UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXVI

Bayesian Nonparametric Dependent Mixtures for Causal Inference with Applications to Air Pollution Epidemiology

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Antonio Canale

Co-supervisore: Prof. Francesca Dominici

Dottoranda: Dafne Zorzetto

30 September 2023

Abstract

Environmental epidemiology research raises intriguing and fascinating causal inference questions, aiming to understand the comprehensive effects of environmental exposures on human health. The complexity of the ties between air pollution exposure, well-being index, demographic and socio-economic characteristics of the population, and air pollution regulations, solicit clear definitions of the causal effects and flexible models. In reply, we propose novel causal models that leverage the desirable characteristics of the Bayesian nonparametric prior, particularly the dependent Dirichlet process. These Bayesian nonparametric mixture models' well-known flexibility and adaptability are not the only reasons. In truth, our proposed models can easily handle two central challenges: the missing data problem, that arises in the causal inference framework of potential outcome, and the clustering structure, that tethers the applied research question with the proposed methodologies engaged in this thesis. Indeed the Bayesian paradigm allows straightforward missing potential outcome imputation, while the mixture structure of the dependent Dirichlet process naturally induces the clustering of the observations, through the latent variable that defines the allocation to the components of the mixture.

In the details, we address two common challenging contexts of causal inference that frequently emerge in observational studies: capture and characterize the heterogeneity in the causal effects and deal with the post-treatment variables, that are affected by the treatment and simultaneously affect the outcome. Both contexts elicit the concept of clustering since the heterogeneous causal effects demand the clarification of the groups and the post-treatment variables induce the constitution of principal strata. These concepts manifest in environmental epidemiology studies as the groups of populations that are differently affected by air pollution exposure or air quality regulations. Indeed, different levels of vulnerability/resilience characterize the population, highlighting the socio-economic disparities in American society.

Our proposed models—the *confounder-dependent Bayesian mixture model* and the *confounders-aware shared-atoms mixture model*—allow us to exploit rich forms of dependence given the confounders and relationship between the variables with different treatment levels, enabling us to (i) define with a flexible structure the probability distribution of outcome/post-treatment variable, (ii) impute the missing data properly, (iii) estimate individual treatment effects, competitively with benchmark models, (iv) identify the groups/strata structure according to the causal estimands of interest, (v) delineate the characteristics of each group/stratum.

Sommario

La ricerca nell'ambito dell'epidemiologia ambientale solleva interrogativi affascinanti e stimolanti nel contesto dell'inferenza causale, mirando a comprendere nel loro complesso gli effetti dell'ambiente sulla salute pubblica. La complessità dei rapporti tra l'esposizione all'inquinamento atmosferico, la salute pubblica, le disuguaglianze demografiche e socio-economiche della popolazione e le normative governative sull'ambiente, richiede definizioni chiare degli effetti causali e modelli flessibili. In risposta a ciò, proponiamo innovativi modelli causali che sfruttano le caratteristiche delle prior non parametriche bayesiane, in particolare il processo dipendente di Dirichlet. La ben nota flessibilità e adattabilità di questi modelli mistura non parametrici bayesiani non sono le uniche ragioni di scelta. Infatti, i modelli da noi proposti possono gestire facilmente due questioni centrali: il problema dei dati mancanti, intrinseco alla definizione di variabili risposte controfattuali nell'inferenza causale, e la struttura di raggruppamento, che lega le domande di ricerca applicata alle metodologie proposte coinvolte in questa tesi. Infatti, il paradigma bayesiano consente un'agevole e corretta imputazione dei dati mancanti della variabile risposta controfattuale, mentre la struttura a mistura del processo dipendente di Dirichlet induce naturalmente il raggruppamento delle osservazioni, attraverso la variabile latente che definisce l'allocazione ai componenti della miscela.

Nel dettaglio, affrontiamo due sfide di inferenza causale che emergono frequentemente negli studi osservazionali: catturare e caratterizzare l'eterogeneità negli effetti causali e gestire le variabili post-trattamento, che sono influenzate dal trattamento e contemporaneamente influiscono la variabile risposta. Entrambi i contesti richiamano il concetto di raggruppamento, poiché gli effetti causali eterogenei richiedono la definizione dei gruppi e le variabili post-trattamento inducono la costituzione di strati principali. Questi concetti si manifestano negli studi di epidemiologia ambientale come i sottogruppi della popolazione che sono influenzati in modo diverso dall'esposizione

all'inquinamento atmosferico o dalle normative sulla qualità dell'aria. In effetti, diversi livelli di vulnerabilità/resilienza caratterizzano la popolazione, mettendo in luce le disparità socio-economiche nella società americana.

I nostri modelli proposti—il *confounder-dependent mixture model* e il *confounders-aware shared-atoms mixture model*—ci consentono di sfruttare elaborate forme di dipendenza con le variabili confondenti e la relazione tra le variabili con diversi livelli di trattamento, consentendoci di (i) definire la distribuzione di probabilità delle variabili risposta/post-trattamento con una struttura flessibile, (ii) imputare correttamente i dati mancanti, (iii) stimare gli effetti individuali del trattamento, in modo competitivo rispetto ai modelli di riferimento in letteratura, (iv) identificare la struttura di gruppi/-strati in base agli stimatori causali di interesse, (v) delineare le caratteristiche di ciascun gruppo/strato.

First and foremost, have fun

Acknowledgements

Thanks to the three people who have been my supervisors and mentors in these last years: Antonio Canale, Francesca Dominici, and Falco J. Bargagli-Stoffi. They have been a great example of professors/research and amazing people, and have guided me during my Ph.D. with immense patience and dedication, and always with the right answers to my questions. A fundamental thanks to Antonio for believing in my potential and for finding a way to transform it into useful abilities, guiding me toward results I couldn't imagine. I will always be grateful to Francesca for her amazing welcome in Boston and for being a role model of success with her enthusiasm and kindness. And a sincere thanks for all the amazing help that Falco has given, contributing to all the parts of this thesis.

Spending almost my Ph.D. in the United States has given me the inestimable opportunity to grow up—not only in the academic path—and to know wonderful people with whom to share peaceful moments, funny evenings, delicious food, and support. Thanks in particular to the best flatmates ever and friends, Léa and Céline, for making me feel at home, and to Federica for being always kind, lively, and supportive (especially during my thesis writing). I am grateful to Max, Sally, and Falco for all the essential academic and life advice.

Among my Paduans colleagues, a special thanks is to Petro e Federica, with whom share great ideas as *ELIU* and trips around the States.

However, nothing of this could have happened without the more enjoyable years in my education path: the bachelor's and master's, with amazing friends that make everything funnier. An important push was given by Ilaria e Fabia, who encouraged me to apply for the Ph.D. when I was unsure and insecure.

Last but not least, thanks to my family: my parents, who make me believe that I can achieve my dreams, always having fun; Tobia, the only person with whom I have tried (and reasonably given up) to compete; and the irreplaceable and lively Giona. And thanks to my favorite person, Elisa e Ilaria, who makes everything funnier and memorable.

Contents

List of Figures

List of Tables

Introduction

Gelman and Vehtari (2021) deliver a critical and thought-provoking overview of the most important statistical ideas of the last half-century. In this interesting list of major statistical topics, two methods emerge: causal inference and overparameterized—also known as nonparametric—models.

The framework of causal inference makes its first steps in various applied fields such as econometrics, psychometrics, and epidemiology (Gelman and Vehtari, 2021). In statistics, the concept of causality has been defined in rigorous statistical terms throughout various causal inference frameworks (Zeng and Wang, 2022). With different perspectives, Neyman (1923) and Rubin (1974) propose the potential outcome framework, while Pearl (2009) introduce the causal inference approach based on structural causal models and graphical models (i.e. directed acyclic graph). Both with the ambitious goal of overtaking the traditional statistical inference approach based on correlation across the variables and allow to make*causal* inferences.

Nonparametric models, limited here in the argument for the Bayesian framework, have experienced substantial growth in the last few decades thanks to the most important simulation tools—i.e. MCMC strategies like the Gibbs Sampler and the Metropolis algorithm, etc.— and the software development (Müller and Walker, 2010). Their considerable degree of flexibility induces a high adaptability to the complexity of real-data applications. Their flourishing literature includes Dirichlet processes mixture models (Lo, 1984), Gaussian process (O'Hagan, 1978), Bayesian Additive Regression Tree (Chipman *et al.*, 2010).

Gelman and Vehtari (2021) provocatively suggests that we might increase progress by combining these two approaches into a single Bayesian nonparametric causal inference approach.

Beyond statistics, a pressing topic nowadays is environmental epidemiology. Undoubtedly, the environment plays an essential role in human well-being.

Specifically, exposure to pollutants has been linked to a range of health issues, including respiratory disease, cardiovascular disorders, and even an increased risk of death

Dominici *et al.* (2014); Pope III *et al.* (2019); Wu *et al.* (2020); Dominici *et al.* (2022).

This thesis covers the above-mentioned aspects and specifically shows how Bayesian nonparametric mixture models can be used to answer causal questions arising in environmental health with policy-relevant effects on socioeconomic inequalities, and air pollution regulation.

Overview

Bayesian nonparametric is known for its flexibility and adaptability to different contexts thanks to the ability to capture complex relationships between variables without imposing rigid modeling assumptions (Escobar and West, 1995; Green and Richardson, 2001; Hjort *et al.*, 2010). In fact, a growing literature has produced a large number of Bayesian infinite mixture models according to the different challenges that arise in real-world data scenarios (see e.g. De Iorio *et al.*, 2004; Jara *et al.*, 2010; Caron *et al.*, 2007; Müller *et al.*, 2004; Chung and Dunson, 2009; Denti *et al.*, 2021).

The application of Bayesian nonparametric models in causal inference (Rubin, 1974) has elicited notable attention in the past decade, even though the amount of work is still limited (see the review by Linero and Antonelli, 2023). The main Bayesian nonparametric contributions in this context include the applications or extensions of the Bayesian Additive Regression Tree (Chipman *et al.*, 2010), and dependent Dirichlet Process mixture models (MacEachern, 2000; Barrientos *et al.*, 2012; Quintana *et al.*, 2020).

Initially introduced into the causal inference framework by Hill (2011) and successively exploited by Hahn *et al.* (2020), the Bayesian additive regression tree has been mostly used to capture the heterogeneity in the causal effects in the context of a binary treatment and a continuous outcome. Dependent Dirichlet Process mixture models have been implemented with the goals of addressing various objectives, such as zero-inflation and skewness in the response variable (Oganisian *et al.*, 2021), handling missing random covariates (Roy *et al.*, 2018), dealing post-treatment outcome (Schwartz *et al.*, 2011), and for mediation analysis (Kim *et al.*, 2017; Roy *et al.*, 2022).

Nevertheless, these models have rarely been applied in environmental epidemiological studies. Though extensive literature highlights the interest in examining the causal link between air pollution and various diseases and mortality, empirical evidence consistently reveals that exposure to higher levels of $\text{PM}_{2.5}$ corresponds to heightened mortality and morbidity risks. (Schlesinger, 2007; Rückerl *et al.*, 2011; Chen *et al.*, 2013; Dominici *et al.*, 2014; Wang *et al.*, 2016; Schwartz *et al.*, 2017; Wu *et al.*, 2019, 2020; Lee *et al.*,

2021; Dominici *et al.*, 2022). With a primary focus on the United States, numerous studies have extensively documented that racial and ethnic minorities, often associated with low income, have a higher risk for adverse health outcomes compared to other population/income groups (Bargagli-Stoffi *et al.*, 2023; U.S. Environmental Protection Agency, 2022d; Di *et al.*, 2017; Kioumourtzoglou *et al.*, 2016; Bell *et al.*, 2013; Jbaily *et al.*, 2022).

For this reason, the U.S. Environmental Protection Agency has set a goal to achieve *environmental justice* (U.S. Environmental Protection Agency, 2022b): no group of people should suffer a disproportionate environmental risk (U.S. Environmental Protection Agency, 2022c). However, only a limited number of studies have investigated the implications of the air quality regulations on the reduction of $PM_{2.5}$ and mortality risk (Zigler *et al.*, 2012; Zigler and Dominici, 2014a; Zigler *et al.*, 2018; Samet, 2011). No study today has proposed a data-driven characterization of the heterogeneous treatment effect of air pollution beside using recursive partitioning methods (Lee *et al.*, 2021; Bargagli-Stoffi *et al.*, 2023).

Main contributions of the thesis

To account for this shortcomings, in Chapter 2 and Chapter 3, we present two methodological contributions, each defining suitable Bayesian nonparametric mixtures. These models exploit various dependent Dirichlet processes (MacEachern, 2000; Barrientos *et al.*, 2012; Quintana *et al.*, 2020) as priors, in two challenging contexts within the potential outcome framework (Rubin, 1974). Specifically, these projects address the challenges posed by heterogeneity in causal effects and by the principal stratification framework. For both of them, we also define and provide estimation guidelines for novel causal estimands.

These innovations hold direct applicability in examining the heterogeneity of the causal effects of air pollutants on mortality risk and achieving the associative and dissociative effects of air quality regulations on public health.

Confounder-Dependent Bayesian Mixture Model for Heterogeneity in Causal Effect

In causal inference studies, some observed characteristics play a key role in the identification of heterogeneity in the treatment effect. The conditional average treatment effect is widely used to describe how the treatment effect varies by the characteristics of the population and group identification is crucial to discover and analyze these explanatory factors.

In Chapter 2, we present a Bayesian nonparametric approach that incorporates the

information carried by the observed characteristics, for imputing the missing potential outcomes in case of heterogeneity in the treatment effect by data-driven discovering. Exploiting the flexibility of the dependent Dirichlet Process, we propose the Confounders-Dependent Bayesian Mixture Model to characterize the distribution of the potential outcomes conditionally to the covariates that allow us to: (i) estimate individual treatment effects, (ii) identify heterogeneous and mutually exclusive population groups defined by similar Group Average Treatment Effects (GATEs), and (iii) estimate and characterize causal effects within each of the identified groups.

Specifically, to define the model we take advantage of the feature of Dependent Probit Stick-breaking Process (Chung and Dunson, 2009; Rodriguez and Dunson, 2011) in order to address the challenges and constraints that arise within the framework of causal inference.

The posterior distribution is approximated via a computationally efficient Gibbs sampling algorithm and a pointwise random partition estimator is proposed following Wade and Ghahramani (2018). We estimate the mutually exclusive groups, each characterized by different GATEs, based on the Cartesian product of the latent clusters of each treatment level.

Simulation studies show that the proposed model has competitive results with causal Bayesian additive regression tree (Hill, 2011) and Bayesian Causal Forest (Hahn *et al.*, 2020) for the estimation of individual treatment effect, and has superior performance broadly for the GATEs with respect to a two-step procedure made of a Bayesian causal forest and a subsequent classification and regression tree analysis Breiman *et al.* (1984), that groups the observations according with their individual treatment effect.

Confounders-Aware Shared-atoms Bayesian Mixture Model for Principal Stratification

In causal inference analyses, researchers are often interested in estimating the causal effect of a treatment on a primary outcome and to what extent this effect might change across values of a post-treatment variable. The framework of principal stratification (Frangakis and Rubin, 2002) deals with the causal effect that splits into *direct*—i.e. the effect of the treatment on the outcomes—and *indirect*—i.e. the effect conveyed through the post-treatment variable. Frangakis and Rubin (2002) and VanderWeele (2011) postulate the existence of principal strata: subgroups of individuals with similar values of joint potential outcomes of the post-treatment variable. In particular, the dissociative stratum is defined as the stratum in which the treatment does not substantially change the post-treatment variable and the associative stratum is the stratum in which the

treatment does substantially affect (positively or negatively) the post-treatment variable.

In Chapter 3, we address this shortcoming by developing a novel approach combining principal stratification principles and Bayesian nonparametric methods to assess the causal effect within each stratum. We introduce three major innovations in the principal stratification framework: (i) we rely on Bayesian nonparametric methodologies for the imputation of missing potential outcomes for the post-treatment variables; (ii) we introduce new conditional estimands for the associative and dissociative effects; (iii) we propose a data-driven methodology to discover heterogeneity in the strata memberships and in the associative and dissociative effects.

We define a Confounders-Aware Shared-atoms Bayesian mixture model that, in particular, describes the distribution of the intermediate variable, conditional on the confounders and treatment variable, with a Bayesian non-parametric model that benefits from the flexibility of dependent Dirichlet Process priors and thoughtful revisions, taking into account the need to share information among the potential intermediate variables.

The posterior distribution is approximated via computationally efficient Gibbs sampling algorithm leveraging Fasano *et al.* (2022) for the probit stick-breaking Rodriguez and Dunson (2011) mixture weights.

Socioeconomic Disparities and Regulations in Air Pollution Epidemiology

Several epidemiological studies have demonstrated the significant effects of long-term exposure to fine particulate matter ($PM_{2.5}$) on human health (see, e.g., Schwartz *et al.*, 2017; Wu *et al.*, 2019; Pope III *et al.*, 2019; Chen and Hoek, 2020; Wu *et al.*, 2020; Lee *et al.*, 2021; Dominici *et al.*, 2022) and additional studies have investigated the causal impact on the health of interventions aimed at reducing the level of pollutants in the air (Zigler *et al.*, 2012, 2016, 2018).

Chapter 4 showcases the analysis of two different datasets, exploiting the methodological contributions of Chapter 2 and 3. Specifically, we first analyze the heterogeneity in the effect of $PM_{2.5}$ on the mortality rate within the elderly population in the state of Texas (USA). Then, we undertake a comprehensive examination of principal causal effects—specifically, the associative and dissociative effects—of the air pollution regulations on the mortality rate among Americans, taking into account the heterogeneity in the causal effects through the $PM_{2.5}$ levels.

Leveraging our novel methodological contributions, we reach a comprehensive perspective on understanding the heterogeneity of causal effects and the distinctive characteristics representing the socioeconomic and demographic disparities in the United

States. Our analysis allows us to (i) confirm the established conclusions regarding the effect of high levels of air pollution exposure on increasing mortality rate and the significant effect of regulations, that aspire to reduce air pollution, on decreasing the mortality rate among Medicare enrollees; (ii) identify and characterize the groups—in the first study—and strata—in the second one—that capture the heterogeneity in the causal effect and characterize the health disparities in the American population.

Chapter 1

Overview: Air Pollution Epidemiology, Causality, and Nonparametric Priors.

1.1 Environmental Health and Socio-Economic Disparities

Commonly air pollutants are grouped into six main categories: carbon monoxide, lead, nitrogen oxides, ground-level ozone, particle pollution—also called particulate matter—, and sulfur oxides. Among them, the researchers are particularly interested in the fine air particulate matter $\text{PM}_{2.5}$, defined as those particles having aerodynamic diameters below $2.5\mu g/m^3$. Therefore it is the smallest pollutant that can directly travel to the lungs.

Several epidemiological studies have provided evidence that long-term exposure to air pollutant, and in particular $\text{PM}_{2.5}$, has a direct effect on various health outcomes, including respiratory and cardiovascular diseases, and even mortality (see, e.g., Schlesinger, 2007; Rückerl *et al.*, 2011; Chen *et al.*, 2013; Dominici *et al.*, 2014; Wang *et al.*, 2016; Schwartz *et al.*, 2017; Wu *et al.*, 2019; Pope III *et al.*, 2019; Chen and Hoek, 2020; Wu *et al.*, 2020; Lee *et al.*, 2021; Dominici *et al.*, 2022). Where the word long-term indicates that the causal effect of exposure to air pollution is computed with respect to health outcomes that occur months or years after the exposure (Schwartz, 2006)

Remarkably, focusing our attention primarily on the United States, some characteristics of the population seem to explain the different degrees of vulnerability or resilience to air pollution, as underlined by the U.S. Environmental Protection

Agency. In particular, the U.S. Environmental Protection Agency identifies race, national origin, sex, education, and/or socioeconomic status as potential explanatory factors (U.S. Environmental Protection Agency, 2022a). This aligns with findings from epidemiological studies, which have consistently reported similar associations (see, e.g., Bell *et al.*, 2013; Kioumourtzoglou *et al.*, 2016; Di *et al.*, 2017; Liu *et al.*, 2021; Jbaily *et al.*, 2022).

Accordingly, it becomes imperative to take decisive action within environmental policies, that take into account the population disparities and vulnerability, and ultimately strive to achieve environmental justice (U.S. Environmental Protection Agency, 2022b). Indeed, environmental justice is an ethical stance of the U.S. Environmental Protection Agency, which wants to achieve the desirable scenario where no group of people should have a disproportionate environmental risk (U.S. Environmental Protection Agency, 2022d).

Some research has focused on investigating the implication of air quality regulations and demonstrating the significant causal impact on health and the reduction of the level of pollutants in the air (Samet, 2011; Zigler *et al.*, 2012, 2016, 2018).

In this environmental epidemiology discussion, the definition of causality and causal effects plays a crucial role in quantifying them from a rigorous statistical perspective and for clear scientific communication.

1.2 Causal Inference

Causality is an open debate in statistics beyond the common agreement on the fact that "correlation does not imply causation".

In this thesis, we follow the potential outcome framework (Neyman, 1923; Rubin, 1974), which embodies the principle "no causation without manipulation" (Rubin, 1974; Holland, 1986). However, other frameworks for causal inference are established, including the causal graphical model (Pearl *et al.*, 2000)—see Zeng and Wang (2022) for a review of different causal approaches.

1.2.1 Potential Outcome Framework

An extended literature reviews the fundamentals of potential outcome framework, from the randomized experiments (Fisher, 1919; Neyman, 1923; Fisher *et al.*, 1936; Rubin, 1974, 1978), to the observational studies (Dorn, 1953; Cochran and Chambers, 1965; Cochran and Rubin, 1973). Of particular interest are knotted: more recently Ding and

Li (2018), Dominici *et al.* (2021), Ding (2023), with peculiar Bayesian prospective Li *et al.* (2023); Linero and Antonelli (2023).

The main interest is to study the causal effect of a defined treatment T—called also intervention or exposure—on a specific outcome Y.

Specifically, considering a set of n study units, let $T_i \in \{0, 1\}$ the observed (binary) treatment, with observed value t_i, and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ the outcome. According to Rubin (1974), the potential outcomes for unit i are defined as $\{Y_i(0), Y_i(1)\} \in \mathbb{R}^2$, for $i = 1, \ldots, n$. Specifically, $Y_i(0)$ is the outcome when the unit i is assigned to the control group, while $Y_i(1)$ is the outcome when it is assigned to the treatment group.

