lambda | T_0)$. This is the normal-Wishart prior for the base distribution G_0, and the parameters to specify are: the mean μ_0, the degrees of freedom of the Wishart distribution ν_0, the positive-definite scale matrix T_0 (denoted as `Lambda` in the code), which is the prior for the precision matrix Λ, and κ_0, which controls the concentration of μ around μ_0. For these values, we follow the approach of Kim et al. (2017): for `mu0` we use the median for continuous and categorical variables and set 0.5 for binary variables; for `Lambda` we use the `RANGE` function for continuous and categorical variables, computing a measure of dispersion, and set 1 for binary variables; for κ_0 we specify a low value such as 0.01 in order to limit the influence of the prior μ_0 and to increase the influence of the data on the posterior value of μ; and for ν_0 we input the degrees of freedom of our data, which is 11 (outcome + mediator + 9 covariates) in our case.

Having specified the model, we then run the following code to fit it:

```
niter <- 44000
burnparam <- 40000

fit1 <- Fit(dp1, its = niter, updatePrior = FALSE, progressBar =
 ↪  TRUE)

Burn1 <- Burn(fit1, burnparam)

obj1 <- create_matrix(Burn1$clusterParametersChain)
```

Here, given a Dirichlet process object dp1 and the number of iterations, the Fit function from the dirichletprocess package uses Algorithm 4 from Neal (2000) to carry out the sampling procedure for a Dirichlet process. Remaining in line with the approach of Kim et al. (2017), we do not update the parameters of the base distribution in the fitting process by setting updatePrior to FALSE. The first 40,000 iterations are discarded. This is because the MCMC algorithm used for sampling needs to converge to its stationary distribution to reliably sample from it. We assume that this convergence, or burn-in, phase is concluded after 40,000 iterations, allowing us to rely on the following 4,000 samples for analysis. At this point, Burn1$clusterParametersChain contains the mean vector and covariance matrix of the outcome, mediator, and covariates sampled at each iteration. In order to proceed with mediation analysis, we extract and store these values in obj1, which is a matrix of dimensions $niter \times (q + \frac{q*(q+1)}{2})$, where each row i is formatted as the mean vector sampled at iteration i followed by the upper triangular entries of the covariance matrix sampled at iteration i.

6.3 Bayesian Nonparametric Mediation Analysis

With the samples at hand, we now proceed to the mediation analysis using the function bnpmediation from the library BNPMediation created by Chanmin Kim and Michael Daniels, two of the authors of the paper Kim et al. (2017). Given means and covariances from our samples, this function calculates the posterior means and 95% credible intervals of the marginal natural indirect effects, natural direct effects, and total effects:

$$NIE = E[Y_i\{A_i = 1, M_i(1)\} - Y_i\{A_i = 1, M_i(0)\}] \tag{27}$$

$$NDE = E[Y_i\{A_i = 1, M_i(0)\} - Y_i\{A_i = 0, M_i(0)\}] \tag{28}$$

$$TE = NIE + NDE = E[Y_i\{A_i = 1, M_i(1)\} - Y_i\{A_i = 0, M_i(0)\}] \tag{29}$$

The main challenge in implementing this function for our analysis has been the fact that it relies on the output of the `DPdensity` function from the `DPpackage` package, which has been orphaned and removed from the CRAN repository. This made it so that the only way to use it is by downloading it and compiling it locally. However, the part of the library written in Fortran is complicated to compile, leading to compilation errors that might be too complex to troubleshoot.