In practice, however, for $i = 1, \ldots, n$, we observe only $y_i \in \mathbb{R}$, that is the realization of the random variable Y_i that is defined, invoking the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1986), as

$$Y_i := (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1).$$

Specifically, SUTVA enforces that each unit's outcome is a function of its treatment only. This is a combination of the following assumptions: (i) no interference among the units, i.e. the potential outcome values from unit i do not depend on the treatment applied to other units, and (ii) consistency, i.e. no different versions of the treatment levels assigned to each unit. Conversely, we can not observe the realization $y_i^{mis} \in \mathbb{R}$ of the random variable Y_i^{mis} defined as $Y_i^{mis} := T_i \cdot Y_i(0) + (1 - T_i) \cdot Y_i(1)$.

For this reason, Holland (1986) viewed the fundamental problem of causal inference as a missing data problem, since for each unit only the potential outcome corresponding to the actual treatment is observed, and the other potential outcomes are missing or counterfactual. Moreover, it is Holland (1986) again that called *Rubin's Causal Model* this causal inference view, by referring to the mathematical modeling proposed by Rubin.

In this context, it is a reasonable consequence that the *individual treatment effect* (ITE), defined for each unit i as

$$ITE = Y_i(1) - Y_i(0), \tag{1.1}$$

cannot be observed.

Additionally, we consider $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ a set of p observed covariates. The observed x_i, for $i = 1, \ldots, n$, is a vector of subject-specific background characteristics, considered as covariates and potential confounders—also called pre-treatment variables. Each vector x_i can contain both categorical and continuous variables.

Therefore, the tuple (y_i, t_i, x_i) for $i = 1, \ldots, n$ represents the observed quantities, and the relation among their respective aleatory variables are represented in Figure 1.1.

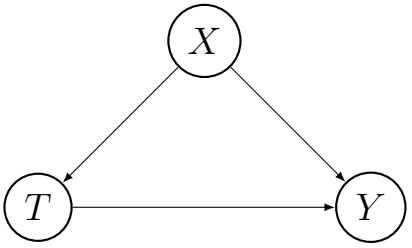

Figure 1.1: Graphical representation of the link between treatment T, outcome Y, and confounders X.

Taking into consideration the variables X, different causal effect are well established in the literature (Dominici *et al.*, 2021; Li *et al.*, 2023), as the expected value of the ITE, that is defined as

$$\bar{\tau}_i = \mathbb{E}[Y_i(1) - Y_i(0)], \tag{1.2}$$

with the expected value respect to the distribution of the outcomes of unit i; and the *population average treatment effect* (PATE)—with attention to distinguish it from *sample average treatment effect* (SATE)— respectively defined as:

$$PATE := \tau^P = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau_i(x)],$$
$$SATE := \tau^S = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i(1) - Y_i(0)\big).$$

While SATE is of interest when the target population is the specific sample, as usual in randomized experiments, PATE is a function of the distribution of the potential outcomes in a population where we just observe a sample (Li *et al.*, 2023).

ITE and PATE can be seen as the two extremes of the definition of *Conditional Average Treatment Effect* (CATE). Indeed, CATE is identified conditionally to a subset of covariates space $C \subseteq \mathcal{X}$, such that

$$CATE := \tau(C) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i \in C]. \tag{1.3}$$

Additionally, an assorted and broad literature is expressed to capture the heterogeneity in the causal effects, defining different estimands and consequently, multifarious methods to compute them (Dahabreh *et al.*, 2016). However, the literature for *heterogeneous treatment effect* (HTE) can be categorize in two main front (Dwivedi *et al.*, 2020; Bargagli-Stoffi *et al.*, 2020): (i) estimating HTEs by examining the CATE; (ii) discovering subgroups of a population classified by HTE and explained by peculiar group-specific

characteristics.

A significant part of them takes advantage of non-parametric machine learning methods such as the random forest (Breiman, 2001) and Bayesian additive regression tree (Chipman *et al.*, 2010) and it has been extended to estimate and capture the heterogeneity in causal effects (Foster *et al.*, 2011; Hill, 2011; Hahn *et al.*, 2020). Moreover, the literature includes two-stage doubly robust CATE estimators by Kennedy (2020); Semenova and Chernozhukov (2021), and various group-base methodologies as Imai and Ratkovic (2013); Qian and Murphy (2011); Kennedy *et al.* (2017); Nie and Wager (2021); Crabbé *et al.* (2022).

1.2.2 Principal Stratification

In experimental and observational studies, researchers are often interested in estimating not only the causal effect of a given treatment on a primary outcome but also to what extent this effect might change across values of a post-treatment variable. Where the causal effect of the treatment on the outcome can be split into *dissociative*——i.e. the effect of the treatment on the outcomes——and *associative*——i.e. the effect of the treatment conveyed through the post-treatment variable to the outcome. Basically, the post-treatment variable is potentially affected by the treatment and also affects the potential outcomes.

This scenario is delineated in the principal stratification framework, introduced by Frangakis and Rubin (2002) (see also Pearl, 2011; Baker *et al.*, 2011; VanderWeele, 2011; Egleston, 2011; Gilbert *et al.*, 2011; Joffe, 2011; Prentice, 2011; Pearl, 2011; Mealli and Mattei, 2012, for interesting discussions and definitions).

Examples of circumstances where intermediate variables may arise in the estimating the causal effect of a given intervention on a given outcome are: in cases of non-compliance with the assigned treatment (i.e. cases where individuals assigned to a particular treatment level fail to comply and receive a different level of treatment); truncation by death (which occurs when an individual exits the study due to death before the outcome is recorded); unintended missing outcome data (which refers to situations where the health records are unintentionally missing for some individuals in the study); or surrogate endpoints (in the case where medium-term trends in health records are used as a substitute measure for long-term health).

Formally, starting from the defined setting in Section 1.2.1, we consider an additional variable: the post-treatment variable P. According with Frangakis and Rubin (2002), we define $P_i \in \mathcal{P} \subseteq \mathbb{R}$, that we can explicit in the *Rubin's Causal Model* as the potential outcome of the post-treatment variables $\{P_i(0), P_i(1)\} \in \mathbb{R}^2$. The vector $\{P_i(0), P_i(1)\}$

represents the collection of the two potential outcomes of the post-treatment variable, specifically $P_i(0)$ is the outcome when the unit i is assigned to the control group while $P_i(1)$ is the outcome when it is assigned to the treatment group.

In practice, as explained for the outcome Y in Section 1.2.1 and invoking SUTVA also for the post-treatment variables, for $i = 1, \ldots, n$, we observe only $p_i \in \mathbb{R}$, that is the realization of the random variable P_i defined as

$$P_i := (1 - T_i) \cdot P_i(0) + T_i \cdot P_i(1).$$

Conversely, we can not observe the realization $p_i^{mis} \in \mathbb{R}$ of the random variable P_i^{mis} defined as $P_i^{mis} := T_i \cdot P_i(0) + (1 - T_i) \cdot P_i(1)$.

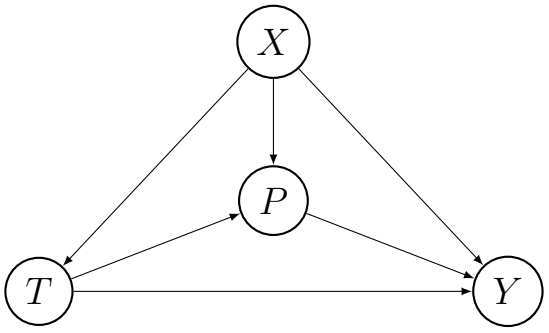

Figure 1.2: Graphical representation of the variables involved in principal stratification framework: treatment T, post-treatment variable P, outcome Y, and confounders X.

In the principal stratification framework, Frangakis and Rubin (2002) postulate the existence of principal strata defined by the joint potential outcomes of the post-treatment variables under treatment and control $\big(P_i(1), P_i(0)\big)$. A principal stratum is simply a subgroup of individuals characterized by similar values in their joint potential outcomes for the post-treatment variables (Mealli and Mattei, 2012; Feller *et al.*, 2017; Ding *et al.*, 2017; VanderWeele, 2011). Consequently, a principal causal effect is a comparison—e.g. the average of the difference—between the potential outcomes within a particular stratum. Therefore, the *expected associative effect* (EAE) is defined as

$$EAE(p_0, p_1) = \mathbb{E}\left[Y_i(1) - Y_i(0) | P_i(0) = p_0, P_i(1) = p_1\right]. \tag{1.4}$$

I.e., EAE is the causal effect of the treatment on the outcome, conditional to the specific values of the potential post-treatment variables. The particular case of $p_0 = p_1$ attracts specific attention, for the reason that this coincides with the dissociative effect, i.e. when the causal effect of the treatment on the outcome does not convey through the

post-treatment variable. It is formalized as the *expected dissociative effect* (EDE):

$$EDE(p) = \mathbb{E}\left[Y_i(1) - Y_i(0) | P_i(0) = P_i(1) = p\right]. \tag{1.5}$$

The definition of the causal estimand of EAE and EDE follows the VanderWeele (2011)'s interpretation of dissociative and associative strata. The *dissociative stratum* is the stratum in which the treatment does not change the post-treatment variable, and the *associative strata* are those groups of units for which the treatment does affect the post-treatment variable, increasing it—positive associative—or decreasing it—negative associative.

However, while these definitions ensure clarity when the post-treatment variable has a discrete nature, different interpretations of the causal effects can arise when the post-treatment variable is continuous.

An important contribution for continuous post-treatment variable is given by Zigler *et al.* (2018), which defines the similarity among post-treatment variable values—and consequently the criteria for the stratification— based on a threshold ξ. Therefore, the EAE and EDE are formalized as

$$\begin{aligned} EAE = \tau_a &= \mathbb{E}\big[Y_i(1) - Y_i(0) \;\big|\; |P_i(1) - P_i(0)| \geq \xi\big], \\ EDE = \tau_d &= \mathbb{E}\big[Y_i(1) - Y_i(0) \;\big|\; |P_i(1) - P_i(0)| < \xi\big]. \end{aligned}$$

However, the discussion revolves around the problem of how to set this threshold ξ. Indeed, this approach can be too sensible for arbitrary choice, where different values of ξ can induce different strata allocation of the analyzed units and, consequently, different estimations of the principal causal effects.

In this thesis, we propose a definition of *similarity* that does not involve thresholds or *a priori* criteria to identify the principal strata and consequently estimate more proper and valid principal causal effects.

Principal stratification framework has some similarities with mediation analysis (Baron and Kenny, 1986; MacKinnon and Dwyer, 1993; MacKinnon, 2012; Imai *et al.*, 2010), however, they generally involve different causal estimands and answer different questions. In particular, while principal stratification is focused on the identification of principal strata and explaining the heterogeneity in the causal effect through them, the mediation analysis is interested in estimating the direct and indirect effects of the treatment on the outcome through the mediators. See Mealli and Mattei (2012) for an overview and comparison of the two methods.

Notably interesting for the topics of this thesis, wide literature in mediation analysis exploits nonparametric Bayesian models, such as Daniels *et al.* (2012) and Roy *et al.* (2018, 2022).

1.3 Bayesian Nonparametrics in Causal Inference

In the potential outcomes framework, there are mainly three models of inference (Ding and Li, 2018): Fisherian randomization test, Neymanian repeated-sampling evaluation, and Bayesian posterior inference. However, the Bayesian paradigm is natural to handle the missing data, therefore Bayesian models allow straightforward missing potential outcome imputation (Li *et al.*, 2023).

In particular, this thesis focuses on Bayesian nonparametric methods, inasmuch as their considerable degree of flexibility induces high adaptability to the complexity of real-data applications and intrinsic mixture structure induces the clustering among the observations, that characterizes the heterogeneity in the distributions of the involved variables. To focus on the latter we avoid digging out Bayesian parametric models. See Li *et al.* (2023) for a recent general review of Bayesian methods and models in causal inference. Bayesian nonparametric methods have recently seen increased interest in the causal inference literature (Linero and Antonelli, 2023), since their flexibility and statistically principled uncertainty quantification, and thanks to the recent advancements in computation and software (Li *et al.*, 2023).

Linero and Antonelli (2023) and Daniels *et al.* (2023) center their discussion on the causal inference challenges that can potentially be addressed through the utilization of Bayesian nonparametric methods. However, exploiting Bayesian nonparametric models and methods in different contexts of causal inference is not just implementing them as in a standard prediction problem, but careful considerations have to be taken for the causal assumptions and the estimands. In fact, regularization induced confounding could occur using Bayesian nonparametric or, in general, high-dimensional methods (Hahn *et al.*, 2018). That is when regularization techniques, introduced to improve the performance or stability of a model, unintentionally introduce confounding factors into the analysis.

Among these main works, Bayesian Additive Regression Tree (Chipman *et al.*, 2010) became widely used, thanks to the first introduction in causal inference by Hill (2011), that models the potential outcomes following the T-learner methods as

$$Y_i(t) = \mu(t, X_i) + \epsilon_i(t),$$

for $t = 0, 1$, where the function μ has Bayesian additive regression tree prior and ϵ are Gaussian errors. Recently, Hahn *et al.* (2020) proposed a reparameterization of the outcome model including the propensity score and defining the Bayesian causal forest as:

$$Y_i = \mu(t, \hat{\pi}(x_i)) + \tau(x_i)t + \epsilon_i,$$

where μ and τ are independent Bayesian additive regression tree priors and $\hat{\pi}(x_i)$ is the estimated propensity score $\pi(x_i) = \mathbb{P}(T_i = 1|x_i)$.

The propensity score is included in the model to further control measured confounding. However, while propensity score is commonly used in the frequentist context, its use is still discussed in the Bayesian framework. Stephens *et al.* (2022) discusses the controversial role of Bayesian approaches to causal inference via propensity score regression and reviews these existing approaches.

Infinite mixture models have also been widely employed in causal inference literature. Notably, among the various dependent Dirichlet process mixture models (MacEachern, 2000; Barrientos *et al.*, 2012; Quintana *et al.*, 2020), the Enriched Dirichlet process mixture model (Wade *et al.*, 2011, 2014) and its modifications have found extensive and promising causal applications (Roy *et al.*, 2018; Oganisian *et al.*, 2020a,b, 2021; Roy *et al.*, 2022), and the Probit Stick-Breaking process mixture model (Rodriguez and Dunson, 2011) has been used in principal stratification framework (Schwartz *et al.*, 2011).

Other Bayesian nonparametric approaches have been employed for causal inference problems as the spike and slab prior distribution (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005; Hahn and Carvalho, 2015) to regularize many of the regression coefficients in high-dimensional settings (Zigler and Dominici, 2014b; Hahn *et al.*, 2018), and the Bayesian bootstrap (Rubin, 1981) to compute different estimands (Linero and Antonelli, 2023).

Since the discussion in this thesis regards approaches based on outcome modeling, it is important to remember that a straightforward application of the standard Bayesian nonparametric priors is sometimes inadequate for causal inference and a desirable prior should accurately reflect uncertainty according to the degree of covariate overlap (Li *et al.*, 2023).

1.4 Bayesian Nonparametric Mixture Models

Among the plethora of Bayesian nonparametric methods and models, we focus on the mixture models that take advantage of the Dirichlet process (Ferguson, 1973,

1974)—and, in the following section, the relative dependent Dirichlet process (Lo, 1984; MacEachern, 2000)—as prior. The choice to concentrate our interest on these models is due to their desirable characteristics, such as flexibility for describing complex distributions and variable relations, adaptability in various real-world applications, ability to capture the clustering structure, and to handle easily missing data.

Given a random variable Y, that takes values y in the space $\mathcal{Y}$, a Dirichlet process mixture model assumes, for the distribution f of Y

$$f(y|G) = \int_{\Psi} \mathcal{K}(y, \psi) dG(\psi); \tag{1.6}$$

where $\mathcal{K}(\cdot, \psi)$ is a continuous density function, for every $\psi \in \Psi$, and G is an almost surely discrete random probability measure. Defining G as a Dirichlet process with parameter $\alpha \in \mathbf{R}_0^+$ and G_0 probability measure on Ψ—i.e. $G|\alpha, G_0 \sim DP(\alpha G_0)$—, then the process can be written with the stick-breaking representation, introduced by Sethuraman (1994):

$$G(B) = \sum_{l=1}^{\infty} \omega_l \delta_{\psi_l}(B),$$

where B is any measurable set, $\delta_{\psi}(\cdot)$ is the Dirac measure at ψ, and $\{\omega_{\psi}\}_{l \geq 1}$ and $\{\psi_l\}_{l \geq 1}$ represent infinite sequences of random weights and random kernel's parameters, respectively. Specifically, the random kernel's parameters are defined as

$$\psi_l | G_0 \overset{iid}{\sim} G_0;$$

and the random weights are defined as

$$\omega_l = V_l \prod_{r<l} (1 - V_l) \text{ with } V_l | \alpha \sim Beta(1, \alpha),$$

where $Beta(\cdot)$ is a beta distribution with two shape parameters such that the strike-breaking ratios $\{V_l|\alpha\}_{l \geq 1}$ are random variables that take values in $[0, 1]$ with mean $\frac{1}{1+\alpha}$ and variance $\frac{\alpha}{(\alpha+1)^2(\alpha+2)}$, with $\alpha > 0$.

Modeling a distribution as an infinite mixture of simpler distributions is useful both as a nonparametric density estimation method and as a way of identifying latent clusters that can explain the behavior of observed variables (Neal, 2000).

There are numerous alternatives and generalizations of the Dirichlet process that have been introduced to face several applied problems in the context of mixture models (see for instance Perman *et al.*, 1992; Pitman and Yor, 1997; Lijoi *et al.*, 2007a,b; De Blasi *et al.*, 2013). Among them, the dependent Dirichlet process mixture model

rises to answers to a common scientific question of interest: taking into account that the distribution of responses often changes with predictors in applied research (MacEachern, 2000; Quintana *et al.*, 2020).

1.4.1 Dependent Nonparametric Mixture Models

The dependent Dirichlet process mixture models can be seen as a generalization of Dirichlet process mixture models, where we want to model the distribution of the response $y \in \mathcal{Y}$ given the predictors $x \in \mathcal{X} \subseteq \mathbb{R}^p$ (Quintana *et al.*, 2020). Many of these dependent Dirichlet process mixture models incorporate dependence on predictors x in the random probability measure G, such that the model definition in (1.6) can be written as

$$f(y|x, G_x) = \int_{\Psi} \mathcal{K}(y; x, \psi) dG_x(\psi),$$

where $\{G_x : x \in \mathcal{X}\}$ is the collection of predictor-dependent mixing probability measures. Specifically, following the stick-breaking representation (Sethuraman, 1994) of each element of the set, we have that

$$\begin{aligned} G_x(\cdot) &= \sum_{l=1}^{\infty} \omega_l(x) \delta_{\psi_l(x)}(\cdot), \\ \omega_l(x) &= V_l(x) \prod_{r<l} [1 - V_r(x)]; \end{aligned}$$

where $\psi_l(x)$ are independent stochastic processes with index set $\mathcal{X}$ and G_x^0 marginal distributions, and $V_l(x)$, with $l \in \mathbb{N}$, are $[0, 1]$-valued independent stochastic processes with index set $\mathcal{X}$ and $Beta(1, M_{\mathcal{X}})$ marginal distributions. $M_{\mathcal{X}} > 0$, similarly to α in the Dirichlet process, is a collection of parameters that controls the precision of the stochastic process. Specifically, the distribution of $V_l(x)$ can differ from the distribution of $V_l(x')$ if $x \neq x'$, for any $l \geq 1$, inducing different weights associated with the components of the mixture $G_x(\cdot)$. The processes associated with the weights and atoms are independent.

Therefore, the dependent Dirichlet process construction is defined by a set of random measures that are marginally—i.e. for every possible predictor value $x \in \mathcal{X}$—Dirichlet process distributed random measures (Quintana *et al.*, 2020).

Almost all of the proposed dependent Dirichlet process mixture models in the literature can be categorized, following Quintana *et al.* (2020), in *single-weights* dependent Dirichlet process models (see De Iorio *et al.*, 2004; Gelfand *et al.*, 2005) and in *single-atoms* dependent Dirichlet process models (see Dunson and Park, 2008; Rodriguez *et al.*,

2008). The first category is defined by MacEachern (2000) as the case of common weights across the values of predictors x, while the atoms depend on them, such that

$$G_x(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{\psi_l(x)}(\cdot) = \sum_{l=1}^{\infty} \left[V_l \prod_{r<l} (1 - V_l) \right] \delta_{\psi_l(x)}(\cdot),$$

where $\{V_l\}_{l\geq 1}$ are iid Beta distribution $Beta(1, \alpha)$, common across all levels of x, as in the Dirichlet process; while the atoms $\{\psi_l(x)\}_{l\geq 1}$ are independent stochastic processes with index set $\mathcal{X}$ and marginal distribution G_x^0.

Included in this definition, is the ANOVA-DDP model introduced by De Iorio *et al.* (2004), where the atom $\psi_l(x)$ is a sequence of an iid random vector of a linear combination of the observed categorical covariates x, and extended to different real-data problems as longitudinal data (De la Cruz-Mesía *et al.*, 2007), multiple treatment testing (Gutiérrez *et al.*, 2019), and nonproportional hazards for survival analysis (De Iorio *et al.*, 2009). Similarly, the linear dependent Dirichlet process model involves a linear combination of a set of covariates (Jara *et al.*, 2010). Gelfand *et al.* (2005) defines the spatial-dependent Dirichlet process mixture model where the weights depend on the spatial location and are defined as a Gaussian process. Caron *et al.* (2007) and Rodriguez and Ter Horst (2008) formulate a dependent Dirichlet process model for dynamic phenomena, where the atoms in the infinite mixture are allowed to change in time.

The second category, the single-atoms dependent Dirichlet process models, is characterized by a set of common atoms across all the values of x, such that the random probability measure takes the form:

$$G_x(\cdot) = \sum_{l=1}^{\infty} \omega_l(x) \delta_{\psi_l}(\cdot) = \sum_{l=1}^{\infty} \left[V_l(x) \prod_{r<l} (1 - V_l(x)) \right] \delta_{\psi_l}(\cdot),$$

where $V_l(x)$, with $l \in \mathbb{N}$, are $[0, 1]$-valued independent stochastic processes with index set $\mathcal{X}$ and $Beta(1, M_{\mathcal{X}})$ marginal distributions, and the parameters $\{\psi_l\}_{l\geq 1}$ are independent with marginal distributions G^0.

Under this definition, all the covariate-dependence is expressed through the weights of the stick-breaking representation, with the implication of obtaining partitions that change with the values of $x \in \mathcal{X}$ (Quintana *et al.*, 2020). This advantage, not shared with the single-weights dependent Dirichlet process, assumes a substantial role when the partition and the latent class are a matter of interest.

Moreover, several models extend this original definition, modifying the stick-breaking representation to allow incorporation of the covariates dependence. Dunson and Park

(2008) propose the kernel stick-breaking process where the weights are a function of beta random variables and random kernel locations. Chung and Dunson (2009) and Ren *et al.* (2011) introduce the probit and logic stick-breaking processes, substituting the Beta random variables with normally distributed random variables with probit and logic link function respectively. A hierarchical structure has been considered in different variations of dependent Dirichlet process, as the hierarchical Dirichlet process mixture model proposed by Müller *et al.* (2004) or that one by Teh *et al.* (2006), the nested Dirichlet process by Rodriguez *et al.* (2008), and the common atom model (Denti *et al.*, 2021).

Chapter 2

Confounder-Dependent Bayesian Mixture Model for Heterogeneity in Causal Effect

The broad interest in capturing the heterogeneity in the causal effect within the Potential Outcome framework (Rubin, 1974)—introduced in Section 1.2.1—is a reflection of the wide number of applications where heterogeneity can arise. However, this notion is conceptualized in different estimands and, consequently, estimated with different methods.

In this chapter, we define the estimand for the Group Average Treatment Effect (GATE) and introduce the novel method that allows us to estimate GATEs: Confounder-Dependent Bayesian Mixture Model.

2.1 Estimands for Heterogeneous Causal Effects

As introduced in Section 1.2.1, the general estimand for the heterogeneous treatment effect is the CATE, defined in (1.3) as the expected value of ITE conditioning to a subset $C \subseteq \mathcal{X}$.

The CATE can be specified at different levels of granularity depending on the complexity of C, where at the highest level of granularity stands the IATE—see (1.2) in Section 1.2.1—, and at a lower level of granularity, the average treatment effect for some subgroups of the population (Bargagli-Stoffi *et al.*, 2022).

This chapter is based on Zorzetto *et al.* (2023a).

We assume that the heterogeneity of the treatment effect is induced by a group structure where the units within a group are characterized by the same individual treatment effect. Let's define G_i as a subject-specific group label indicator and g as one of the labels. Then, the *Group Average Treatment Effect* (GATE) for group g is defined as

$$\tau_g = \mathbb{E}\left[Y_i(1) - Y_i(0) | G_i = g\right]. \tag{2.1}$$

As a specific case of CATE, GATE depends on subsets of the covariates space $\mathcal{X}$. However, most of the approaches for estimating the CATE or GATE require defining these covariates values *a priori*. Such approaches lead to disadvantages as (i) they can be subject to the *cherry-picking* problem of reporting results only for groups with extremely high/low treatment effects (Cook *et al.*, 2004); (ii) they must define the groups a priori, which in turn requires a good understanding of the treatment effects, possibly from previous literature and, thus, may fail to identify unexpected, yet important, heterogeneous groups.

Aiming to overcome these limitations, we propose, in the following section, a novel nonparametric method that allows us to identify mutually exclusive groups, that are driven by the data, identified by different conditional treatment effects, and distinguished by different values of covariates. Specifically, as a feature of the proposed model, partitions of the data are induced by definition, without a priori assumptions on the probability of the units being grouped together.

Consistently with the causal inference framework (Rubin, 1980) the following assumption is required, in addition to SUTVA explained in Section 1.2.1, to identify any causal effects from the observed data.

Strong Ignorability. Given the observed covariate vector x_i, the treatment assignment is strongly ignorable if

$$\{Y_i(1), Y_i(0)\} \perp T_i \mid X_i = x_i,$$

and $0 < \mathbb{P}(T_i = 1 \mid X_i = x_i) < 1$, for all $i = 1, \ldots, n$. This assumption states that: (i) we have a random treatment assumption in each group conditional on some covariates values; (ii) all units have a positive chance of receiving the treatment.

If the SUTVA and strong the ignorability assumption hold, the estimand in (1.2) can be expressed as

$$\begin{aligned} \tau(x) &= \mathbb{E}[Y_i(1) \mid X_i = x] - \mathbb{E}[Y_i(0) \mid X_i = x] \\ &= \mathbb{E}[Y_i \mid X_i = x, T_i = 1] - \mathbb{E}[Y_i \mid X_i = x, T_i = 0]; \end{aligned} \tag{2.2}$$

and consequently also the estimand of GATE (2.1) can be defined as:

$$\begin{aligned}
\tau_g &= \mathbb{E}[Y_i(1) \mid G_i = g] - \mathbb{E}[Y_i(0) \mid G_i = g] \\
&= \int_x \mathbb{E}[Y_i(1) \mid G_i = g, X_i = x] \Pr(X_i = x \mid G_i = g) dx \\
&\quad - \int_x \mathbb{E}[Y_i(0) \mid G_i = g, X_i = x] \Pr(X_i = x \mid G_i = g) dx \\
&= \int_x \Big(\mathbb{E}[Y_i \mid G_i = g, X_i = x, T_i = 1] - \mathbb{E}[Y_i \mid G_i = g, X_i = x, T_i = 0] \Big) \\
&\quad \times \Pr(X_i = x \mid G_i = g) dx; \qquad (2.3)
\end{aligned}$$

where the first equality invokes the properties of expectation and the second equality is due to the no unmeasured confounding and SUTVA assumption. The expected values and the probability can be estimated by the outcome model proposed in the following section.

2.2 Confounder-Dependent Bayesian Mixture Model

The estimation of the causal effects can be seen as a missing data problem where, for each subject, we observe just one of the potential outcomes while the other potential outcome is always missing. Likewise, Rubin (1974) refers to the missing potential outcomes as counterfactual outcomes (Holland, 1986; Imbens and Rubin, 1997). From (2.3), we know that under strong ignorability—that is, under a sufficiently rich collection of control variables—treatment effect estimation reduces to the estimation of the conditional expectations of $\mathbb{E}[Y_i \mid X_i = x, T_i = 1]$ and $\mathbb{E}[Y_i \mid X_i = x, T_i = 0]$. Provided the excellent predictive performance of Bayesian methodologies, Bayesian nonparametric models have been widely used for this task (Sivaganesan *et al.*, 2017; Hill, 2011; Roy *et al.*, 2018; Hahn *et al.*, 2020; Oganisian *et al.*, 2021).

Here we propose a Bayesian nonparametric approach for the expectation of conditional outcomes. In particular, we exploit a dependent nonparametric mixture prior—inspired by the dependent Dirichlet process (MacEachern, 2000; Barrientos *et al.*, 2012;

Quintana *et al.*, 2020). Formally, we assume for each $i = 1, \dots, n$:

$$\begin{aligned} \{Y_i \mid x_i, t\} &\sim f^{(t)}(\cdot \mid x_i), \\ f^{(t)}(\cdot \mid x_i) &= \int_{\Psi} \mathcal{K}(\cdot; \psi) dG_{x_i}^{(t)}(\psi), \\ G_{x_i}^{(t)} &\sim \Pi_{x_i}^{(t)}, \end{aligned} \tag{2.4}$$

where $\mathcal{K}(\cdot; \psi)$ is a continuous density function, for every $\psi \in \Psi$, and $G_{x_i}^{(t)}$ is a random probability measure depending on the confounders x_i associated to an observation assigned to treatment level t. A priori we assume $G_{x_i}^{(t)} \sim \Pi_{x_i}^{(t)}$ where $\Pi_{x_i}^{(t)}$ is a treatment- and confounder-dependent nonparametric process. Following a single-atom dependent Dirichlet process (Quintana *et al.*, 2020) characterization of the random measure $G_{x_i}^{(t)}$, we can write:

$$G_{x_i}^{(t)} = \sum_{l \geq 1} \omega_l^{(t)}(x_i) \delta_{\psi_l^{(t)}}, \tag{2.5}$$

where $\{\omega_l^{(t)}(x_i)\}_{l \geq 1}$ and $\{\psi_l^{(t)}\}_{l \geq 1}$ represent infinite sequences of random weights and random kernel's parameters, respectively. Notably, both random sequences depend on treatment level t while the weights also depend on the confounders values x_i.

Furthermore, the sequence of dependent weights is defined through a stick-breaking representation (Sethuraman, 1994),

$$\omega_l^{(t)}(x_i) = V_l^{(t)}(x_i) \prod_{r<l} \{1 - V_r^{(t)}(x_i)\}, \tag{2.6}$$

where $\{V_l^{(t)}(x)\}_{l \geq 1}$ are $[0, 1]$-valued independent stochastic processes. The sequence of random parameters $\{\psi_l^{(t)}\}_{l \geq 1}$ are independent and identically distributed from a base measure $G_0^{(t)}$.

The discrete nature of the random probability measure $G_{x_i}^{(t)}$ allows us to introduce the latent categorical variables $S_i^{(t)}$, which identifies the cluster allocation for each unit $i \in \{1, \dots, n\}$, whose clusters are defined by heterogeneous responses to the treatment level t. Assuming $\mathbb{P}\{S_i^{(t)} = l\} = \omega_l^{(t)}(x_i)$, we can write model in (2.4), exploiting conditioning on $S_i^{(t)}$, as

$$\{Y_i | x_i, t, \psi^{(t)}, S_i^{(t)} = l\} \sim \mathcal{K}(\cdot \mid \psi_l^{(t)}), \quad \psi_l^{(t)} \sim G_0^{(t)}.$$

where $\psi^{(t)}$ represents the infinite sequence $\{\psi_l^{(t)}\}_{l \geq 1}$, for $t = \{0, 1\}$.

Among the plethora of dependent nonparametric processes (see the recent review by Quintana *et al.*, 2020, for a detailed description), we focused on the dependent Probit

Stick-Breaking process for its success in applications, good theoretical properties, and ease of computation (Rodriguez and Dunson, 2011). Consistently with this, we specify:

$$V_l^{(t)}(x_i) = \Phi(\alpha_l^{(t)}(x_i)), \quad \alpha_l^{(t)}(x_i) \sim \mathcal{N}(\beta_{0l}^{(t)} + x_i^T \beta_l^{(t)}, 1), \tag{2.7}$$

where $\Phi(\cdot)$ is the CDF of a standard Normal distribution and $\{\alpha_l^{(t)}(x_i)\}_{l\geq 1}$ has Gaussian distributions with mean a linear combination of the p covariates x_i.

As commonly done, we assume the kernel $\mathcal{K}$ to be a Gaussian, so that model (2.4)–(2.5) specifies to

$$\{Y_i | x_i, t, S_i^{(t)} = l, \eta^{(t)}, \sigma^{(t)}\} \sim \mathcal{N}(\eta_l^{(t)}, \sigma_l^{2(t)}). \tag{2.8}$$

where $\eta^{(t)}$ and $\sigma^{(t)}$ represent the infinite sequences $\{\eta_l^{(t)}\}_{l\geq 1}$ and $\{\sigma_l^{(t)}\}_{l\geq 1}$, respectively.

Prior elicitation is completed by assuming for the regression parameters in (2.7) the conjugate priors

$$\beta_{ql}^{(t)} \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2),$$

for $t = 0, 1$, $l \geq 1$,and $q = 0, 1, \ldots, p$ and for the parameters $\eta_l^{(t)}$ and $\sigma_l^{(t)}$ in (2.8)

$$\eta_l^{(t)} \overset{iid}{\sim} \mathcal{N}(\mu_\eta, \sigma_\eta^2), \text{ and } \sigma_l^{(t)} \overset{iid}{\sim} \text{InvGamma}(\gamma_1, \gamma_2).$$

where $\text{InvGamma}(\gamma_1, \gamma_2)$ represents the inverse-gamma distribution with shape parameter $\gamma_1 \in \mathbb{R}^+$ and scale parameter $\gamma_2 \in \mathbb{R}^+$, and mean equal to $\frac{\gamma_2}{\gamma_1 - 1}$ and variance $\frac{\gamma_2^2}{(\gamma_1-1)^2(\gamma_1-2)}$.

In observation studies, it is fundamental the control the confounding bias. Therefore, in the real-data application, the study design is required before the fitting of the CDBMM.

2.2.1 Posterior Inference

Rodriguez and Dunson (2011) proves that the finite truncation of the dependent Probit Stick-Breaking process is a good approximation; therefore, we can rewrite (2.5) as a finite mixture to $L < \infty$ components with L a reasonable conservative upper bound. Rodriguez and Dunson (2011)'s proof is a key point that allows us to provide a simpler algorithm without losing the robustness of the model.

The Gibbs sampling algorithm for model fitting, which we define below, is inspired by the algorithm proposed by Rodriguez and Dunson (2011) to obtain draws from the posterior distribution. Following the steps in the Algorithm 1, in each iteration

$r = 1, \ldots, R$, we use the observed data (y, t, x) to update the parameters and the augmented variables and impute the missing potential outcomes y^{mis}.

Cluster Allocation. The latent variables $S_i^{(t)}$ identifies the cluster allocation for each units $i \in \{1, \ldots, n\}$ at the treatment level t. Its posterior distribution is a multinomial distribution where

$$\mathbb{P}\{S_i^{(t)} = l\} \propto \omega_l^{(t)}(x_i)\mathcal{N}(y_i; \eta_l^{(t)}, \sigma_l^{(t)2}),$$

for $i = 1, \ldots, n$ and $l = 1, \ldots, L$, with $\omega_l^{(t)}$ defined as:

$$\omega_l^{(t)}(x_i) = \Phi(\alpha_l^{(t)}(x_i)) \prod_{r<l} (1 - \Phi(\alpha_l^{(t)}(x_i))),$$

for $l = 1, \ldots, L-1$ and with $\Phi(\alpha_L^{(t)}(x_i)) = 1$.

Cluster Specific Parameters. Thanks to the latent variables $S_i^{(t)}$, that cluster the units by the value of their outcome for the treatment level t, we know for each cluster $l \in \{1, \ldots, L\}$, the allocated units and we can update the values of the parameters from their posterior distributions:

$$\eta_l^{(t)} \sim \mathcal{N}\left(V_l^{-1} \times \left(\frac{\sum_{\{i:S_i^{(t)}=l\}} y_i(t)}{\sigma_l^{(t)2}} + \frac{\mu_\eta}{\sigma_\eta^2}\right), V_l^{-1}\right), \text{ for } l = 1, \ldots, L;$$

$$\sigma_l^{(t)2} \sim \text{InvGamma}\left(\gamma_1 + \frac{n_l^{(t)}}{2}, \gamma_2 + \frac{\sum_{\{i:S_i^{(t)}=l\}} (y_i(t) - \eta_l^{(t)})^2}{2}\right), \text{ for } l = 1, \ldots, L;$$

where $V_l = n_l^{(t)}/\sigma_l^{(t)2} + 1/\sigma_\eta^2$ and $n_l^{(t)}$ is the number of units allocated in the l-th cluster.

Augmentation Scheme. In order to sample from $\{\alpha_l^{(t)}(x)\}_{l=1}^L$ and the corresponding weights $\{\omega_l^{(t)}(x)\}_{l=1}^L$, we need a data augmentation scheme. The idea was developed by Albert and Chib (2001) and borrowed by Rodriguez and Dunson (2011) to obtain exact Bayesian inference for binary regression and computationally easy to include it in the Gibbs sampling (Albert and Chib, 2001). We can impute the augmented variables $Z_l^{(t)}(x_i)$ by sampling from its full conditional distribution (Rodriguez and Dunson, 2011):

$$Z_l^{(t)}(x_i)|S_i^{(t)}, \alpha_l^{(t)}(x_i) \sim \begin{cases} \mathcal{N}(\alpha_l^{(t)}(x_i), 1)\mathbb{I}_{\mathbb{R}^+} \text{ if } S_i^{(t)} = l, \\ \mathcal{N}(\alpha_l^{(t)}(x_i), 1)\mathbb{I}_{\mathbb{R}^-} \text{ if } S_i^{(t)} < l. \end{cases}$$

The mean, $\alpha_l^{(t)}(x_i)$, of the previous normal distributions is obtained from:

$$\alpha_l^{(t)}(x_i) = \phi\left(\frac{\omega_{il}^{(t)}(x_i)}{\prod_{r<l}(1-\Phi(x_i^T\beta_l^{(t)})}\right) = \phi\left(\frac{\omega_{ir}^{(t)}(x_i)}{1-\sum_{r<l}\omega_{ir}^{(t)}(x_i)}\right);$$

where $\phi(\cdot)$ is the continuous density function of Gaussian distribution.

Confounder-Dependent Weights. To conclude the for-loop, the $\{\beta_{ql}^{(t)}\}_{q=0}^{p} = (\beta_{0l}^{(t)}, \beta_l^{(t)})$, for $l = 1, \ldots, \max(S_i^{(t)}, L-1)$, are updated for the posterior distribution:

$$\beta_{0l}^{(t)} \sim \mathcal{N}((1/\sigma_\beta^2 + n)^{-1} \times (\mu_\beta/\sigma_\beta^2 + 1_n^T\tilde{\mathbf{Z}}), (1/\sigma_\beta^2 + n)^{-1});$$
$$\beta_l^{(t)} \sim \mathcal{N}_p(W^{-1} \times (\mu_\beta/\sigma_\beta^2 + (\tilde{X})^T\tilde{\mathbf{Z}}), W^{-1});$$

where 1_n is a n vector of ones, $W = I_p/\sigma_\beta^2 + (\tilde{X})^T\tilde{X}$, I_p is a $p \times p$ diagonal matrix, $\tilde{X}$ is a matrix such that it is composed by the rows i in X, such that $S_i^{(t)} \leq l$, and $\tilde{\mathbf{Z}}$ is a vector composed by the $z_l^{(t)}(x_i)$ for the units i such that $S_i^{(t)} \leq l$.

While the Gibbs sampler allows us to recover the posterior distribution of the random variable, it is crucial to identify the point estimation of the quantities of interest. According to the goal to estimate the GATEs, the point estimation of the partition for latent variables $S^{(0)}$ and $S^{(1)}$, respectively, is the first step. As explained with further details and intuitions in the following section, this estimation is obtained through the Wade and Ghahramani (2018)'s methods. Successively, the group allocation is identified by the Cartesian product of the two partitions of $S^{(0)}$ and $S^{(1)}$. The GATEs are computed conditional to the group allocation.

2.2.2 Groups & Clusters

One of the advantages of Bayesian nonparametric mixtures is their ability to cluster the observations. Consistently with our goal of defining heterogeneous causal effects, herein we discuss how to estimate mutually exclusive *groups* of observations, each characterized by different GATEs.

Consistently with the Bayesian nonparametric literature, we call *clusters* the sets defining the estimated partition for each treatment level $t \in \{0, 1\}$. We then combine these clusters to estimate the *groups* into which the observations are divided, and for each group, we calculate a different GATE. Note that we use the term *cluster* differently from *group*. The former refers to the sets defined for a specific t as a byproduct of the

Algorithm 1 Estimation Confounder-Dependent Mixture Model (CDBMM)

Inputs:
- the observed data (y, t, x).

Outputs:
- posterior distributions of parameters: η, σ, and β;
- posterior distribution over the space of partitions of the units.

Procedure:
Initialization of all parameters and latent variables.
For $r \in \{1, \ldots, R\}$ **:**
  For $t \in \{0, 1\}$ **:**
    Compute $\omega^{(t)}(x_i)$ for $i = 1, \ldots, n$;
    Draw $S_i^{(t)}$ for $i = 1, \ldots, n$;
    Draw $\eta^{(t)}$ and $\sigma^{(t)}$;
    Compute $\alpha^{(t)}(x_i)$ for $i = 1, \ldots, n$;
    Draw $z^{(t)}(x_i)$ for $i = 1, \ldots, n$;
    Draw $\beta^{(t)}$.
  End
End

infinite mixture model specification obtained through Wade and Ghahramani (2018) procedure, while the latter refers to the final groups for which we computed different GATE.

The model specification introduced in the previous section allows us to define a group of observation based on the Cartesian product of the latent categorical variables $\{S_i^{(0)}, S_i^{(1)}\}$, for each unit $i = 1, \ldots, n$. Under our fully Bayesian approach, $S_i = \{S_i^{(0)}, S_i^{(1)}\}$ are couples of random variables for which we can characterize the associated posterior distribution—reported in the previous section. This is customary in all Bayesian infinite mixture models, which are inherently associated with random partition models (Quintana, 2006). Under these settings, the posterior of the random partition reflects uncertainty in the clustering structure given the data (Wade and Ghahramani, 2018).

A general problem associated with the huge dimension of the space of partitions is the appropriate summarization of the posterior through a point estimate. Wade and Ghahramani (2018) propose a solution based on decision theory—i.e. the optimal point estimate is that which minimizes the posterior expectation of a loss function using either on Binder's loss (Binder, 1978) or Variation of Information (Meilă, 2007).

In the following analyses, we use the approach proposed by Wade and Ghahramani (2018) to divide the observations into separate clusters for each value of $t \in \{0, 1\}$. By using this approach, we can associate different treatments with varying numbers of clusters, which allows for a highly flexible model. For instance, in the simulations presented

in Section 2.3, we designed Scenario 2 to have a single cluster with a constant response for the treated subjects and different clusters with varying responses depending on the confounders for the untreated group. Our proposed approach effectively accommodates such scenarios.

Our proposed method estimates the GATE of each group without the need for a predefined selection of a partition of the confounder space. This is achieved by directly obtaining the groups from the posterior of the model. Given that the groups are mutually exclusive, it is straightforward to identify the characteristics of the units in each group, such as the average of observed confounders or modal categories for continuous and categorical confounders, respectively.

We define the posterior distribution of the GATE for each group as the mean of the posterior distribution of the ITE for units in that group. In the following analyses, we use the mean as the posterior point estimation of GATE for each group, but other measures, such as the median, can also be used. Bayesian credible intervals for the GATE can be obtained as well.

However, our proposed method allows us to define and estimate any functions of the potential outcomes conditional to the group allocation. An example is the *Group Average Risk Ratio* (GARR), used in Chapter 4, that wants to characterize the risk ratio between the potential outcomes given the group g. Specifically, the causal estimand of GARR is defined as follows:

$$GARR_g = \mathbb{E}\left[\frac{Y_i(1)}{Y_i(0)}\middle| G_i = g\right].$$

Under strong ignorability and properties of expectation, the statistical estimand results to be

$$\begin{aligned} GARR_g &= \frac{\mathbb{E}[Y_i(1) \mid G_i = g]}{\mathbb{E}[Y_i(0) \mid G_i = g]} \\ &= \frac{\int_x \mathbb{E}[Y_i(1) \mid G_i = g, X_i = x] \Pr(X_i = x \mid G_i = g) dx}{\int_x \mathbb{E}[Y_i(0) \mid G_i = g, X_i = x] \Pr(X_i = x \mid G_i = g) dx} \end{aligned} \tag{2.9}$$

where G_i is the Cartesian product of the point estimate of the cluster partition of $S_i^{(0)}$ and $S_i^{(1)}$.