As it is not possible to observe the exact contents of the output produced by `DPdensity`, one can analyze the source code of the original `bnpmediation` function to understand the elements extracted from the output of `DPdensity` to compute the causal effects. It was observed that the `bnpmediation` is programmed to extract only the mean vector and covariance matrix sampled in each iteration from the output of `DPdensity`, ignoring the other entries in the output. Because of this, it was possible to modify the `bnpmediation` function to be able to work directly with a matrix that just contains the mean vector and the upper triangular entries of the covariance matrix for each iteration, such as the `obj1` matrix described above. As the covariance matrix is symmetric, it can be reconstructed inside the function from these upper triangular entries. The code for the modified `bnpmediation` function is as follows:

```
bnpmediation <- function(obj1, obj0, q, NN = 100, n1, n0,
 ↪  extra.thin = 5) {
  library(mnormt)
  Len.MCMC <- 1:dim(obj0)[1]
```

```
if (extra.thin != 0) {Len.MCMC <- Len.MCMC[seq(1,
 ↪  length(Len.MCMC), extra.thin)]}
Y1 <- obj1[, 1]
Y0 <- obj0[, 1]
Y11 <- Y1[Len.MCMC]
Y00 <- Y0[Len.MCMC]
mat.given.ij <- function(x, y) ifelse(x <= y, (q - 1) * (x - 1)
 ↪  + y - x * (x - 1) / 2, (q - 1) * (y - 1) + x - y * (y - 1)
 ↪  / 2)
mat <- function(q) outer(1:q, 1:q, mat.given.ij)
pb <- txtProgressBar(min = 0, max = length(Len.MCMC), style =
 ↪  3)
Y10 <- NULL
index <- 0
for (j in Len.MCMC) {
  index <- index + 1; mu2 <- obj0[j, 2:q]
  sigma22 <- matrix(obj0[j, seq(2 * q + 1, q + (q * (q + 1) /
   ↪  2))][mat(q - 1)], q - 1, q - 1, byrow = TRUE)
  joint0 <- do.call("rbind", replicate(NN,
   ↪  data.frame(sapply(1:n0, function(x) {
    sigma22_x <- sigma22
    rmnorm(1, mu2, sigma22_x)
  }))))
  b01 <- NULL
  Weight.num0 <- matrix(nrow = 1, ncol = n0 * NN)
  B0 <- matrix(nrow = 1, ncol = n0 * NN)
  mu1 <- obj1[j, 1]
  mu2 <- obj1[j, 2:q]
  sigma1 <- obj1[j, q + 1]
  sigma12 <- obj1[j, (q + 2):(2 * q)]
  sigma22 <- matrix(obj1[j, (2 * q + 1):(q + (q * (q + 1) /
   ↪  2))][mat(q - 1)], q - 1, q - 1, byrow = TRUE)
  Weight.num0[1, 1:(n0 * NN)] <- dmnorm(joint0, mu2, sigma22)
  b01[1] <- mu1 - sigma12 %*% solve(sigma22) %*% t(t(mu2))
  B0[1, 1:(n0 * NN)] <- sigma12 %*% solve(sigma22) %*%
   ↪  t(joint0)
  Weight <- apply(Weight.num0, 2, function(x) x / sum(x))
  test <- Weight * (b01 + B0)
  Y10[index] <- mean(apply(test, 2, sum))
  setTxtProgressBar(pb, index)
```

```
    }
    z <- list(Y11=Y11, Y00=Y00, Y10=Y10,
              ENIE=mean(Y11-Y10), ENDE=mean(Y10-Y00),
              ETE=mean(Y11-Y00),
              TE.c.i=c(sort(Y11-Y00)[length(Len.MCMC)*0.025],
                       sort(Y11-Y00)[length(Len.MCMC)*0.975]),
              IE.c.i=c(sort(Y11-Y10)[length(Len.MCMC)*0.025],
                       sort(Y11-Y10)[length(Len.MCMC)*0.975]),
              DE.c.i=c(sort(Y10-Y00)[length(Len.MCMC)*0.025],
                       sort(Y10-Y00)[length(Len.MCMC)*0.975]))
    z$call <- match.call()
    class(z) <- "bnpmediation"
    return(z)
  }
```

This function computes `Y11` and `Y00`, the estimated expectations of $Y_i\{A_i = 1, M_i(1)\}$ and $Y_i\{A_i = 0, M_i(0)\}$, respectively, by extracting, for each MCMC iteration, the mean parameter for the outcome from the sampled mean vectors of the distributions of (Y_i^1, M_i^1, X_i^1) and (Y_i^0, M_i^0, X_i^0). It then uses the remaining entries of the sampled mean vectors and covariance matrices in order to sample from the posterior distribution of $Y_i\{A_i = 1, M_i(0)\}$ and obtain estimates of $E[Y_i\{A_i = 1, M_i(0)\}]$ for each MCMC iteration, storing them as `Y10`. These samples are then used to calculate the estimated natural indirect effects, `ENIE`, estimated natural direct effects, `ENDE`, and estimated total effects, `ETE`, along with their respective 95% credible intervals. The natural indirect effects are calculated by subtracting `Y10` from `Y11`, as defined in (27); the natural direct effects are calculated by subtracting `Y00` from `Y10`, as defined in (28); and the total effects are calculated by subtracting `Y00` from `Y11`, as defined in (29). The point estimates for these quantities are then calculated by taking the mean over all MCMC samples, and the credible intervals are calculated by ordering each sample of these quantities and calculating the 2.5th and 97.5th percentiles.