2.3 Simulation Study

The performances of the proposed Confounder-Dependent Bayesian Mixture Model (CDBMM) are assessed through a simulation study. Our objective is to investigate the

model's ability to (i) accurately estimate ITEs—i.e. evaluating bias and mean square error (MSE) of the SATE—, (ii) correctly identify the groups of data that describe the heterogeneity in the effects, and, as a result, accurately estimate GATEs. To achieve this, we conduct simulations under seven different data-generating models and analyze the results to understand the model's behavior in different scenarios.

Specifically, in each scenario, we assume that the confounders and the treatment variable are binary. We simulate them respectively, for $i = 1, \dots, n$, from $X_{1i} \sim \text{Be}(0.4)$, $X_{2i} \sim \text{Be}(0.6)$, $X_{3i} \sim \text{Be}(0.3)$, $X_{4i} \sim \text{Be}(0.5)$, $X_{5i} \sim \text{Be}(0.2)$ and $T_i \sim \text{Be}(\text{expit}(0.4X_{1i} + 0.6X_{2i}))$, for all scenarios except for Scenario 5 where the treatment is defined as function of all the five confounders, such that $T_i \sim \text{Be}(\text{expit}(0.4X_{1i}+0.6X_{2i}-0.3X_{3i}+0.2X_{4i}X_{5i}))$, where $\text{Be}(\theta)$ represents a Bernoulli random variable with success probability θ.

Each scenario assumes a different configuration of the groups, defining different characterization of the heterogeneity in the causal effects. Each group is obtained by introducing, for each unit i, categorical variables S_i with k categories, that allocate the unit in one of the k groups according to covariates values. Note that we use the words *group* and *cluster* consistently with the previous section. Conditionally on $S_i = s$ we simulate both potential outcomes as $Y_i(0)|S_i = s \sim \mathcal{N}(\eta_s^{(0)}, \sigma_s^{(0)})$ under control and $Y(1)|S_i = s \sim \mathcal{N}(\eta_s^{(1)}, \sigma_s^{(1)})$ under treatment.

In each setting, the sample size is fixed to $n = 500$. For each scenario, we simulate 100 samples. We set $k = 3$ for Scenarios 1-2-3-6, $k = 4$ for Scenario 4, $k = 5$ for Scenario 5, and $k = 1$ for Scenario 7—i.e. to simulate the degenerative case of heterogeneity: homogeneity.

Scenario 1: We investigate a situation in which the expected value of the outcome decreases with the treatment, but the intensity of this decrease varies across different groups. We set $\eta^{(0)} = (2, 4, 6)$ and $\eta^{(1)} = (0, 3, 6)$, and assume that the variance within each group is constant, with $\sigma_s^{(0)} = \sigma_s^{(1)} = 0.3$ for $s = 1, 2, 3$. This results in group average treatment effects of $(-2, -1, 0)$ respectively in the three groups. The units are allocated in the three groups according to the covariates values: $s = 1$ when $X_{1i} = X_{2i} = 0$, $s = 2$ when $X_{1i} = 1$, and $s = 3$ when $X_{1i} = 0$ and $X_{2i} = 1$.

Scenario 2: We examine a typical study setting where the population has different outcomes under the control group but similar outcomes under treatment. To achieve this, we set $\eta^{(0)} = (0, 2.2, 4.4)^T$, $\eta_s^{(1)} = 0$ for each $s \in 1, 2, 3$, and $\sigma_s^{(0)} = \sigma_s^{(1)} = 0.2$. Consequently the group average treatment effects is $\tau(s) = (0, -2.2, -4.4)$ respectively for each group. The group allocation is the same as Scenario 1.

Scenario 3: We focus on a case where the groups are less separated, with the location

parameters $\eta^{(0)}$ and $\eta^{(1)}$ being closer to each other and different variances between groups. Specifically, we set $\eta^{(0)} = (1, 2, 3)$, $\eta^{(1)} = (0, 1.5, 3)$, $\sigma^{(0)} = (0.2, 0.25, 0.25)$, and $\sigma^{(1)} = (0.25, 0.3, 0.2)$. This results in group average treatment effects of $(-1, -0.5, 0)$. The group allocation is the same as Scenario 1.

Scenario 4: We consider a situation in which there are four groups, combining features from the previous scenarios. Specifically, we consider the values within $\eta^{(0)}$ and $\eta^{(1)}$ to be close to each other, as in Scenario 3, and assume a different behavior under treatment and control in terms of heterogeneity as in Scenario 2. Specifically, the outcomes under treatment show four different clusters, while the outcomes under control underline three different clusters. This is achieved by setting $\eta^{(0)} = (1, 2, 3, 3)$, $\eta^{(1)} = (0, 1.5, 3, 4.5)$, and $\sigma_s^{(0)} = \sigma_s^{(1)} = 0.2$ for $s = 1, 2, 3, 4$. The group average treatment effect results from $(-1, -0.5, 0, 1.5)$ respectively for four groups. The units are allocated in the four groups according with the covariates values: $s = 1$ when $X_{1i} = 0$ and $X_{2i} = 1$, $s = 2$ when $X_{1i} = X_{2i} = 0$, $s = 3$ when $X_{1i} = X_{2i} = 1$, and $s = 4$ when $X_{1i} = 1$ and $X_{2i} = 0$.

Scenario 5: We consider all the five confounders X as well as characterization of the groups. Moreover, in this scenario we also increase the number of groups up to 5, describing a more complex and heterogeneous setting. The cluster-spesific parameters are set to $\eta^{(0)} = (2, 2, 3, 4.5, 6.5)$, $\eta^{(1)} = (0, 1, 2.5, 5, 7.5)$, $\sigma_s^{(0)} = \sigma_s^{(1)} = 0.2$ for $s = 1, 2, 3, 4, 5$. We obtain group average treatment effects equal to $(-2, -1, -0.5, 0.5, 1)$. The units are allocated in the five groups according to the values of the five covariates: $s = 1$ when $X_{1i} = X_{2i} = 1$, $s = 2$ when $X_{1i} = 0$ and $X_{3i} = 1$, $s = 3$ when $X_{1i} = X_{3i} = 0$ and $X_{4i} = 1$, $s = 4$ when $X_{1i} = X_{3i} = X_{4i} = 0$, and $s = 5$ otherwise.

Scenario 6: This is similar to Scenario 1 and Scenario 3, but with groups that are even closer and with bigger variance, such that the marginal distributions for the treated and control outcomes, respectively, are not multimodal. In particular, we have three groups with $\eta^{(0)} = (1.5, 2, 2.5)$, $\eta^{(1)} = (1, 1.75, 2.5)$, and $\sigma_s^{(0)} = \sigma_s^{(1)} = 0.3$ for $s = 1, 2, 3$. The group allocation is the same as Scenario 1.

Scenario 7: We study the degenerative case of heterogeneity, such that we have only one group, with $\eta^{(0)} = 2$, $\eta^{(0)} = 3$, and $\sigma_s^{(0)} = \sigma_s^{(1)} = 0.5$. The average treatment effect is equal to 1.

We choose the same hyperparameters for each setting such that the prior is non-informative and in common for all the settings. For the regression parameters in (2.7)

and for the parameters $\eta_l^{(t)}$ and $\sigma_l^{(t)}$ in (2.8) we use the following conjugate priors

$$\beta_{ql}^{(t)} \sim \mathcal{N}(0, 20),\ \eta_l^{(t)} \sim \mathcal{N}(0, 10),\ \text{and}\ \sigma_l^{(t)} \sim \text{InvGamma}(5, 1).$$

for $t \in \{0, 1\}$, $l \in \{1, \ldots, 20\}$—see Section 2.2.1 for the choice of finite truncation of the dependent Probit Stick-Breaking process—, and q according with the covariates considered in different setting.

The performance of the proposed approach is compared to those obtained with the Bayesian additive regression trees model by Hill (2011)—using the `R` package `bartCause`—and with the Bayesian causal forest approach by Hahn *et al.* (2020)—using the package `bcf` available in GitHub. We chose the Bayesian additive regression tree and Bayesian causal forest as a benchmark as these models have shown particular flexibility and an excellent performance—with the need of no or little hyper-parameter tuning—in both prediction tasks (Linero and Yang, 2018; Linero, 2018; Hernández *et al.*, 2018) and in causal inference applications (Hill, 2011; Hahn *et al.*, 2020; Logan *et al.*, 2019; Nethery *et al.*, 2019; Bargagli-Stoffi *et al.*, 2022). These methods do not have a direct characterization of the heterogeneity of the causal effect, but the groups can be obtained with second-step, where the ITE are grouped by classification and regression trees, introduced by Breiman *et al.* (1984). In particular, the benchmark for group identification is Bayesian causal forest and classification and regression tree analysis combo, where we use the `R` package `rpart` for the classification and regression tree analysis.

First, we analyze the result for the SATE. As illustrated in Figure 2.1- 2.2, the results obtained from the three models are quite similar in terms of both bias and mean square error. In particular, the bias is close to zero, as reported in Figure 2.1, where the median of the boxplots is close to the red horizontal line that indicates a bias of zero; and the variability among the censorious is small and correlated to the variability of the simulated data—e.g. the scenario 7 has boxplot with longer tails than the other scenarios, due to a bigger simulated values for the parameters $\sigma^{(0)}$ and $\sigma^{(1)}$. Additionally, the mean square errors, in Figure 2.2, reflect the simulated variability in each of the seven scenarios. The MSE results for the three models are comparable, with smaller median values for CDBMM, even though, there are some outliers for scenario 4 for this model.

The proposed method not only produces accurate average treatment effect estimates but also excels in estimating the GATEs, that directly depend on the identified partition of the groups. To evaluate the estimated partition we use the adjusted Rand index (ARI)—using the `R` package `mcclust`. ARI is a cluster comparison measure, that informs

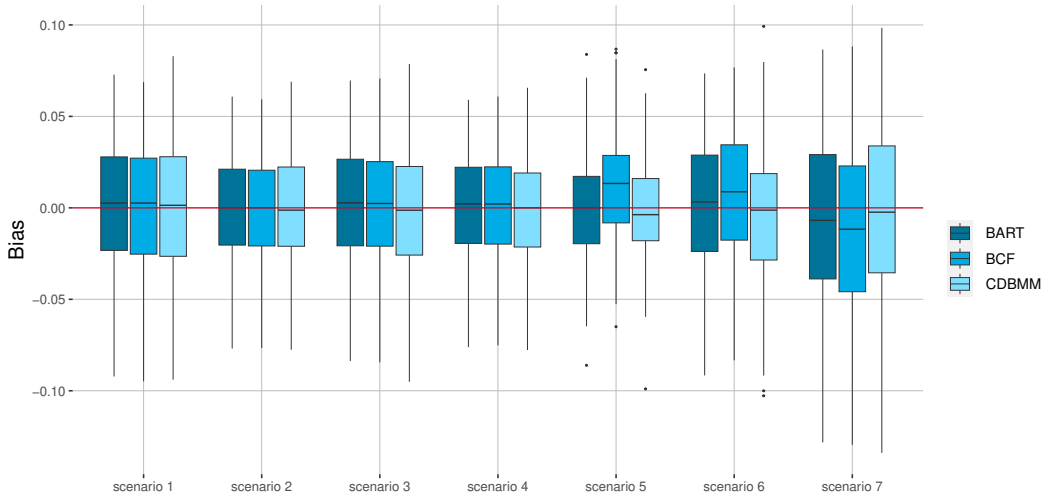

FIGURE 2.1: Comparison of estimation of the average of individual treatment effects between Bayesian additive regression tree (BART), Bayesian causal forest (BCF), and CDBMM: bias.

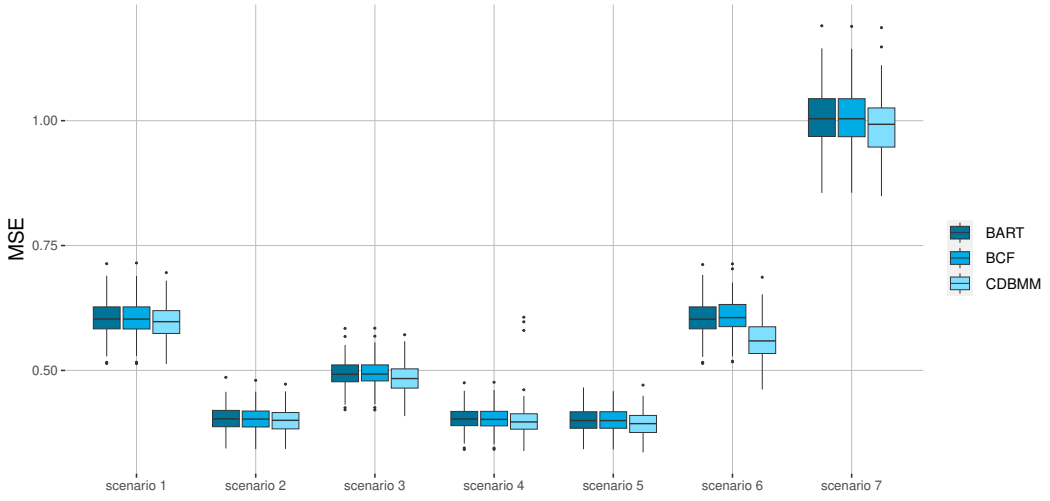

FIGURE 2.2: Comparison of estimation of the average of individual treatment effects between Bayesian additive regression tree (BART), Bayesian causal forest (BCF), and CDBMM: mean square error (MSE).

Table 2.1: Mean and empirical standard deviation of the adjusted Rand index computed for the seven settings with CDBMM and Bayesian causal forest and classification and regression tree analysis (BCF+CART) combo.

	CDBMM		BCF + CART	
	mean	sd	mean	sd
Scenario 1	0.9995	0.0016	0.8203	0.0185
Scenario 2	0.9910	0.0390	0.8203	0.0185
Scenario 3	0.9981	0.0046	0.8158	0.0416
Scenario 4	0.9926	0.0320	1.0000	0.0000
Scenario 5	0.9905	0.0199	0.7935	0.0283
Scenario 6	0.9757	0.0425	0.7351	0.1341
Scenario 7	1.0000	0.0000	1.0000	0.0000

of the goodness of the estimated partition, compared with the simulated partition. It takes values in $[0, 1]$, with 0 indicating that the two partitions do not agree on any pair of units and 1 indicating that the partitions perfectly match. We utilize as a benchmark the Bayesian causal forest and classification and regression tree analysis combo, and the compared results are reported in Table 2.1. The proposed method has superior performance, in mean and variability, broadly concerning the Bayesian causal forest and classification and regression tree analysis combo in Scenarios 1–6 where heterogeneity is present. In the case of homogeneity (Scenario 7), the methods are comparable as both methods correctly find a single group.

Concluding the simulation study, we show the estimated GATEs for each group in the seven simulated scenarios in Figure 2.3. The boxplots underline that the medians of each GATE are in close agreement with the true simulated values, reported as the light-blue dot-dashed lines, confirming the model's capability to identify and estimate the GATEs with high precision.

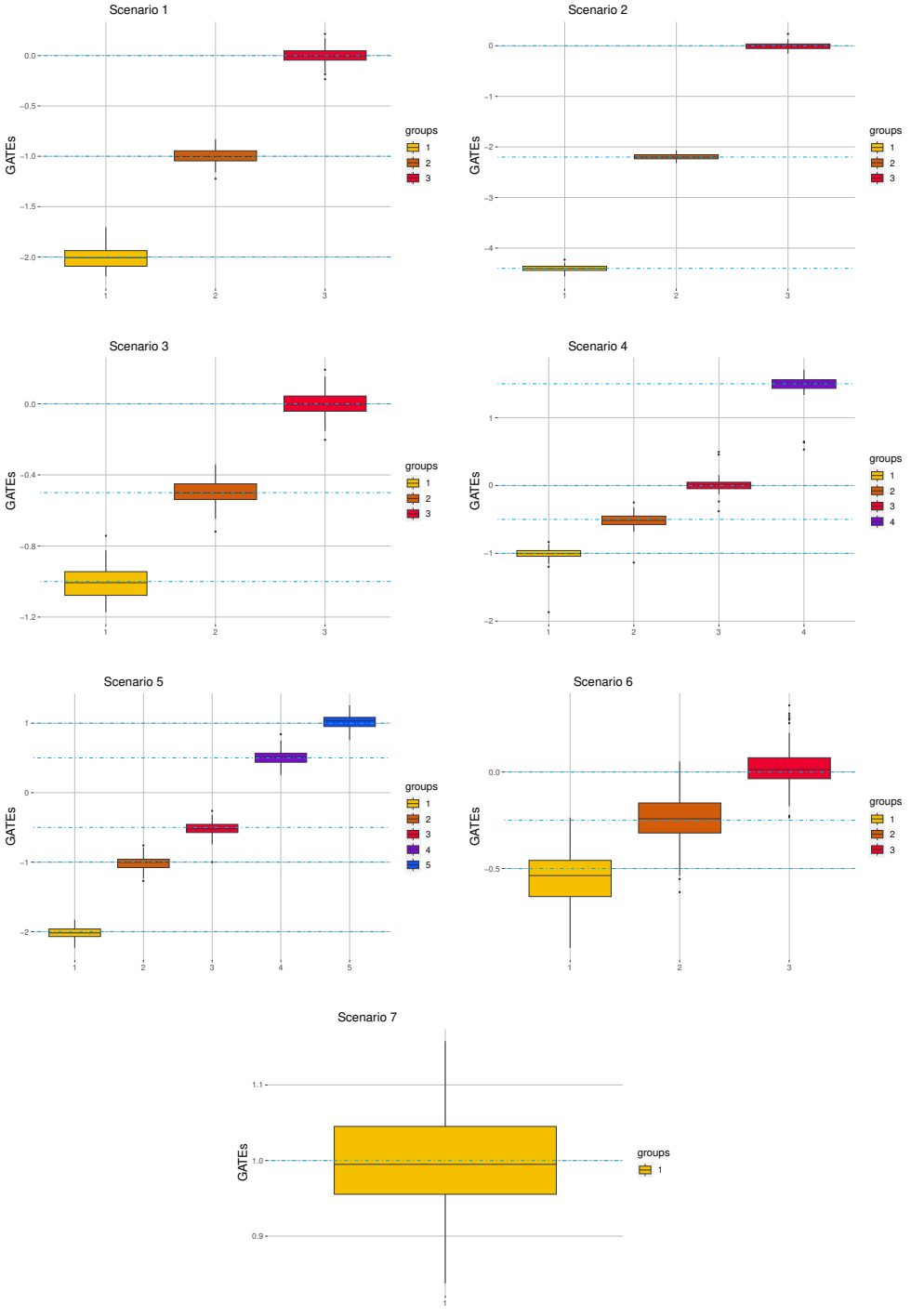

FIGURE 2.3: GATEs for the groups in the seven simulated scenarios. The light-blue dot-dashed lines show the true values.

Chapter 3

Confounders-Aware Shared-atoms Bayesian Mixture Model for Principal Stratification

Bayesian nonparametric mixture models can handle the challenges that emerge in the principal stratification framework (Frangakis and Rubin, 2002). First and foremost, our primary objective is to elucidate the causal estimands, taking into account the continuous nature of the post-treatment variable and its impact on the outcome. Secondly, we aim to harness the adaptability offered by the dependent Dirichlet process prior to defining a model that captures the distribution of the potential post-treatment variables and the complex causal relations.

In pursuit of these aims, in this chapter, firstly we define novel principal strata estimands and secondly, we introduce a new method that allows us to learn these estimands a Confounders-Aware SHared-atoms BAyesian mixture model denoted by the acronym CASBAH, with the intentional alteration of the letter "H" for improved terminological conciseness.

3.1 Principal Strata Estimands

We have elucidated the principal stratification framework in Section 1.2.2, including the main estimand currently used in the literature. The continuous nature of the post-treatment variable makes more complex the definition of the standard expected dissociative effect (EDE) and expected associative effect (EAE) for the discrete post-treatment

This chapter is based on Zorzetto *et al.* (2023b).

variable. As reported in Section 1.2.2, Zigler *et al.* (2018) suggests a criteria for the stratification based on a threshold ξ.

We believe that a priori choice of the threshold ξ can be too sensible for arbitrary choice. However, we believe that the strata are expressions of heterogeneity in the causal effect of the treatment on the post-treatment variable. Specifically, we assume that the two potential post-treatment variables $\{P_i(1), P_i(0)\}$ have distributions that are a mixture of parametric kernels, and consequently also $P_i(1) - P_i(0)$—i.e. the causal effect of the treatment on the post-treatment variable—has a mixture distribution, such that each of its kernels has a finite first moment.

In other words, we assume that the units can be clustered into strata, where each stratum has a finite average treatment effect on the post-treatment variable. Specifically, if the a unit has

- $\mathbb{E}[P_i(1) - P_i(0)] = 0$, it belongs to the *dissociative stratum*;
- $0 < \mathbb{E}[P_i(1) - P_i(0)] < \infty$, it belongs to the *associative positive stratum*;
- $-\infty < \mathbb{E}[P_i(1) - P_i(0)] < 0$, it belongs to the *associative negative stratum*.

In regard to the environmental epidemiology application in Section 4.2, we can define dissociative stratum as those counties where the air pollution regulation does not significant modify the level of $\text{PM}_{2.5}$. Oppositely, a county, where the reduction of the level of $\text{PM}_{2.5}$ is significant greater when the air pollution regulation is apply that when it is not, is allocated in the associative negative stratum.

Therefore, we propose novel estimands conditionally on latent variables V_i that define the allocation of each unit $i \in \{1, \dots, n\}$, in one of the three strata previously described. Specifically $V_i = 0$ in the dissociative stratum, while it is equal to ± 1 for the associative positive and negative strata, respectively.

Therefore, conditionally on V_i, we can define the causal estimands for EDE and EAEs as

$$\begin{aligned} EDE &= \tau_0 = \mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = 0], \\ EAE_+ &= \tau_{+1} = \mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = +1], \\ EAE_- &= \tau_{-1} = \mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = -1]. \end{aligned} \tag{3.1}$$

With respect to the real-data application, the principal causal estimands capture the heterogeneity in the causal effect of the air pollution regulation on the mortality rate conditional to the different behaviour of the level of $\text{PM}_{2.5}$. I.e., we want to estimate

the difference on the mortality rate in case of application and no application of the regulation, given the principal strata induced by the $PM_{2.5}$.

In order to identify the causal estimands introduced above, the following assumption has to be made, in addition to SUTVA invoked for both the outcome variable and the post-treatment variable, as explained in Section 1.2.2:

Strongly Ignorable Treatment Assignment. Given the observed covariates x_i,

$$\{\big(Y_i(1), Y_i(0)\big), \big(P_i(0), P_i(1)\big)\} \perp\!\!\!\perp T_i \mid X_i,$$

$$0 < Pr\left(T_i = 1 \mid X_i = x\right) < 1 \ \forall\, x \in \mathcal{X},$$

where $\mathcal{X}$ is the features' space. Therefore as reported also in Section 2.1, (i) we have a random treatment assumption in each group conditional on some covariates values; (ii) all units have a positive chance of receiving the treatment.

Under these assumptions, the causal estimands in (3.1) can be written as follows. For a clearer notation, we use the notation $\mathbb{E}_A\mathbb{E}[B \mid A]$ to indicate $\int_{\mathcal{A}} \mathbb{E}[B \mid A = a] \Pr(A = a)da$ where $\mathcal{A}$ is the support of the variable A.

$$\begin{aligned}
&\mathbb{E}\big[Y_i(1) - Y_i(0) \mid V_i = v\big] \\
&\quad = \mathbb{E}\big[Y_i(1) \mid V_i = v\big] - \mathbb{E}\big[Y_i(0) \mid V_i = v\big] \\
&\quad = \mathbb{E}_{X_i|V_i=v}\mathbb{E}\big[Y_i(1) \mid V_i = v, X_i = x\big] - \mathbb{E}_{X_i|V_i=v}\mathbb{E}\big[Y_i(0) \mid V_i = v, X_i = x\big] \\
&\quad = \mathbb{E}_{X_i|V_i=v}\mathbb{E}\big[Y_i \mid T_i = 1, V_i = v, X_i = x\big] - \mathbb{E}_{X_i|V_i=v}\mathbb{E}\big[Y_i \mid T_i = 0, V_i = v, X_i = x\big],
\end{aligned}$$

for each stratum $v \in \{0, \pm 1\}$, where the first and second equality invoke the properties of expectation and the third equality is due to the no unmeasured confounding and SUTVA assumption. The inner expectation of each integral, for the treatment value $t \in \{0, 1\}$, can be decomposed as

$$\begin{aligned}
&\mathbb{E}\big[Y_i \mid T_i = t, V_i = v, X_i = x\big] \\
&\quad = \mathbb{E}_{P_i(1),P_i(0)|T_i=t,V_i=v,X=x}\mathbb{E}\big[Y_i \mid T_i = t, V_i = v, X_i = x, P_i(1) = p_1, P_i(0) = p_0\big] \\
&\quad = \mathbb{E}_{P_i(1),P_i(0)|T_i=t,V_i=v,X_i=x}\mathbb{E}\big[Y_i \mid T_i = t, X_i = x, P_i(1) = p_1, P_i(0) = p_0\big],
\end{aligned}$$

such that $p_1 = p_0$ when $v = 0$, $p_1 > p_0$ when $v = +1$, and $p_1 < p_0$ when $v = -1$. The inner expectations $\mathbb{E}\big[Y_i \mid T_i = t, X_i = x, P_i(1) = p_1, P_i(0) = p_0\big]$ is straightforward to calculate since we have an outcome model for $Y_i \mid T_i, X_i, P_i(1), P_i(0)$.

Hence the joint probability of the potential post-treatment variables, involved in the previous integral can be re-written as

$$\begin{aligned}
&\Pr\big(P_i(1)=p_1, P_i(0)=p_0 \mid T_i=t, V_i=v, X_i=x\big) \\
&\quad= \frac{\Pr\big(P_i(1)=p_1, P_i(0)=p_0, T_i=t, V_i=v, X_i=x\big)}{\Pr\big(T_i=t, V_i=v, X_i=x\big)} \\
&\quad= \frac{\Pr\big(V_i=v \mid P_i(1)=p_1, P_i(0)=p_0, T_i=t, X_i=x\big)\Pr(P_i(1)=p_1, P_i(0)=p_0 \mid T_i=t, X_i=x)}{\Pr\big(V_i=v \mid T_i=t, X_i=x\big)} \\
&\quad= \frac{\Pr(P_i(1)=p_1, P_i(0)=p_0 \mid T_i=t, X_i=x)\mathbb{I}_{\{V_i=v\}}}{\Pr\big(V_i=v \mid T_i=t, X_i=x\big)}.
\end{aligned}$$

Therefore, the statistical estimands for (3.1) are

$$\begin{aligned}
\tau_v &= \mathbb{E}_{X_i|V_i=v}\mathbb{E}\big[Y_i \mid T_i=1, V_i=v, X_i=x\big] - \mathbb{E}_{X_i|V_i=v}\mathbb{E}\big[Y_i \mid T_i=0, V_i=v, X_i=x\big] \\
&= \int_x\int_{p_0p_1} \mathbb{E}\big[Y_i \mid T_i=1, X_i=x, P_i(1)=p_1, P_i(0)=p_0\big]\frac{\Pr(P_i(1)=p_1, P_i(0)=p_0 \mid T_i=1, X_i=x)}{\Pr\big(V_i=v \mid T_i=1, X_i=x\big)} \\
&\quad\times \mathbb{I}_{\{V_i=v\}}\Pr(X_i=x \mid V_i=v)dp_0p_1dx - \int_x\int_{p_0p_1} \mathbb{E}\big[Y_i \mid T_i=0, X_i=x, P_i(1)=p_1, P_i(0)=p_0\big] \\
&\quad\times \frac{\Pr(P_i(1)=p_1, P_i(0)=p_0 \mid T_i=0, X_i=x)}{\Pr\big(V_i=v \mid T_i=0, X_i=x\big)}\mathbb{I}_{\{V_i=v\}}\Pr(X_i=x \mid V_i=v)dp_0p_1dx,
\end{aligned} \tag{3.2}$$

for $v \in \{0, \pm 1\}$, where the $\Pr(P_i(1)=p_1, P_i(0)=p_0 \mid T_i=t, X_i=x)$ and the weights $\frac{\mathbb{I}_{\{V_i=v\}}\Pr(X_i=x|V_i=v)}{\Pr\left(V_i=v|T_i=t,X_i=x\right)}$, for $t=\{0,1\}$, can be calculate by the post-treatment variable model $P_i(1), P_i(0) \mid T_i, X_i$, defined in the following section.

3.2 Confounders-Aware Shared-atoms Bayesian Mixture Model

Following the Bayesian paradigm, the joint probability distribution of the involved variables is defined as

$$\Pr(Y_i(0), Y_i(1), P_i(0), P_i(1), T_i, X_i) = \int_\Theta \Pr(Y_i(0), Y_i(1), P_i(0), P_i(1), T_i, X_i, \theta)p(\theta)d\theta$$

for $i=1,\ldots,n$, where $p(\theta)$ is the prior distribution for all the involved parameters θ that take values in the parametric space Θ, and the inner probability can be factorized

as

$$\Pr(T_i \mid Y_i(0), Y_i(1), P_i(0), P_i(1), X_i, \theta) \times \Pr(Y_i(0), Y_i(1) \mid P_i(0), P_i(1), X_i, \theta) \\ \times \Pr(P_i(0), P_i(1) \mid X_i, \theta) \times \Pr(X_i \mid \theta). \tag{3.3}$$

Following the assumptions in the previous section, the conditional probability of the treatment variable can be written as $\Pr(T_i \mid Y_i(0), Y_i(1), P_i(0), P_i(1), X_i, \theta) = \Pr(T_i \mid X_i, \theta)$. Moreover, we condition on the empirical distribution of covariates so that $\Pr(X_i \mid \theta) = \Pr(X_i)$. Both the treatment and covariates distributions do not need to be modeled since they are observed (Schwartz *et al.*, 2011).

However the remaining two probabilities in equation (3.3), have to be modeled: (i) the distribution of potential outcomes of the response conditional on the potential outcome of the post-treatment variable and covariates and (ii) the distribution of the potential outcome of the post-treatment variable conditional on the covariates.

We focus our attention on the latter, that is the distribution of the potential outcome of the post-treatment variable conditional on the covariates for which we propose a novel Bayesian nonparametric approach. As our primary focus lies in the post-treatment variable distribution, for the sake of clarity in our discussion, we make the simplifying assumption of employing a more straightforward parametric model for the conditional response distribution following Schwartz *et al.* (2011)'s settings. While the model for the outcome is assumed to be a linear regression model, it can also be easily generalized to a more complex and flexible model.

For the post-treatment variable, we exploit a dependent nonparametric mixture prior, following the approach adopted in the previous chapter of this thesis and more broadly the dependent Dirichlet process (MacEachern, 2000; Barrientos *et al.*, 2012; Quintana *et al.*, 2020). Notably, our model shares some similarities also with the hierarchical Dirichlet process (Teh *et al.*, 2004, 2006; Teh and Jordan, 2009) and the more recent common atoms model of Denti *et al.* (2023).

Specifically, we assume for each $i = 1, \ldots, n$:

$$\begin{aligned} \{P_i(t) \mid x_i, t\} &\sim f^{(t)}(\cdot \mid x_i), \text{ for } t = \{0, 1\}, \\ f^{(t)}(\cdot \mid x_i) &= \int_{\Psi} \mathcal{K}(\cdot; x_i, \psi) dG_{x_i}^{(t)}(\psi), \\ G_{x_i}^{(t)} &\sim \Pi_{x_i}, \end{aligned} \tag{3.4}$$

where $\mathcal{K}(\cdot; x, \psi)$ is a continuous density function, for every $\psi \in \Psi$, and $G_x^{(t)}$ is a random

probability measure depending on the confounders x_i associated to an observation assigned to treatment level t. In particular, the distributions belonging to $G_x^{(0)}$ and $G_x^{(1)}$ are characterized by specific weights assigned to a common set of atoms induced by the common process Π_x.

Following the characterization of the random probability measures we can write:

$$G_{x_i}^{(t)} = \sum_{l \geq 1} \pi_l^{(t)}(x_i)\delta_{\psi_l}, \tag{3.5}$$

where the sequences $\{\pi_l^{(t)}(x_i)\}_{l\geq 1}$ for $t = \{0, 1\}$ represent infinite sequences of random weights, and $\{\psi_l\}_{l\geq 1}$ is an infinite sequence of random kernel's parameters—independent and identically distributed from a base measure H—shared among potential outcome of post-treatment variable under both the treatment level t.

This definition appears similar to the prior distribution introduced in the previous chapter. However, the implication of the difference between (2.5) and (3.5) is remarkable and coherent with the different goals of the two proposed approaches. Basically, while in the CDBMM the atoms $\{\psi_l^{(t)}\}_{l\geq 1}$ for $t \in \{0, 1, \}$ are independent conditionally to the treatment level to give maximum flexibility to the model, in the CASBAH the information between the different treatment levels are shared, such that we have only one infinite sequence of random kernel's parameters $\{\psi_l\}_{l\geq 1}$. This implies that (i) the potential post-treatment variables $\{P(1), P(0)\}$ are not independent conditionally to the covariates X, while in CDBMM the potential outcomes given X are independent, and (ii) we have the desirable feature that there is a not null probability that, for any i, $P_i(0)$ and $P_i(1)$ share the same atom—definition that we exploit to identify the dissociative stratum.

However, the specification of the sequences of weights follows the same stick-breaking representation (Sethuraman, 1994) reported in (2.6). In particular, we choose again the dependent Probit Stick-Breaking process (Rodriguez and Dunson, 2011)—see details in Section 2.2—for its convenient theoretical and computational properties.

The discrete nature of the random probability measure $G_{x_i}^{(t)}$, for $t = \{0, 1\}$, allows us to introduce the latent categorical variables $S_i^{(t)}$, for $t = \{0, 1\}$, describing clusters of units defined by heterogeneous responses to the treatment level t. Assuming $\mathbb{P}\{S_i^{(t)} = l\} = \pi_l^{(t)}(x_i)$, we can write model in (3.4), exploiting conditioning on $S_i^{(t)}$, as

$$\{P_i(t)|x_i, t, \psi, S_i^{(t)} = l\} \sim \mathcal{K}(\cdot \mid x_i, \psi_l), \quad \psi_l \sim H.$$

where ψ represents the infinite sequence $\{\psi_l\}_{l\geq 1}$.

As commonly done, we assume the kernel $\mathcal{K}$ to be a Gaussian, so that model (3.4)–(3.5) specifies to

$$\{P_i(t)|x_i, S_i^{(t)} = l, \eta, \sigma\} \sim \mathcal{N}(\eta_l, \sigma_l^2). \tag{3.6}$$

where η and σ represent the infinite sequences of location parameters $\{\eta_l\}_{l\geq 1}$ and scale parameters $\{\sigma_l\}_{l\geq 1}$, respectively, such that $\psi_l = (\eta_l, \sigma_l)$.

Prior elicitation is completed by assuming for the parameters η_l and σ_l in (3.6)

$$\eta_l \overset{iid}{\sim} \mathcal{N}(\mu_\eta, \sigma_\eta^2), \text{ and } \sigma_l \overset{iid}{\sim} \text{InvGamma}(\gamma_1, \gamma_2).$$

where $\text{InvGamma}(\gamma_1, \gamma_2)$ represents the inverse-gamma distribution.

For the regression parameters in the dependent Probit Stick-Breaking process, we assume the following multivariate Gaussian prior:

$$\beta^{(t)} \sim \mathcal{N}_{(p+1)(L-1)}(\xi, \Omega) \tag{3.7}$$

for $t = 0, 1$ and $l \geq 1$. According to Fasano *et al.* (2022), the Gaussian prior leads to straightforward posterior computation as discussed in the next section.

Sampling from the posterior joint distribution is straightforward via Gibbs sampling. In particular, the algorithm—described in detail in Section 3.2.2—takes inspiration from those proposed by Teh *et al.* (2004, 2006); Teh and Jordan (2009), for the hierarchical structure, and by Fasano *et al.* (2022) for the probit regression in the weights.

However, in observation studies, it is fundamental the control the confounding bias. Therefore, in the real-data application, the study design is required before the fitting of the CASBAH.

3.2.1 Conjugate Prior for Multinomial Probit Regression

Assuming a multivariate Gaussian distribution as prior of the probit regression parameters—Eq. (3.7)—is done not only because it is the most natural choice but also because it enjoys interesting computational properties if paired with a probit likelihood as first discussed by Durante (2019). The Gaussian, indeed, is a special case of the SUN distribution (Arellano-Valle and Azzalini, 2006) and consequently, adapting the solutions presented in Fasano *et al.* (2022), it allows us (i) to obtain an efficient step in the Gibbs sampler and (ii) to avoid data augmentation that has usually same drawbacks.

Starting from the general distribution of the multivariate Gaussian distribution for $\beta^{(t)}$ in Eq. (3.7), i.e. the SUN density distribution, the probability function is defined, for $q = (p+1)(L-1)$, as

$$\mathbb{P}(\beta^{(t)}) = \phi_q(\beta^{(t)} - \xi; \Omega) \frac{\Phi_h(\gamma + \Delta^T \bar{\Omega}^{-1} \omega^{-1}(\beta^{(t)} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_h(\gamma; \Gamma)}, \tag{3.8}$$

where $\phi_q(\beta^{(t)} - \xi; \Omega)$ is a probability density function of a q-variate Gaussian distribution with ξ vector of location parameters and Ω the covariance matrix, such that $\Omega = \omega \bar{\Omega} \omega$ where $\bar{\Omega}$ is the correlation matrix and $\omega = (\Omega \bigodot \mathbf{1}_q)^{1/2}$ where $\bigodot$ is the element-wise Hadamard product. The second part of the formula introduces a skewness mechanism, driven by the cumulative distribution function, computed at $\gamma + \Delta^T \bar{\Omega}^{-1} \omega^{-1}(\beta^{(t)} - \xi) \in \mathbb{R}^h$ of an h-variate Gaussian with mean vector 0 and $h \times h$ covariance matrix $\Gamma - \Delta^T \bar{\Omega}^{-1} \Delta$. The quantity $\Phi_h(\gamma; \Gamma)$ is the normalizing constant, which coincides with the cumulative distribution function, evaluated at $\gamma \in \mathbb{R}^h$, of an h-variate Gaussian with mean vector 0 and $h \times h$ covariance matrix Γ.

The amount of skewness in the prior is mainly controlled by the $q \times h$ matrix Δ, and when all the entries in Δ are 0, the prior for β_t coincides with the density of a q-variate Gaussian distribution with ξ vector of location parameters and Ω the covariance matrix (Fasano *et al.*, 2022).

Arellano-Valle and Azzalini (2006) show that if $\beta^{(t)} \sim \text{SUN}_{q,h}(\xi, \Omega, \Delta, \gamma, \Gamma)$ then

$$\begin{aligned}
&\beta^{(t)} \overset{d}{=} \xi + \omega(B_0^{(t)} + \Delta \Gamma^{-1} B_1^{(t)}), \\
&B_0^{(t)} \sim \mathcal{N}_q(0, \bar{\Omega} - \Delta \Gamma^{-1} \Delta^T), \\
&B_1^{(t)} \sim TN_h(-\gamma; 0, \Gamma),
\end{aligned}$$

where $TN_h(-\gamma; 0, \Gamma)$ denotes an h-variate Gaussian with zero mean, covariance matrix Γ and truncation below $-\gamma$. A simple mechanism that helps in the simulation of SUN variables.

The multinomial probit distribution for the weights $\pi^{(t)} = \{\pi_l^{(t)}\}_{l=1}^L$—defined in Eq. (**??**)-(**??**)—can be rewrite as

$$\mathbb{P}(S_i^{(t)} = l | \beta^{(t)}, X_i) = \Phi(x_i^T \beta_l^{(t)}) \prod_{k=1}^{l-1} [1 - \Phi(x_i^T \beta_k^{(t)})] = \prod_{k=1}^{l} \Phi\left((2\bar{s}_{ik}^{(t)} - 1) x_i^T \beta_k^{(t)}\right) = \Phi_l(x_i^T \beta^{(t)}; \mathbf{I}_l)$$

for $t = \{0, 1\}$ and $l = 1, \ldots, L-1$, and where $x_i = (1, x_{i1}, \ldots, x_{ip}^T)$ is the vector of the p covariates and intercept for the unit i, $\bar{s}_i^{(t)} = (0_{S_i^{(t)}-1}^T, 1)^T$ if $S_i^{(t)} \leq L-1$ and $\bar{s}_i^{(t)} = 0_{L-1}$ if $S_i^{(t)} = L$, and $\mathbf{I}_l$ refers to the $l \times l$ identity matrix.

Consequently the probability over the observation $i = 1, \ldots, n_t$ for $t = \{0, 1\}$ is

$$\mathbb{P}(S^{(t)}|\beta^{(t)}, X) = \prod_{i=1}^{n^{(t)}} \mathbb{P}(S^{(t)}|\beta^{(t)}, X_i) = \Phi_{\bar{n}_t}(\bar{X}^{(t)}\beta^{(t)}, \mathbf{I}_{\bar{n}^{(t)}}) \tag{3.9}$$

where $\bar{n}^{(t)} = n_1^{(t)} + \cdots + n_n^{(t)}$ with $n_i^{(t)} = \min(s_i^{(t)}, L-1)$, $\bar{X}^{(t)}$ is a $\bar{n}^{(t)} \times [(p+1)(L-1)]$ matrix with row blocks $\bar{X}_{[i]}^{(t)} = X_i^{(t)}$ and $X_i^{(t)} = (diag(2\bar{s}_i^{(t)} - 1) \bigotimes x_i^T, 0_{(n_i^{(t)} \times [(p+1)(L-1-n_i^{(t)})])})$.

Considering with the prior (3.8) and the likelihood (3.9), the posterior distribution for $\beta^{(t)}$ is

$$\mathbb{P}(\beta^{(t)}|S^{(t)}, X) = \phi_q(\beta^{(t)} - \xi; \Omega) \frac{\Phi_{h+\bar{n}_t}(\gamma_{pst} + \Delta_{pst}^T \bar{\Omega}^{-1}\omega^{-1}(\beta^{(t)} - \xi); \Gamma_{pst} - \Delta_{pst}^T \bar{\Omega}^{-1}\Delta_{pst})}{\Phi_{h+\bar{n}_t}(\gamma_{pst}; \Gamma_{pst})}, \tag{3.10}$$

where $\Delta_{pst} = (\Delta, \bar{\Omega}\omega(\bar{X}^{(t)})^T d^{-1})$, $\gamma_{pst} = (\gamma^T, \xi^T(\bar{X}^{(t)})^T d^{-1})$, Γ_{pst} is an $(h + \bar{n}_t) \times (h + \bar{n}_t)$ covariance matrix with blocks $\Gamma_{pst[11]} = \Gamma$, $\Gamma_{pst[22]} = d^{-1}(\bar{X}^{(t)}\Omega(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}})d^{-1}$, and $\Gamma_{pst[12]} = \Gamma_{pst[21]} = d^{-1}\bar{X}^{(t)}\omega\Delta$, where $d = [(\bar{X}^{(t)}\Omega(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}}) \bigodot \mathbf{I}_{\bar{n}^{(t)}}]^{1/2}$.

In the particular case in which the prior for $\beta^{(t)}$ is a multivariate Gaussian distribution—i.e.$h = 0$—, then the posterior is still the SUN distribution in Eq. (3.10) with $\Delta_{pst} = \bar{\Omega}\omega(\bar{X}^{(t)})^T d^{-1}$, $\gamma_{pst} = d^{-1}\bar{X}^{(t)}\xi$, and $\Gamma_{pst} = d^{-1}(\bar{X}^{(t)}\Omega(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}})d^{-1}$.