The parameters `obj1` and `obj0` represent the input, which has to be in the format described above; `q` is the number of variables, including the outcome and mediator; `n1` and `n0` are, respectively, the number of observations under $A_i = 1$ and $A_i = 0$. The parameter `extra.thin` denotes the interval for the thinning of the samples, i.e., we "thin" the sample by extracting and using only each "extra.thin"th sample. Finally, `NN` denotes the number of samples used to estimate `Y10` at each MCMC iteration. To obtain the results of our analysis, we run the below code:

```
resultsbnpmed <- bnpmediation(obj1 = obj1, obj0 = obj0, q = 11,
 ↪   NN = 1000, n1 = Burn1$n, n0 = Burn0$n, extra.thin = 4)
```

The point estimates of `Y11`, `Y00` and `Y00` from the first six iterations are shown below as an example:

```
> head(resultsbnpmed$Y11)
[1] 1.720205 1.720308 1.720413 1.720467 1.720302 1.720337
> head(resultsbnpmed$Y00)
[1] 1.783626 1.783534 1.783586 1.783179 1.783848 1.783641
> head(resultsbnpmed$Y10)
[1] 1.740345 1.736331 1.741264 1.752202 1.740417 1.743881
```

The point estimates of `NIE`, `NDE`, and `TE` are as follows:

```
> resultsbnpmed$ENIE)
[1] -0.0250476
> resultsbnpmed$ENDE
[1] -0.03829482
> resultsbnpmed$ETE
[1] -0.06334242
```

Finally, the 95% credible intervals of these effects are as follows:

```
> resultsbnpmed$IE.c.i
[1] -0.03534164 -0.01587757
> resultsbnpmed$DE.c.i
[1] -0.04771600 -0.02794744
> resultsbnpmed$TE.c.i
[1] -0.06408899 -0.06266709
```

The credible intervals can be interpreted in this way: given the prior information and the observed data, there is a 95% probability that the true value of the effects lies within the interval.

- There is a 95% probability that the true natural indirect effects are in the interval $[-0.03534164, -0.01587757]$.
- There is a 95% probability that the true natural direct effects are in the interval $[-0.04771600, -0.02794744]$.
- There is a 95% probability that the true total effects are in the interval $[-0.06408899, -0.06266709]$.

These results show that it is highly probable that both the direct and indirect effects of the treatment are not negligible, as 0 is not contained in any of the above intervals. The job training intervention seems to have a positive influence on the mental health of the participants, that is, it is associated with a decrease in post-treatment depression `depress2`. This decrease can be interpreted as the total effects of the treatment, and these total effects can be decomposed into its direct effects and its indirect effects through the `job_seek` variable.

The training intervention mainly consisted of the teaching of job-search skills and informing participants about coping strategies for dealing with setbacks in the job-search process. With this information, we can hypothesize that the direct effect on mental health may be a result of increased awareness about the coping strategies for setbacks learned during the intervention and that the indirect effects may be a result of increased self-confidence in job-search efficacy, the mediator, which is increased thanks to the job-search skills learned during the intervention.

6.4 Assumptions and Considerations

It is important to note that by using `bnpmediation`, we assumed that the data satisfied the sequential ignorability assumptions. Although the fact that the data comes from a randomized study might help satisfy (21), as we mentioned in Section 5.1, these assumptions are not directly verifiable from the observed data.