Moreover a reasonable assumption for $\beta^{(t)}$ prior is the independence among the q elements, such that $\Omega = \omega^2 \cdot \mathbf{I}_q$—i.e. the correlation matrix $\bar{\Omega} = \mathbf{I}_q$. Therefore, following again the Arellano-Valle and Azzalini (2006)'s results, the posterior distribution of $\beta^{(t)}$ can be drawn from

$$\begin{aligned}
\beta^{(t)} &\stackrel{d}{=} \xi + \omega(B_{0,pst}^{(t)} + \Delta_{pst}\Gamma_{pst}^{-1}B_{1,pst}^{(t)}),\\
B_{0,pst}^{(t)} &\sim \mathcal{N}_q(0, \mathbf{I}_q - \Delta_{pst}\Gamma_{pst}^{-1}\Delta_{pst}^T),\\
B_{1,pst}^{(t)} &\sim TN_{h+\bar{n}^{(t)}}(-\gamma_{pst}; 0, \Gamma_{pst}),
\end{aligned}$$

with $\Delta_{pst} = \omega(\bar{X}^{(t)})^T d^{-1}$, $\gamma_{pst} = d^{-1}\bar{X}^{(t)}\xi$, and $\Gamma_{pst} = d^{-1}(\omega^2\bar{X}^{(t)}(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}})d^{-1}$ where $d = [(\omega^2\bar{X}^{(t)}(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}}) \bigodot \mathbf{I}_{\bar{n}^{(t)}}]^{1/2}$.

3.2.2 Posterior Inference

In this section, we describe the Gibbs sampling algorithm for model fitting that allows us to draw from the posterior distribution. Following the steps in the algorithm 2, in each iteration $r = 1, \ldots, R$, we use the observed data (y, p, t, x) to update the parameters and the augmented variables and impute the missing post-treatment variable P^{mis} and missing outcome Y^{mis}.

The Gibbs sampling algorithm is divided into three parts: the estimation of the shared atoms mixture model for the post-treatment variables (divided in the estimation of cluster allocation, cluster-specific parameters, and confounder-dependent weights), the imputation of the missing post-treatment variables, and the estimation of the outcome model.

As already discussed, the outcome model is not our main concern, therefore we assume a linear model. In particular, in the following Gibbs sampler, we consider the outcome model used in the simulation study—i.e., (3.11)—where only the potential post-treatment variables are included. This choice is driven by the purpose of focusing attention on the essential definition of the relation between the post-treatment variable and outcome, which is crucial to impute the missing post-treatment variable P^{mis}. However, the algorithm can be easily modified to include the covariates X in the linear regression or to consider a more complex and flexible model.

Cluster Allocation. The latent variables $S_i^{(t)}$ identifies the cluster allocation for each units $i \in \{1, \dots, n\}$ at the treatment level t. Its posterior distribution is a multinomial distribution where

$$\mathbb{P}\{S_i^{(t)} = l\} \propto \pi_l^{(t)}(x_i)\mathcal{N}(p_i; \eta_l, \sigma_l^2),$$

for $i = 1, \dots, n$ and $l = 1, \dots, L$, with $\omega_l^{(t)}$ defined as:

$$\pi_l^{(t)}(x_i) = \Phi(\alpha_l^{(t)}(x_i)) \prod_{r<l} (1 - \Phi(\alpha_r^{(t)}(x_i))),$$

for $l = 1, \dots, L-1$ and with $\Phi(\alpha_L^{(t)}(x_i)) = 1$.

Cluster Specific Parameters. Thanks to the latent variables $\{S_i^{(0)}, S_i^{(1)}\}$, that cluster the units by the value of their outcome, we know for each cluster $l \in \{1, \dots, L\}$, the allocated units and we can update the values of the parameters from their posterior distributions:

$$\eta_l \sim \mathcal{N}\left(V_l^{-1} \times \left(\frac{\sum_{\{i:S_i^{(0)}=l\}} p_i(0) + \sum_{\{i:S_i^{(1)}=l\}} p_i(1)}{\sigma_l^2} + \frac{\mu_\eta}{\sigma_\eta^2}\right), V_l^{-1}\right);$$

$$\sigma_l^2 \sim \text{InvGamma}\left(\gamma_1 + \frac{n_l}{2}, \gamma_2 + \frac{\sum_{\{i:S_i^{(0)}=l\}} (p_i(0) - \eta_l)^2 + \sum_{\{i:S_i^{(1)}=l\}} (p_i(1) - \eta_l)^2}{2}\right);$$

for $l = 1, \dots, L$ and where $V_l = n_l/\sigma_l^2 + 1/\sigma_\eta^2$ and n_l is the number of units allocated in the l-th cluster.

Confounder-Dependent Weights. The $\{\beta_{ql}^{(t)}\}_{q=0}^{p} = (\beta_{0l}^{(t)}, \beta_{l}^{(t)})$, for $l = 1, \ldots, \max(S_i^{(t)}, L-1)$, are updated for the posterior distribution:

$$\begin{aligned}\beta^{(t)} &\stackrel{d}{=} \xi + \omega(B_{0,pst}^{(t)} + \Delta_{pst}\Gamma_{pst}^{-1}B_{1,pst}^{(t)}),\\ B_{0,pst}^{(t)} &\sim \mathcal{N}_q(0, \mathbf{I}_q - \Delta_{pst}\Gamma_{pst}^{-1}\Delta_{pst}^{T}),\\ B_{1,pst}^{(t)} &\sim TN_{h+\bar{n}^{(t)}}(-\gamma_{pst}; 0, \Gamma_{pst}),\end{aligned}$$

with $\Delta_{pst} = \omega(\bar{X}^{(t)})^T d^{-1}$, $\gamma_{pst} = d^{-1}\bar{X}^{(t)}\xi$, and $\Gamma_{pst} = d^{-1}(\omega^2\bar{X}^{(t)}(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}})d^{-1}$ where $d = [(\omega^2\bar{X}^{(t)}(\bar{X}^{(t)})^T + \mathbf{I}_{\bar{n}^{(t)}}) \bigodot \mathbf{I}_{\bar{n}^{(t)}}]^{1/2}$. More details for $\bar{X}^{(t)}$ and ω definitions in Section 3.2.1.

Imputation Missing Post-Treatment Variables. For each unit $i \in \{1, \ldots, n\}$, we impute the missing post-treatment variable P_i^{mis}. Firstly, drawing the relative cluster-allocation variable $S_i^{(1-t)}$—where t is the observed treatment of the unit i—from a multinomial distribution with

$$\mathbb{P}\{S_i^{(1-t)} = l\} \propto \pi_l^{(1-t)}(x_i),$$

for $l = 1, \ldots, L$. Where $\pi_l^{(1-t)}(x_i) = \Phi(\alpha_l^{(1-t)}(x_i)) \prod_{r<l}(1 - \Phi(\alpha_r^{(1-t)}(x_i)))$, for $l = 1, \ldots, L-1$ and with $\Phi(\alpha_L^{(1-t)}(x_i)) = 1$.

Successively, drawing the missing post-treatment variable P_i^{mis}, conditioned to the allocation to the cluster l and the observed outcome variables $Y_i(t)$. For each i such that the observed treatment level is $T = 1$, $P_i(1-t)$ is drawn from

$$\{P_i^{mis} | S_i^{(1-t)} = l, \eta, \sigma^2, P_i, Y_i\} \sim \mathcal{N}\left(v^{-1}\left(\frac{\eta_l}{\sigma_l^2} + \frac{m_1}{v_1}\right)\right);$$

where

$$v = \frac{1}{\sigma_l^2} + \frac{1}{v_1}, \quad m_1 = \frac{Y_i(1) - \theta_{10} - \theta_{11}P_i(1)}{\theta_{12} + \theta_{13}P_i(1)}, \quad v_1 = \frac{e^{\lambda_0 + \lambda_1 P_i(1)}}{(\theta_{12} + \theta_{13}P_i(1))^2}.$$

While for each i such that the observed treatment level is $T = 0$, $P_i(1-t)$ is drawn from

$$\{P_i^{mis} | S_i^{(1-t)} = l, \eta, \sigma^2, P_i, Y_i\} \sim \mathcal{N}(\eta_l, \sigma_l^2).$$

Outcome Model. The $\theta^{(t)}$ parameters are independent for the treatment level t, therefore the posterior distributions are, respectively for $t = 0, 1$:

$$\begin{aligned}\theta^{(t)} &\sim \mathcal{N}_{q^{(t)}}\left((V^{(t)})^{-1}M^{(t)}, (V^{(t)})^{-1}\right);\\ M^{(t)} &= (\tilde{P}^{(t)})^T\Xi^{(t)}\tilde{P}^{(t)} + (\sigma_\theta^2)^{-1}\mathbf{I}_{q^{(t)}};\\ V^{(t)} &= (\tilde{P}^{(t)})^T\Xi^{(t)}Y(t) + \frac{\mu_\theta}{\sigma_\theta^2}.\end{aligned}$$

For the treatment level $t = 0$: $q^{(0)} = 2$; $\tilde{P}^{(0)}$ is a matrix $n_0 \times n_0$ such that $\tilde{P}^{(0)} = [1_{n_0}, P(0)]$ with 1_{n_0} a vector of 1 and $P(0)$ the vector of observed values of post-treatment variable for the units n_0 assigned at the control group—i.e. $t = 0$; and $\Xi^{(0)}$ is a diagonal matrix $n_0 \times n_0$ with value $\exp(\lambda_0)$ in the diagonal. In similar way, for the treatment level $t = 1$: $q^{(1)} = 4$; $\tilde{P}^{(1)}$ is a matrix $n_1 \times n_1$ such that $\tilde{P}^{(1)} = [1_{n_0}, P(1), P(0), P(1) \cdot P(0)]$ with 1_{n_1} a vector of 1, $P(1)$ the vector of observed values of post-treatment variable for the units n_1 assigned at the treated group—i.e. $t = 0$—and $P(0)$ the vector of imputed values of post-treatment variable; and $\Xi^{(1)}$ is a diagonal matrix $n_1 \times n_1$ with the values $exp(\lambda_0 + \lambda_1 P(1))$ in the diagonal.

The parameters in the variance of the Y-model, λ_0 and λ_1, do not have conjugate priors, therefore a independent Metropolis proposal step is necessary. At each iteration $r \in \{1, \ldots, R\}$, λ_0^* and λ_1^* are drown from the proposal distribution $\mathcal{N}(\mu_{\lambda_0}, \sigma^2_{\lambda_0})$ and $\mathcal{N}(\mu_{\lambda_1}, \sigma^2_{\lambda_1})$ respectively. Then, at iteration r the value of the parameter are updated as following: $\lambda_0^{(r)} = \lambda_0^*$ with probability

$$\frac{\prod_{i\in n}\mathcal{N}(Y_i|\mu_Y^{(t)}, \exp(\lambda_0^* + \mathbb{I}_{(T_i=1)}\lambda_1^{(r-1)}P_i(1))}{\prod_{i\in n}\mathcal{N}(Y_i|\mu_Y^{(t)}, \exp(\lambda_0^{(r-1)} + \mathbb{I}_{(T_i=1)}\lambda_1^{(r-1)}P_i(1))},$$

otherwise $\lambda_0^{(r)} = \lambda_0^{(r-1)}$; and $\lambda_1^{(r)} = \lambda_1^*$ with probability

$$\frac{\prod_{i\in n_1}\mathcal{N}(Y_i|\mu_Y^{(t)}, \exp(\lambda_0^{(r-1)} + \lambda_1^* P_i(1))}{\prod_{i\in n_1}\mathcal{N}(Y_i|\mu_Y^{(t)}, \exp(\lambda_0^{(r-1)} + \lambda_1^{(r-1)}P_i(1))},$$

otherwise $\lambda_1^{(r)} = \lambda_1^{(r-1)}$; where $\mu_Y^{(0)} = \theta_{00} + \theta_{01}P_i(0)$ and $\mu_Y^{(1)} = \theta_{10} + \theta_{11}P_i(1) + \theta_{12}P_i(0) + \theta_{13}P_i(0)P_i(1)$.

While the Gibbs sampler allows us to recover the posterior distribution of the random variable, it is crucial the identify the point estimation of the quantities related with the strata. First at all, the posterior distribution of the latent variable V_i, for $i \in \{1, \ldots, n\}$, that describe the probability of the strata allocation, is estimated following the definition in the following section. While its point estimation is obtain with the mode. Given the

strata allocation, the estimation of the causal effect on the post-treatment variable and the EDE and EAEs is straightforward.

Algorithm 2 Confounders-Aware Shared-atoms Bayesian Mixture Model

Inputs:
- the observed data (y, p, t, x).

Outputs:
- posterior distributions of parameters: η, σ, β, θ, and λ;
- imputed values for P^{mis};
- posterior distribution over the space of partitions of the units.

Procedure:
Initialization of all parameters and latent variables.
For $r \in \{1, \ldots, R\}$ **:**
→ *Estimation of Shared Atoms Mixture Model:*
Compute $\omega^{(t)}(x_i)$ for $i = 1, \ldots, n$ and $t = 0, 1$;
Draw $S_i^{(t)}$ for $i = 1, \ldots, n$ and $t = 0, 1$;
Draw η and σ;
Compute $\alpha^{(t)}(x_i)$ for $i = 1, \ldots, n$ and $t = 0, 1$;
Draw $\beta^{(t)}$ for$t = 0, 1$.
→ *Imputation of Missing Post-Treatment Variables:*
Draw P_i^{mis} for $i = 1, \ldots, n$ and $t = 0, 1$.
→ *Estimation of Outcome Model:*
Draw $\theta^{(t)}$ for $t = 0, 1$;
Draw λ_0 and λ_1.
End

3.2.3 Discovery of Principal Strata

As already underlined in Section 2.2.2, one of the advantages of Bayesian nonparametric mixtures is their ability to cluster the observations, thanks to the latent categorical variables, $\{S_i^{(0)}, S_i^{(1)}\}$ for $i = 1, \ldots, n$, that describe the probability of each unit to be allocated in the components of the mixture.

Consistently with our goal to estimate the EAEs and EDE, which are conditional to the principal strata, we define the *principal strata* as those observations that have a particular combination of latent categorical variables for the post-treatment variable mixture.

Under our fully Bayesian approach the couples $\{S_i^{(0)}, S_i^{(1)}\}$ are random variables for which we can characterize the associated posterior distribution. This is customary in all Bayesian infinite mixture models which are inherently associated with random partition models (Quintana, 2006). Moreover, the proposed CASBAH model is defined

such that the atoms $\{\psi_l\}_{l\geq 1}$ are shared between the two potential outcomes of the post-treatment variable of the same unit. Therefore, the model informs us about the posterior probability of $\{S_i^{(0)}, S_i^{(1)}\}$ to be associated with the same atoms or different atoms, for each unit i.

In particular, we define the variable V_i, for each unit $i \in \{1, \ldots, n\}$, as the latent categorical variable, function of $\{S_i^{(0)}, S_i^{(1)}\}$, that defines the allocation in one of the three strata: dissociative, associative positive, and associative negative.

- The unit i that belongs to the *dissociative stratum* has $V_i = 0$ when $S_{1i} = S_{0i}$, independently on the actual values of the atoms $\{\psi_l\}_{l\geq 1}$. Note that when $S_i^{(1)} = S_i^{(0)}$, the latent variables are allocated in the same cluster and thus the expectation $\mathbb{E}[P_i(1) - P_i(0) \mid V_i = 0]$ is null.
- When $S_i^{(1)} \neq S_i^{(0)}$—i.e. the latent variables indicate different clusters—and the value of the corresponding expectation for the latent variable S_{1i} is greater than the value of the corresponding quantity for the latent variable $S_i^{(0)}$—$\psi_{S_i^{(1)}} > \psi_{S_i^{(0)}}$—the unit i belongs to the *associative positive stratum*.
- Similarly to the previous point, the *associative negative stratum* is composed by the units with $S_i^{(1)} \neq S_i^{(0)}$ and $\psi_{S_i^{(1)}} < \psi_{S_i^{(0)}}$, corresponding to $V_i = -1$.

Under these settings, the posterior distribution on V_i, for each i, reflects the posterior uncertainty in the strata allocation, taking into account the heterogeneity in the post-treatment variables induce by the confounders X.

3.3 Simulation Study

The performances of the proposed CASBAH mixture model are assessed through a simulation study. Our objective is to investigate the model's ability to (i) accurately impute the missing post-treatment and outcome variables—i.e. evaluating the bias of the average of $P_i(1) - P_i(0)$ and for the average of $Y_i(1) - Y_i(0)$ over the units $i = 1 \ldots, n$—, (ii) correctly identify the principal strata. To achieve this, we conduct simulations under five different data-generating models and analyze the results to understand the model's behavior in different scenarios.

Specifically, we assume a linear regression model for the outcome model, defined as following:

$$\begin{bmatrix} Y(0) \\ Y(1) \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} \theta_{00} + \theta_{01} P(0) \\ \theta_{10} + \theta_{11} P(1) + \theta_{12} P(0) + \theta_{13} P(0) P(1) \end{bmatrix}, \begin{bmatrix} e^{\lambda_0} & 0 \\ 0 & e^{\lambda_0 + \lambda_1 P(1)} \end{bmatrix} \right). \quad (3.11)$$

Assuming as prior distribution for the parameters

$$\theta^{(0)} = (\theta_{00}, \theta_{01}) \sim \mathcal{N}_2(\mu_\theta, \sigma_\theta^2 \mathbf{I}_2) \text{ and } \theta^{(1)} = (\theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}) \sim \mathcal{N}_4(\mu_\theta, \sigma_\theta^2 \mathbf{I}_4);$$
$$\lambda_0 \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2) \text{ and } \lambda_1 \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2);$$

with $\mu_\theta = 0$, $\sigma_\theta = 10$, $\mu_\lambda = 0$, and $\sigma_{\lambda_1} = 2$.

We simulate two Bernoulli confounders (X_1, X_2) for the scenarios 1-4 and five Bernoulli confounders $(X_1, X_2, X_3, X_4, X_5)$ for scenario 5, and a binary treatment variable, such that $T_i \sim \text{Be}(\text{expit}(0.4X_{1i} + 0.6X_{2i}))$ for the scenarios $1 - 4$ and $T_i \sim \text{Be}(\text{expit}(0.4X_{1i} + 0.6X_{2i} - 0.3X_{3i} + 0.2X_{4i}X_{5i}))$ for scenario 5, for $i = 1, \ldots, n$.

Each scenario assumes a different conformation of the strata for the continuous post-treatment variable $P_i = (P_i(0), P_i(1)) \in \mathbb{R}^2$. Each stratum is obtained by introducing, for each unit, categorical variables $S_i^{(0)}$ and $S_i^{(1)}$, for control and treatment levels respectively, with the vector of probabilities that depend on the values of the confounders and allocates the unit in different clusters. Conditionally on the cluster allocation $S_i^{(t)} = s$, with s that has the same support for control and treatment levels, we simulate both potential post-treatment variables—under control and under treatment, respectively—as

$$P_i(0)|S_i^{(0)} = s \sim \mathcal{N}(\eta_s, \sigma_s^2), \quad P_i(1)|S_i^{(1)} = s \sim \mathcal{N}(\eta_s, \sigma_s^2).$$

The continuous potential outcomes $(Y_i(0), Y_i(1))$, for $i = 1, \ldots, n$, are simulated following the model (3.11) with common values for $\lambda_0 = -0.5$ and $\lambda_1 = 0.1$, while $\theta^{(0)}$ and $\theta^{(1)}$ are different for each scenario. In each setting, the sample size is fixed to $n = 500$. For each scenario, we simulate 100 samples.

Scenario 1: We investigate a situation in which there are two strata: one with a dissociative effect and one with a positive associative effect. In particular, for $S_i^{(0)} = S_i^{(1)} = 1$ $P_i(0), P_i(1) \sim \mathcal{N}(1, 0.05)$, and for $S_i^{(0)} = 2$ and $S_i^{(1)} = 3$ $P_i(0) \sim \mathcal{N}(2, 0.05)$ and $P_i(1) \sim \mathcal{N}(3, 0.05)$. The regression parameters for Y-model are $\theta^{(0)} = (1, 2)$ and $\theta^{(1)} = (1, 2, -1, 0.5)$.

Scenario 2: We focus on a case where we have a dissociative stratum—for $S_i^{(0)} = S_i^{(1)} = 1$ both $P(0)$ and $P(1)$ are simulated from $\mathcal{N}(2, 0.05)$—, an associative stratum with a positive effect—for $S_i^{(0)} = 2$ and $S_i^{(1)} = 3$ $P_i(0) \sim \mathcal{N}(2, 0.05)$ and $P_i(1) \sim \mathcal{N}(3, 0.05)$—, and dissociative stratum with a negative effect—for $S_i^{(0)} = 2$ and $S_i^{(1)} = 1$ $P_i(0) \sim \mathcal{N}(2, 0.05)$ and $P_i(1) \sim \mathcal{N}(1, 0.05)$. The regression parameters for Y-model are $\theta^{(0)} = (1, 2)$ and $\theta^{(1)} = (1, 1.2, -1, 1)$.

Scenario 3: This scenario corresponds to Scenario 1 with closer atoms for the strata

and different variances. In particular, the dissociative stratum has $(P_i(0)|S_i^{(0)} = 1) = (P_i(1)|S_i^{(1)} = 1) \sim \mathcal{N}(1.5, 0.12)$, and the associative stratum has $(P_i(0)|S_i^{(0)} = 2) \sim \mathcal{N}(2, 0.1)$ and $(P_i(1)|S_i^{(1)} = 3) \sim \mathcal{N}(2.5, 0.08)$. The regression parameters for Y-model are $\theta^{(0)} = (1, 2)$ and $\theta^{(1)} = (1, 1.2, -0.8, 0.5)$.

Scenario 4: This scenario corresponds to Scenario 2 with closer atoms for the strata and different variances. In particular, the dissociative stratum has $(P_i(0)|S_i^{(0)} = 1) = (P_i(1)|S_i^{(1)} = 1) \sim \mathcal{N}(1.5, 0.12)$, the associative positive stratum has $(P_i(0)|S_i^{(0)} = 2) \sim \mathcal{N}(2, 0.1)$ and $(P_i(1)|S_i^{(1)} = 3) \sim \mathcal{N}(2.5, 0.08)$, and the associative negative stratum has $(P_i(0)|S_i^{(0)} = 2) \sim \mathcal{N}(2, 0.1)$ and $(P_i(1)|S_i^{(1)} = 1) \sim \mathcal{N}(1.5, 0.12)$. The regression parameters for Y-model are $\theta^{(0)} = (1, 2)$ and $\theta^{(1)} = (1, 1.2, -0.8, 0.5)$.

Scenario 5: We investigate the scenario with the three strata when the number of confounders increases, in particular the treatment variable and cluster allocation variables that depend on five confounders. The dissociative stratum has $(P_i(0)|S_i^{(0)} = 1) = (P_i(1)|S_i^{(1)} = 1) \sim \mathcal{N}(2, 0.05)$, the associative positive stratum has $(P_i(0)|S_i^{(0)} = 3) \sim \mathcal{N}(3, 0.05)$ and $(P_i(1)|S_i^{(1)} = 4) \sim \mathcal{N}(4, 0.05)$, and the associative negative stratum has $(P_i(0)|S_i^{(0)} = 2) \sim \mathcal{N}(2, 0.05)$ and $(P_i(1)|S_i^{(1)} = 1) \sim \mathcal{N}(1, 0.05)$. The regression parameters for Y-model are $\theta^{(0)} = (1, 2)$ and $\theta^{(1)} = (1, 1.2, -1, 0.5)$.

We choose the same hyperparameters for each setting such that the prior is non-informative and in common for all the settings. For the regression parameters in (**??**) and for the parameters η_l and σ_l in (3.6) we use the following conjugate priors

$$\beta^{(t)} \sim \mathcal{N}_{(p+1)(L-1)}(0, 20 \times \mathbf{1}_{(p+1)(L-1)}),$$
$$\eta_l \sim \mathcal{N}(0, 20), \text{ and } \sigma_l \sim \text{InvGamma}(2, 0.5),$$

for $t \in \{0, 1\}$, $l \in \{1, \ldots, 20\}$, and p according with the covariates considered in different settings, and where $\mathbf{1}_q$ is a diagonal matrix $q \times q$.

The performance of the proposed approach is compared to those obtained with the Schwartz *et al.* (2011)'s model (hereinafter referred to as SLM)

Table 3.1 reports the median and interquartile range (IQR) of the bias for the expected value of the posterior distribution of sample average of $P_i(1) - P_i(0)$ and $Y_i(1) - Y_i(0)$, for $i \in \{1, \ldots, n\}$. The results for the five scenarios for the proposed model CASBAH show a good ability of the proposed model to impute the missing variables and capture the true distribution, indeed the medians of the bias are close to zero and the interquartile range is reasonable according to the simulated variability. The comparison with the Schwartz *et al.* (2011)'s model, SLM, underlines the superiority of

CASBAH, which obtains medians of the bias closer to zero and a more contained IQR for $\mathbb{E}[P_i(1) - P_i(0)]$ in all scenarios and almost of them for $\mathbb{E}[Y_i(1) - Y_i(0)]$.

TABLE 3.1: Median and interquartile range (IQR) of the bias for the expected value of the posterior distribution of sample average of $P_i(1) - P_i(0)$ and $Y_i(1) - Y_i(0)$, for $i \in \{1, \ldots, n\}$. Values reported for CASBAH and SLM.

		Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Post-treatment variable						
CASBAH	median	-0.0164	0.0027	-0.0318	0.0027	0.0011
	IQR	0.0315	0.0254	0.0462	0.0325	0.0183
SLM	median	-0.0637	0.3388	0.1254	0.3280	0.3572
	IQR	0.0989	0.0986	0.0956	0.0963	0.1216
Outcome						
CASBAH	median	0.0024	0.0256	0.0067	0.0419	0.0344
	IQR	0.2235	0.1550	0.4661	0.1357	0.1316
SLM	median	-1.9891	-1.8265	-1.4363	-1.3115	-1.8856
	IQR	0.2546	0.2110	0.1904	0.1836	0.3727

To evaluate the correct identification of the principal strata, we use the adjusted Rand index (ARI)—defined in Chapter 2. The values of ARI for the five scenarios are reported in Table 3.2. For all the scenarios, the index is close to 1 and confirms that the proposed model CASBAH can identify correctly the principal strata, and combined with the good missing data imputation, allow us to estimate the expected associative and dissociative effects. The SLM model does not identify the principal strata according to our definition or simultaneously to the model estimation, in opposition to CASBAH.

TABLE 3.2: Adjusted rand index for the five simulated scenarios computed on the point estimated partitions obtained with the proposed model CASBAH. mean and empirical standard deviation (sd) are reported.

	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
mean	0.9850	0.9906	0.9706	0.9717	0.9154
sd	0.0997	0.0597	0.1021	0.0892	0.1577

The Figure 3.1 reports the results for the five simulated scenarios obtained with CASBAH. In the left, the boxplots show the distribution over the simulated samples of the expected values of the difference of the post-treatment variables under treatment and under control in each stratum. The graphics confirm the ability of our proposed model to (i) identify correctly the number of strata—two strata in the simulated scenario 1 and 3, and three in the others—and (ii) capture the definition of the associative/dissociative strata without an *a priori* criteria—the dissociative stratum is always around zero for $\mathbb{E}[P(1) - P(0)]$, while the dissociative stratum do not include the zero. In the boxplots in the right, there are the distribution of the principal causal effect: the statistical

estimate of EAE_-, EDE, and EAE_+. It is clear that the different strata identify different treatment effects on the outcome, allowing us to characterize the heterogeneity in the causal effects. Few outliers are observed, however they are founded in particular in the Scenario 5 that describes a more complex relation among variables and strata.

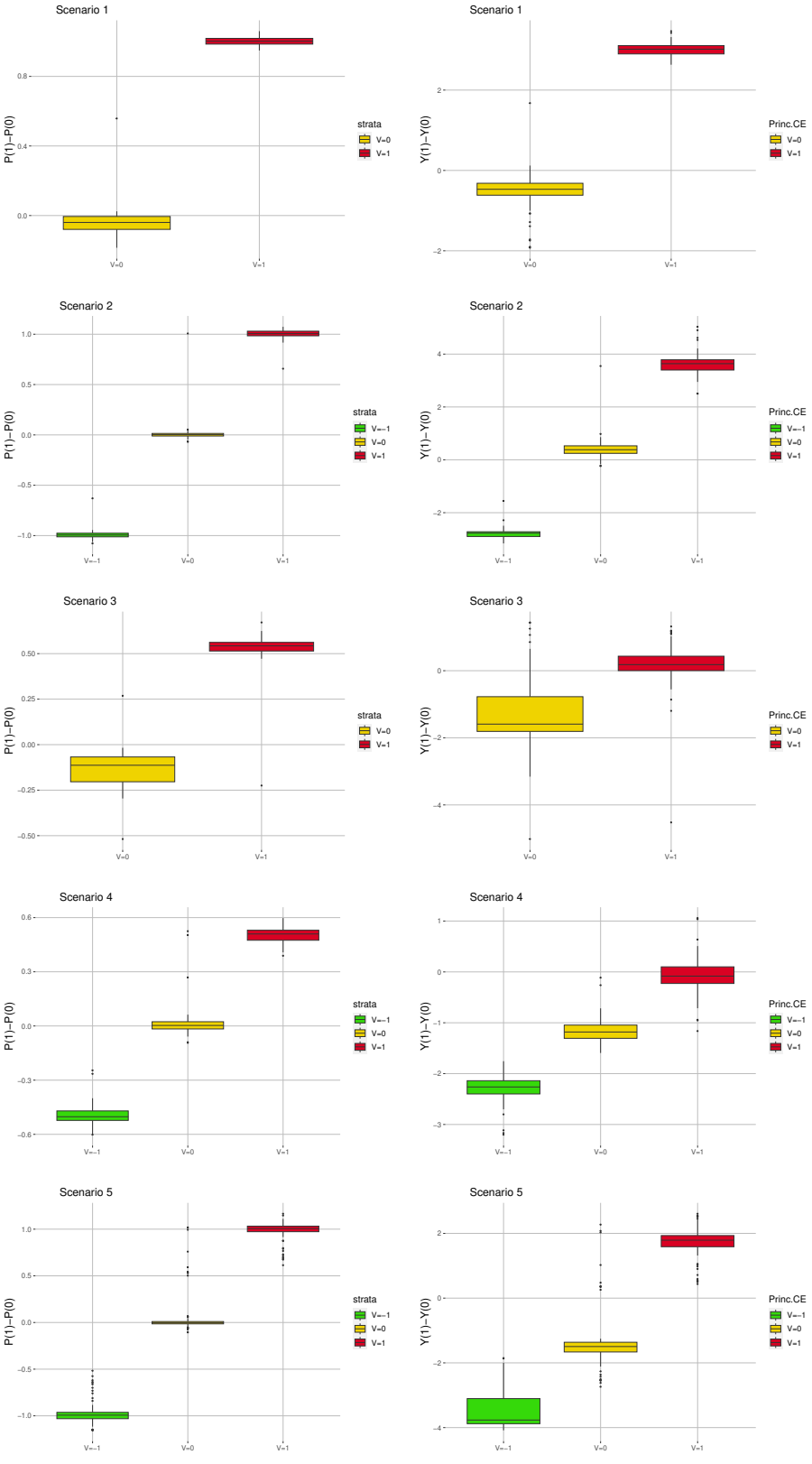

FIGURE 3.1: Representations of the five simulated scenarios. (Left) The expected value of the difference of the post-treatment variables under treatment and under control given the strata allocation. (Right) The expected associative/dissociative causal effects. In green the associative negative stratum—i.e. corresponding to the latent variable $V = -1$—, in yellow the dissociative stratum—i.e., $V = 0$—, and in red the associative positive stratum—i.e., $V = +1$.

Chapter 4

Socio-Economic Disparities and Regulations in Air Pollution Epidemiology

In this last chapter, we address the questions that lead us to formulate the causal Bayesian nonparametric models introduced in the previous chapters: comprehend and capture the complexity of the causal link among the real-data variables. Specifically, we study the causal link between (i) long-term $PM_{2.5}$ exposure and the mortality rate, and (ii) air pollution regulations and variation of mortality rate, conveyed through variation of $PM_{2.5}$. With the main focus on the United States, we take into account the heterogeneity in the demographic and socio-economic characteristics.

This chapter is based on the data applications in Zorzetto *et al.* (2023a) and Zorzetto *et al.* (2023b), respectively included in Section 4.1 and Section 4.2.

4.1 Socio-Economic Disparities in $PM_{2.5}$ Exposure

With a focus on uncovering vulnerability/resilience in the causal effects in the context of an environmental study, we apply the proposed model in Chapter 2—the confounder-dependent Bayesian mixture model—to discover the heterogeneity in the health effects of exposure to higher levels of air pollution in Texas for the elderly population. Texas is a crucial case study for understanding air pollution vulnerability because of its unique demographic makeup and exposure disparities. First, recent literature has shown that, in Texas, black and low-income groups are exposed to higher levels of air pollution, whereas college graduates and high-income groups are exposed to lower levels (Li *et al.*,

2019). Second, Texas has a high proportion of Hispanic residents, which makes it a valuable case study for examining the health impacts of air pollution on this demographic group. According to the US Census Bureau, Hispanic residents make up over 39% of the Texas population (U.S. Census Bureau, 2020). Studies have shown that low-income and Hispanic communities are more likely to be exposed to higher levels of $PM_{2.5}$ compared to wealthier and non-Hispanic communities (Jbaily *et al.*, 2022).

Using the information on Texan Medicare enrollees (i.e., individuals older than 65), we investigate the heterogeneous causal link between long-term $PM_{2.5}$ exposure and mortality. Our analysis depicts how our method can discover mutually exclusive groups, estimate the heterogeneity in the effects of long-term exposure to $PM_{2.5}$ on the mortality rate, and identify the social-economical characteristics that distinguish the different groups.

4.1.1 Data Description

We conduct our analysis at the ZIP code level (1,929 units), where we have data on the following variables: the average $PM_{2.5}$ levels during the years 2010 and 2011; the mortality rate in the 5 follow-up years; census variables such as the percentage of residents for different races/ethnicities (in particular, categorized as Hispanics, blacks, whites, and other races); the age of each Medicare enrollee (≥ 65 years of age) and their sex (female/male); the average household income; the average home value; the proportion of residents in poverty; the proportion of residents with a high school diploma; the population density; the proportion of residents that own their house; the average body mass index; the smoking rate; the percentage of people who are eligible for Medicare (this variable is a proxy of low social-economic status and is reported as S.E.S.). Moreover, we also have access to meteorological variables: the averages of maximum daily temperatures and the relative average of humidity during summer (June to September) and winter (December to February).

The distribution of the population in Texas during 2010, represented in the map (a) in Figure 4.1, is clearly concentrated around the main cities, such as Dallas, San Antonio, Austin, and Houston, while expansive areas are quite empty, due to the desert ecosystem. Consequently, we consider only the ZIP codes with a population density different from zero and more than 10 Medicare enrollees, such that we have enough records in the Medicare dataset for those ZIP codes. For these ZIP codes, the observed values of $PM_{2.5}$ and mortality rates in Texas are illustrated in Figure 4.1—(b) and (c), respectively. Not surprisingly, the highest values on record for $PM_{2.5}$ are primarily

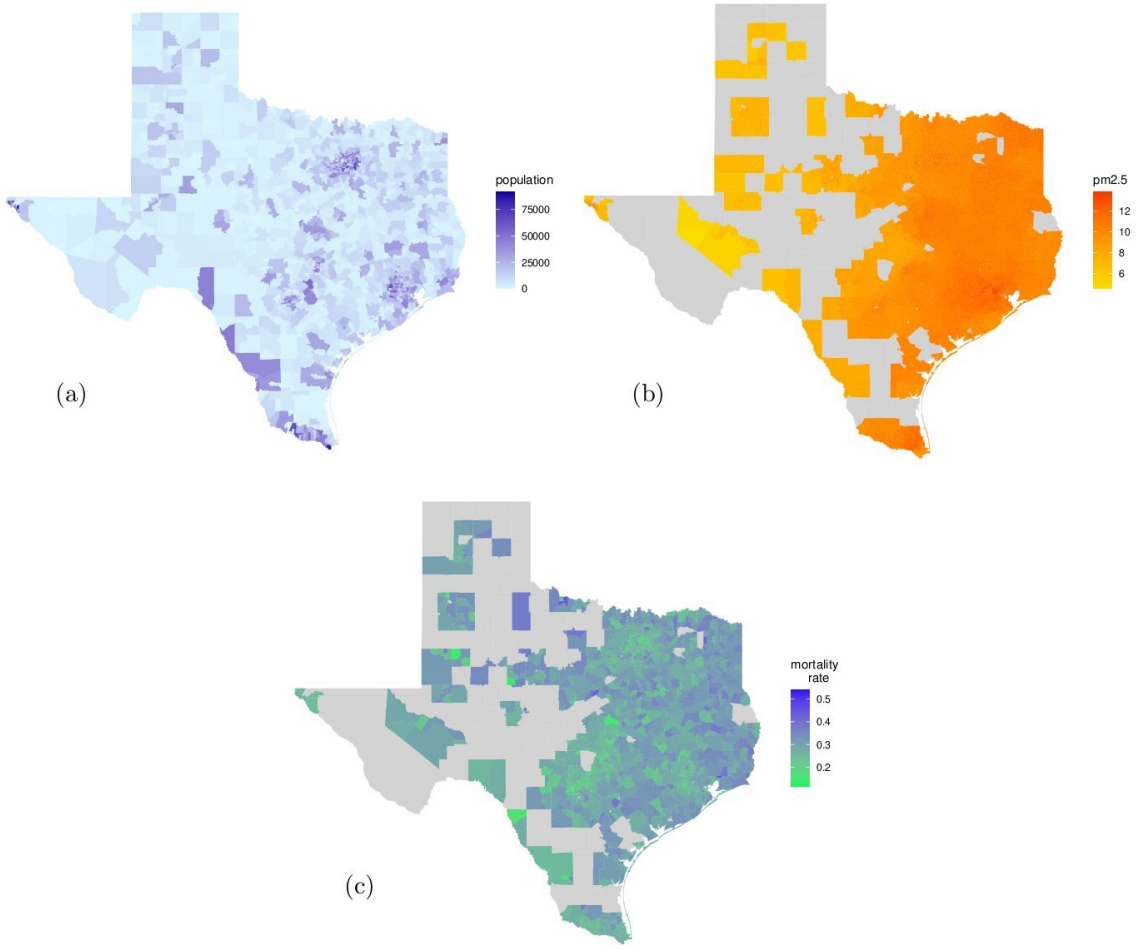

FIGURE 4.1: (a) Population density during 2010 in Texas (USA) (b) Average of long-term $PM_{2.5}$ exposure in 2010-2011 in Texas. The data are aggregated by ZIP codes. (c) The mortality rate at 5 follow-up years for each Texan ZIP code. The gray areas indicate the ZIP codes with population density different from zero and more than 10 Medicare enrollees.

concentrated in the urban areas, while the mortality rate doesn't show any particular pattern.

4.1.2 Study Design

We define the treatment variable as $T = 1$ if the average $PM_{2.5}$ in 2010 and 2011 is above the threshold of $10\mu g/m^3$—corresponding to 701 out of 1,929 considered zip codes—and $T = 0$ otherwise—corresponding to 1,228 zip codes. The choice of $10\mu g/m^3$ as a threshold aligns with the new proposal for the National Ambient Air Quality Standard (NAAQS) established by the U.S. Environmental Protection Agency

(U.S. Environmental Protection Agency, 2022a).

Our proposed model is applied to a matched dataset, where the census and meteorological variables are used for the matching. Matching is commonly used in observational studies to adjust for potential measured confounding bias (Rosenbaum and Rubin, 1983). In this context, similarly to what has already been done in the literature on air pollution effects on health (see, e.g., Lee *et al.*, 2021; Wu *et al.*, 2020), we decide to use matching before running our model to make our analyses as robust as possible with respect to potential measured confounding bias.

We employ a 1-to-1 nearest neighbor propensity score matching without replacement, obtaining 1,402 selected units. The reduction of units is due to the different sample sizes of the treated and control groups in the original data, and 1-to-1 matching creates a sample with the same size for the treated and control groups since it finds a matched control unit for each treated unit. Using matching greatly improves the covariates balance, where the covariate balance is evaluated based on the difference in standardized means of the covariates. In Figure 4.2, we depict the covariate balance before and after the matching. Before the matching, there was a significant difference between the difference in standardized means of the observed values of some covariates, like poverty, education, or median household income, for the treated and control groups. These imbalances in the data might have led to spurious discoveries of effect heterogeneity. After the matching, the mean standardized differences of the covariates and the propensity score are included in the interval $[-0.1, 0.1]$, which is usually used as a rule-of-thumb for good quality matches (Ho *et al.*, 2007; Austin, 2011).

4.1.3 Results

We analyze the data with the confounders-dependent Bayesian mixture model, proposed in Chapter 2, considering as covariate, among all the confounders, the percentage of males, the percentage of white, Hispanic, black, and other races, the average age among the Medicare enrollees, and the percentage of people who are eligible for Medicare. CDBMM identifies six mutually exclusive groups in the matched ZIP codes: four where exposure to higher levels of $PM_{2.5}$ increases the mortality rate, and two where exposure to higher levels of $PM_{2.5}$ decreases the mortality rate. Figure 4.3 presents the posterior distribution of the GATEs and the GARRs for each identified group. We present results for both GATE and GARR as the information they provide is complementary, and furnishes a deeper insight into the heterogeneity in the causal effects in the case of our application. The vertical black lines in each figure represent the null causal effect, which is indicated by a GATE equal to 0 and by a GARR equal to 1.

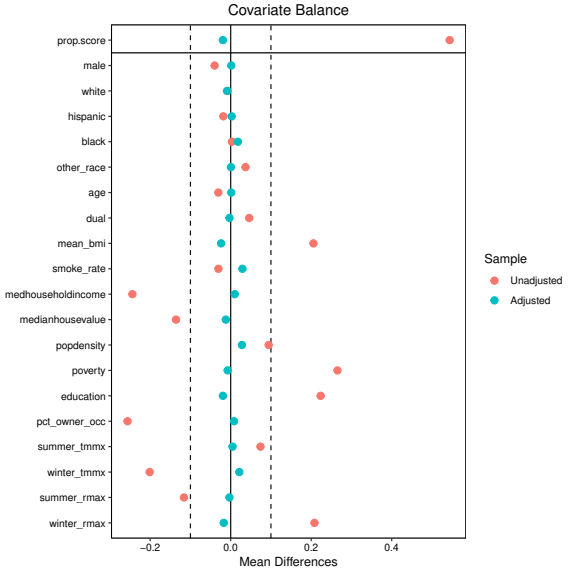

FIGURE 4.2: Comparison of the covariate balance between before and after the nearest neighbor propensity score matching 1-to-1. The continuous black vertical line indicates the value 0, while the two dotted lines are for the values -0.1 and 0.1, respectively.

The four groups identified by CDBMM where exposure to higher levels of $PM_{2.5}$ increases the mortality rate have positive GATE values (in order from highest to lowest: (f) 0.143, (e) 0.097, (d) 0.039, (c) 0.006) and have GARR values greater than 1 (in order: (f) 1.892, (e) 1.604, (d) 1.148, (c) 1.021). While the groups where exposure to higher levels of $PM_{2.5}$ decreases the mortality rate have negative GATE values (in order from lowest to highest: (a) -0.040 and (b) -0.007) and have GARR values less than 1 (in order: (a) 0.866 and (b) 0.979).

The majority of the population (90% of ZIP codes) is included in the groups (b), (c), and (d). In the bigger group (c)—including $45,4\%$ of the ZIP codes— the mortality rate of the population increases by 2% under a high level of $PM_{2.5}$ (corresponding to an increment of 0.006); the group (d)—including 13% of the ZIP codes—has an increment of 15% in the mortality (i.e., an increment of 0.039). Conversely, group (b), including $30,5\%$ of the ZIP codes, has a decrement of mortality by 2%—corresponding to a decrement of 0.007—under a high level of $PM_{2.5}$. Three small groups (a), (e), and (f) are also discovered (10% of ZIP codes). More extreme effects characterize these subgroups. An increment of up to 89% of the mortality rate when the sub-population in (f) is exposed to a high level of $PM_{2.5}$ and an increment of 60% for the group (e), while the group (a) has a decrement of 13% of the mortality rate when it is exposed to a high level of $PM_{2.5}$ instead of a lower level. I.e., the increment of the mortality rate is 0.143, 0.097, and -0.040, respectively for groups (f), (e), and (a). The percentages of ZIP codes allocated

in the various groups are reported in Figure 4.4.

Moreover, the uncertainty of the GATEs and GARRs can be quantified by the posterior distributions. In particular, all the posterior distributions of GATEs and GARRs are concentrated around the means, with light tails, and the 95% credible intervals of group (b) and group (c) do not include the value zero, for GATEs, and the value 1, for the GARR,—values that indicate the null causal effect of the $PM_{2.5}$ exposure on the mortality rate.

We would like to remind that there is also an uncertainty induced by the group allocation. Thanks to the choice of using the Wade and Ghahramani (2018)'s method allows us to know the posterior of the random partition, as explained in Section 2.2.2.