The need to maintain this assumption was the main reason behind our choice of using `treat` as the variable indicating the treatment instead of `comply`. To make inferences about the causal effects of the treatment, it might seem more reasonable to consider only the people who actually received the treatment ($\texttt{comply} = 1$) as treated and not those who were selected to receive it ($\texttt{treat} = 1$), as out of the 600 people selected for participation in the program, only 372 of them actually participated (number of individuals with $\texttt{treat} = 1$, $\texttt{comply} = 1$). However, the problem with using `comply` is the loss of randomization and the possible introduction of selection bias, which would compromise the sequential ignorability assumption by not properly accounting for the relationship between the treatment, mediator, and outcome.

Although this approach can show the treatment's efficacy under ideal conditions, the possible violation of sequential ignorability would introduce a significant amount of bias in our analysis, as the `bnpmediation` function largely depends on this set of assumptions. On the other hand, by using `treat` and conserving randomization, we can be more confident that the effect is correctly identified, although the non-compliance of some individuals may lead to a dilution of the estimates. In other words, we give more importance to the fact that the direction of the effect is correctly identified, even if the magnitude is understated. Our approach of including in the treatment group all the people who were selected for the treatment is called *Intention-to-Treat* analysis. In contrast, the approach where only the people who actually received treatment are included in the treatment group is called *Per-Protocol* analysis.

We conclude by pointing out that the modified `mediation` code assumes that the DP model only creates one cluster from the data. This was done in order to simplify the code for the application, as the `DirichletProcessMvnormal` function consistently created only one cluster when fitted with the `jobs2` data. This might occur in cases where the data has relatively homogeneous characteristics, leading the model to group all data points into a single cluster.

References

[Banerjee et al., 2007] Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2007). Model selection through sparse maximum likelihood estimation. *CoRR*, abs/0707.0704.

[Baron and Kenny, 1986] Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182.

[Casella, 2001] Casella, G. (2001). Empirical bayes gibbs sampling. *Biostatistics*, 2:485–500.

[Chipman et al., 1998] Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.

[Chipman et al., 2010] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298.

[Cressie and Johannesson, 2008] Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

[Datta et al., 2016] Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

[Dorie et al., 2018] Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2018). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition.

[Escobar and West, 1994] Escobar, M. and West, M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90.

[Fan et al., 2023] Fan, L., Peng, D., and Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381:20220153.

[Ferguson, 1983] Ferguson, T. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, 24:287–302.

[Ferguson, 1973] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230.

[Friedman et al., 2000] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337 – 407.

[Hahn et al., 2020] Hahn, P., Murray, J., and Carvalho, C. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056. Publisher Copyright: © 2020 International Society for Bayesian Analysis.

[Hahn et al., 2018] Hahn, P. R., Carvalho, C. M., He, J., and Puelz, D. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

[Jara et al., 2011] Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011). DPpackage: Bayesian non- and semi-parametric modelling in R. *Journal of Statistical Software*, 40:1–30.

[Katzfuss and Guinness, 2021] Katzfuss, M. and Guinness, J. (2021). A general framework for vecchia approximations of gaussian processes. *Statistical Science*, 36(1).

[Kim et al., 2017] Kim, C., Daniels, M. J., Marcus, B. H., and Roy, J. A. (2017). A framework for bayesian nonparametric inference for causal effects of mediation. *Biometrics*, 73(2):401–409.

[Linero, 2022] Linero, A. R. (2022). Simulation-based estimators of analytically intractable causal effects. *Biometrics*, 78(3):1001–1017.

[Linero and Antonelli, 2022] Linero, A. R. and Antonelli, J. (2022). The how and why of bayesian nonparametric causal inference. *WIREs Computational Statistics*, 15.

[Müller et al., 2015] Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer-Verlag.

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

[Oganisian and Roy, 2021] Oganisian, A. and Roy, J. A. (2021). A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2):518–551.

[Park and Casella, 2008] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

[Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

[Rubin, 1974] Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66.

[Rubin, 1981] Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130 – 134.

[Splawa-Neyman et al., 1990] Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

[Taddy, 2008] Taddy, M. A. (2008). *Bayesian Nonparametric Analysis of Conditional Distributions and Inference for Poisson Point Processes*. PhD thesis, University of California, Santa Cruz.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[Zhang et al., 2015] Zhang, B., Sang, H., and Huang, J. Z. (2015). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statistica Sinica*, 25(1):99–114.

[Zhou et al., 2019] Zhou, T., Elliott, M., and Little, R. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114:1–19.