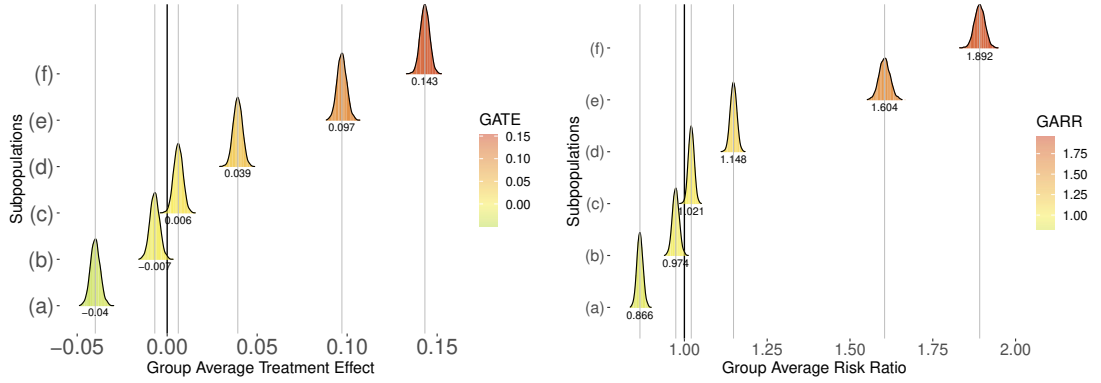

Figure 4.3: (Left) Posterior distribution of GATEs for the six estimated groups. (Right) Posterior distribution of CARR for the six estimated groups in the ZIP codes. In both plots, the black line identifies the null causal effects and the gray lines are the mean of each posterior distribution for the GATEs and the GARRs, respectively.

The demographic and socio-economic characteristics of the ZIP codes help us understand the differences in the causal effects in each identified group. The estimated model identifies mutually exclusive groups. Therefore, we can describe the distribution of these characteristics for the ZIP codes allocated in the different groups.

In the spider-plots reported in Figure 4.4, we can observe, for each group, the different distribution of the variables sex (close to the center indicates a higher percentage of women in the population of the ZIP codes, far to the center a higher percentage of men), percentage of white, Hispanic, black and other race (where smaller percentages are closer to the center), old (where the ZIP codes with age mean close to 65 years for Medicare enrollees are close to the center, and older population far from the center), and poor (the center of the spider-plot indicates a population with high income—i.e., a lower percentage of dually eligible individuals as proxy—, and far from the center lower

income). Each group is identified with a different color, while the grey area reports the mean of these variables among all the analyzed ZIP codes after matching.

The groups where exposure to higher levels of $PM_{2.5}$ increases the mortality rate are characterized by poorer populations for (c) and (d) and a higher percentage of black and/or other races for all four groups. Consistently with the intuition, these groups are characterized by higher percentages of people from minority groups for long-term exposure to $PM_{2.5}$. This could be likely due to the fact that minorities, often associated with low income, are structurally exposed to higher levels of air pollution. Exposure effects might accumulate over time—as is likely the case with $PM_{2.5}$ —leading to increased mortality rates (see, e.g., Pope III *et al.*, 2019; Liu *et al.*, 2021; Jbaily *et al.*, 2022). The groups (e) and (f) are characterized by a young (close to 65 years) and rich population compared to the mean among all the analyzed ZIP codes, specifically with white and Hispanic women in group (e) and male for group (f).

In juxtaposition, the groups (a) and (b) where exposure to higher levels of $PM_{2.5}$ decreases the mortality rate is mainly composed of a population with a higher percentage of Hispanics compared to the mean among all the analyzed ZIP codes and the other identified groups. In particular, group (a) is also composed of a big community of blacks and other races. This decrease in mortality when being exposed to higher levels of air pollution, while surprising, has already been documented in the literature (Liu *et al.*, 2021; Jbaily *et al.*, 2022). This finds an explanation in potential survival bias (see, e.g., Mayeda *et al.*, 2018; Shaw *et al.*, 2021). Survival bias happens in cohort studies that start later, leading to the most vulnerable individuals in certain groups dying before entering the cohort. In this case, the individuals entering the cohort are the most resilient ones and might depict a decreasing mortality effect even when exposed to higher levels of pollutants. This is likely to be the case for these two groups.

Moreover, Figure 4.5 investigates the spatial distribution of the discovered cluster in Texas, identified via our CDBMM model. In particular, we find that the clusters characterized by higher vulnerability are mostly located in southern Texas. The higher-vulnerability clusters are also found in suburban areas and along interstate highways between cities. Conversely, resilient clusters can be found in more sparsely populated areas. Gray areas could not be matched and thus are excluded from our analysis.

Figure 4.4: Representation of the characteristics of the identified groups. Each spider plot reports in the colored area the group-specific characteristics—the mean of the analyzed covariates—and in the gray area the collective characteristics—the mean of the covariates among all the analyzed Texan ZIP codes. We can consider the gray area as the benchmark to understand how the characteristics of each group differ from the collective characteristics of the analyzed Medicare enrollees in Texas.

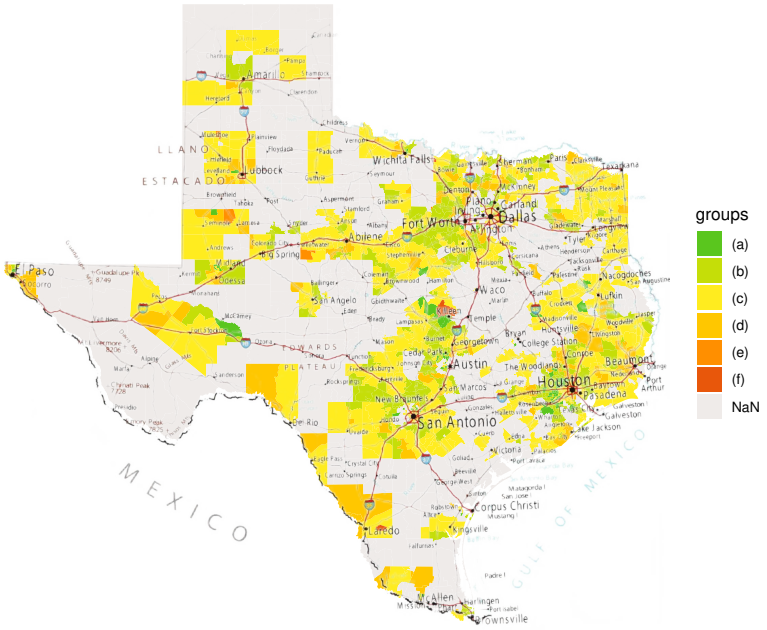

FIGURE 4.5: Representation of the identified clusters on the map of Texas.

4.2 Effect of Air Pollution Regularization on Mortality Risk

Aware of the causal link between air pollution and mortality and of the presence of groups in the population with different degrees of vulnerability to air pollution, due to socioeconomic disparities, the attainment of environmental justice seems imperative. The U.S. Environmental Protection Agency declares the goal to promote environmental justice, defined as "no group of people should bear a disproportionate burden of environmental harms and risks" (see U.S. Environmental Protection Agency, 2022c, page 116).

To achieve environmental justice and safeguard public well-being, the U.S. Environmental Protection Agency established the National Ambient Air Quality Standards (widely known as NAAQS). These air quality standards impose limits on the atmospheric concentration of air pollutants, with the goal of maintaining these pollutant concentrations below specified levels to preserve the overall health of the population in both the short and long term. In doing so, the U.S. Environmental Protection Agency designates as in *non-attainment* those counties whose annual average pollutant concentration exceeds the National Ambient Air Quality Standards levels. These counties are mandated to reduce the levels of pollutants in the subsequent years, applying air

quality regulations. Hence, it is critically important to understand how air quality regulations benefit health outcomes and how these effects can vary within different groups of individuals.

Following the 2003 report from the Health Effects Institute (HEI Accountability Working Group, 2003), the studies, that analyze the assessment of the extent to which air pollution impacts health outcomes, are defined as *accountability* studies. Accountability studies have been categorized into two branches: *indirect* and *direct* assessment studies (Zigler and Dominici, 2014a). On the one side, indirect accountability studies have focused on understanding what is the causal effect on the health of exposure to levels of air pollution (Wu *et al.*, 2020). On the other side, direct accountability studies have investigated the causal impact on the health of interventions aimed at reducing the level of pollutants in the air (Zigler *et al.*, 2018; Nethery *et al.*, 2020).

In this thesis, we (i) analyze the causal effects of the air quality regulations on the mortality rate, considering the variation of $PM_{2.5}$ levels as post-treatment variable, under the principal stratification framework, (ii) take advantage of the flexible Bayesian nonparametric mixture model for post-treatment variable—defined in Chapter 3—to understand how these effects can vary within strata, and (iii) characterize the identified different groups of individuals.

4.2.1 Data Description

To answer our research question we have merged two datasets: the information about the air pollution levels and the demographic and socio-economic characteristics in the counties in the Eastern United States, used in the Zigler *et al.* (2018)'s analysis, and the information about the age adjusted mortality rate in these counties available by the Center for Disease Control and Prevention of United States.

In particular, Zigler *et al.* (2018)'s dataset is composed of 384 counties in the Eastern United States, where national monitoring networks have detected the $PM_{2.5}$ concentration. In 2005, the U.S. Environmental Protection Agency designated as *nonattainment* of the NAAQS these counties where the average of the $PM_{2.5}$ concentration was above $15\mu g/m^3$, or otherwise *attainment*. States containing counties designated as nonattainment were required to develop or revise State Implementation Plans outlining how a nonattainment area will attain the standards with strategies to reduce ambient concentrations of $PM_{2.5}$.

As represented in Figure 4.6, the study design identifies the *baseline period* of 2000-2005 and the *follow-up period* of 2010-2016 during which we have the following information: (i) the average ambient concentration of $PM_{2.5}$, from Zigler *et al.* (2018)'s dataset;

(ii) the age-adjusted mortality rate of all-cause mortality, public available in the website of Center for Disease Control and Prevention of United States.

Moreover, for each county, we have census variables such as the percentage of Hispanic and black residents; the average household income; the percentage of females; the average house value; the proportion of residents in poverty; the proportion of residents with a high school diploma; the smoking rate; the population; the percentage of residents in urban area; the employment rate (the percentage of the workforce employed); the percentage of the move in the last 5 years. We also have access to meteorological variables: the averages of daily temperatures and the relative average of humidity; the dew point (the temperature at which air becomes saturated with moisture, leading to the formation of dew or condensation).

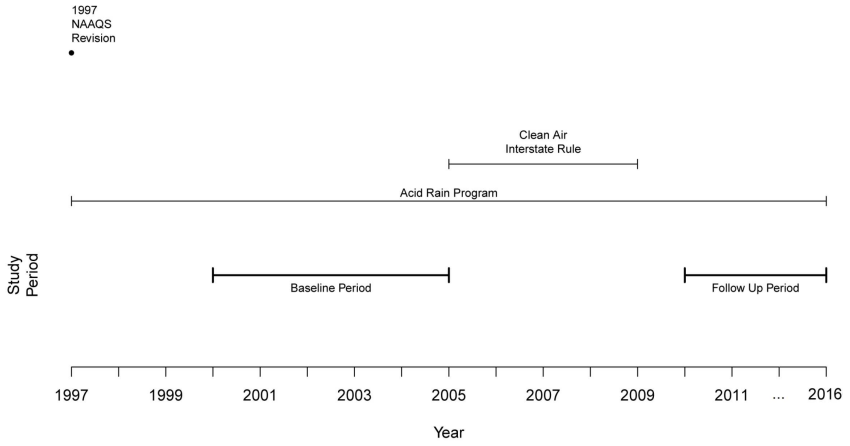

FIGURE 4.6: Timeline of National Ambient Air Quality Standards revisions, baseline and follow-up period (Zigler *et al.*, 2018).

We apply the principal stratification framework introduced in Section 1.2.1 using the following treatment, post-treatment variable, and outcome variable. The binomial treatment variable is the status designated by the U.S. Environmental Protection Agency for each county, such that $T = 1$ if the county is *nonattainment* and it had to develop or revise State Implementation Plans and $T = 0$ otherwise. In the top map in Figure 4.7, the red points visualize the treated counties, that seem to be closer to the main cities, such as Chicago, New York City, Washington DC, or Cleveland. The continuous post-treatment variable is the variation—i.e. the difference—of the $PM_{2.5}$ level between in the follow-up period and the baseline period, reported in the bottom left map in Figure 4.7. The outcome variable is defined as the variation of the age-adjusted mortality rate between the follow-up period and the baseline period, and it is visualized in the bottom

right map in Figure 4.7. As reported in the maps, both the post-treatment variable and the outcome have almost all negative values, underlining a general trend of decreasing the $PM_{2.5}$ level and an improvement in the quality of life in the last decade.

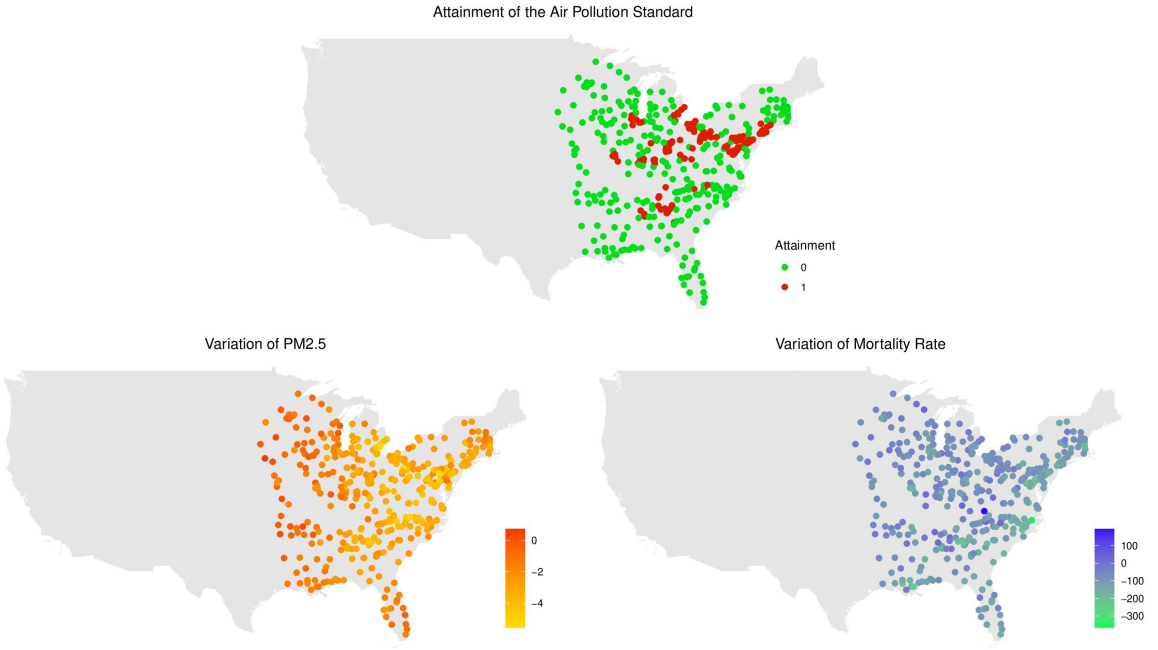

Figure 4.7: Considered counties in the Easter United States. (top) Attainment of the air pollution standard ($15\mu m$ for $PM_{2.5}$) and consequently application of air pollution regulations: 0 if the county was below the threshold (no pollution regulation had to be applied), 1 if the county was above the threshold and had to apply air pollution regulation for $PM_{2.5}$ reduction. (bottom left) Variation of the average of long-term $PM_{2.5}$ exposure between baseline and the follow-up period. (bottom right) Variation of the average of the age-adjusted mortality rate between baseline and the follow-up period, value per 100.000.

Before the employment of the model proposed in Chapter 3, an analysis of the covariate balance is due to understand the presence of the potential confounding bias. The analysis of the difference in standardized means of the covariates between the treated and control group shows a not significant confounding bias since all the values are included in the interval $[-0.1, 0.1]$, the rule-of-thumb for good quality matches (Ho *et al.*, 2007; Austin, 2011). Moreover, the use of matching techniques, such as 1-to-1 matching without replacement which is commonly used with this type of data, reduces the sample size that is already limited without improving significantly the covariate balance. Therefore, we chose to do not use any matching technique in this particular real-data analysis.

4.2.2 Results

We apply the proposed confounders-aware shared-atoms Bayesian mixture model to the previously described 384 counties in the Eastern United States, including all the covariates—census and meteorological variables—in the weights of the post-treatment variables mixture, while for the outcome model we use the linear model in (3.11). The model identifies the three strata: the dissociative stratum with 115 counties (30% of the total analyzed counties), the associative positive stratum with 48 of counties (12.5%), and the associative negative stratum with 221 counties (57.5%). According to the definition of the three strata and as visualized in the left image in Figure 4.8, the dissociative stratum—identified with the color yellow—is composed of counties where the application of the environmental plans do not substantially impact the level of $PM_{2.5}$, in fact, the expected value of $\mathbb{E}[P_i(1) - P_i(0)|V_i = 0]$ for the counties allocated to this strata has median close to zero and first and third quantiles of $-0.29\mu g/m^3$ and $0.07\mu g/m^3$. The associative negative stratum—identified with the color green—is composed of counties where the implementation of environmental plans decreases significantly the $PM_{2.5}$ levels. Specifically, in these counties the air quality regulation can reduce by $0.9\mu g/m^3$ in the median the $PM_{2.5}$ level, with first and third quantiles of $-1.46\mu g/m^3$ and $-0.38\mu g/m^3$. The associative positive stratum—identified with the color red—is composed of counties where the implementation of environmental plans increases the $PM_{2.5}$ levels by $0.8\mu g/m^3$ in the median. The latest stratum has the biggest uncertainly for $\mathbb{E}[P_i(1) - P_i(0)|V_i = -1]$, with values of the first and third quantiles equal to $0.13\mu g/m^3$ and $1.34\mu g/m^3$, respectively.

The corresponding distributions of the expected dissociative/associative effects are reported in the right image in Figure 4.8 and show the effect of the implementation of environmental plans on the mortality rate conditional to the three strata, i.e. conditional the heterogeneity in the causal effect in the level of pollution. The EDE and the EAE_+ has a similar increment in the median by $2.7‱$ of the mortality rate when the environmental plans are applied. While the means of these two effects are $0.8‱$ and $1.3‱$, respectively. For both the principal causal effects, the interquantile interval include the zero value. This mean that the counties where the implementation of environmental plans does not affect the $PM_{2.5}$ level or where this level increase, the mortality rate is not been affect by the implementation of environmental plans. The EAE_-—in color green in the right image of Figure 4.8- assumes negative values, indicating that the implementation of environmental plans, in the counties where the regulations affect the level of $PM_{2.5}$, reducing it significantly, reduces also the median age-adjusted mortality rate. Specifically, the mortality rate is decreased by $11.9‱$ in median and with the

interquantile interval of $[-15.2‰, -9.1‰]$.

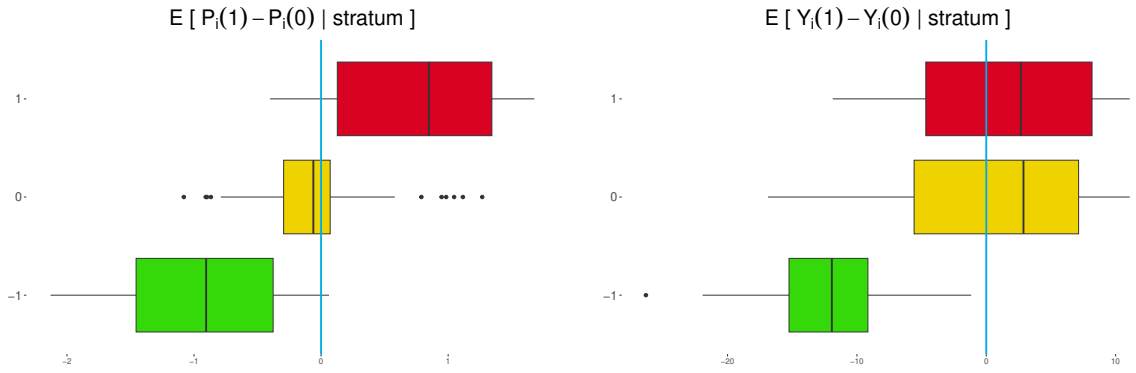

Figure 4.8: Boxplots of the three identified strata for (left) the conditional average of the difference of the post-treatment variables, and (right) the expected associative/dissociative effects. The light-blue vertical lines show the value zero, identifying the null effects.

Additionally, it is our interest to characterize the heterogeneity in the strata. Figure 4.9 visualizes the average of the observed covariates within each stratum—reported as the colored lines—and compares them with the average of the covariates among all the 348 observed counties in the Eastern United States.

The associative positive stratum and the dissociative stratum are composed of rural counties with a smaller population density, where the population has lower income with respect to the mean of the overall counties, and with a small employment rate. Differently, in the countries of the associative positive stratum, the percentage of Hispanics is smaller than the average, in favor of a bigger percentage of whites and Asians, while the dissociative stratum is characterized by a bigger percentage of Hispanics and a smaller community of black people, with a higher level of education. Moreover, the associative positive stratum is distinguished by a high smoke rate. In opposition, the associative negative stratum identifies urban counties with high population density. This population is characterized by higher income but also a bigger percentage of poor people and lower levels of education.

The meteorological variable, such as the averages of daily temperatures, the relative average of humidity, and the dew point seems to play an important role in the distinction of the strata, i.e. in the characterization of the different effects of the implementation of environmental plans on the level of $PM_{2.5}$.

Moreover, our proposed approach allows us to quantify the uncertainly of the strata allocation, indeed for each county we know the probability to be allocated in each of the three strata, in addition to the estimation of its allocation. Moreover, we can visualize

FIGURE 4.9: Representation of the characteristics of the identified strata. Each spider plot reports in the colored area the strata-specific characteristics—the mean of the analyzed covariates—and in the gray area the collective characteristics—the mean of the covariates among all the analyzed counties in the Eastern United States. We can consider the gray area as the benchmark to understand how the characteristics of each stratum differ from the collective characteristics of the analyzed population.

these information on the US map. Specifically, the first three maps in Figure 4.10 visualize the probability of each county to be allocated to the three different strata. As already underline Figure 4.9, the counties with higher probability to be allocated in the associative negative stratum and in the dissociative stratum are far from the biggest cities, differently to the associative positive stratum. Moreover the Western counties seem to have a small probability to be allocated in the associative negative stratum. The fourth map—bottom right in Figure 4.10—reports the partition point estimation of the strata, partition that is used to estimated the principal causal effects.

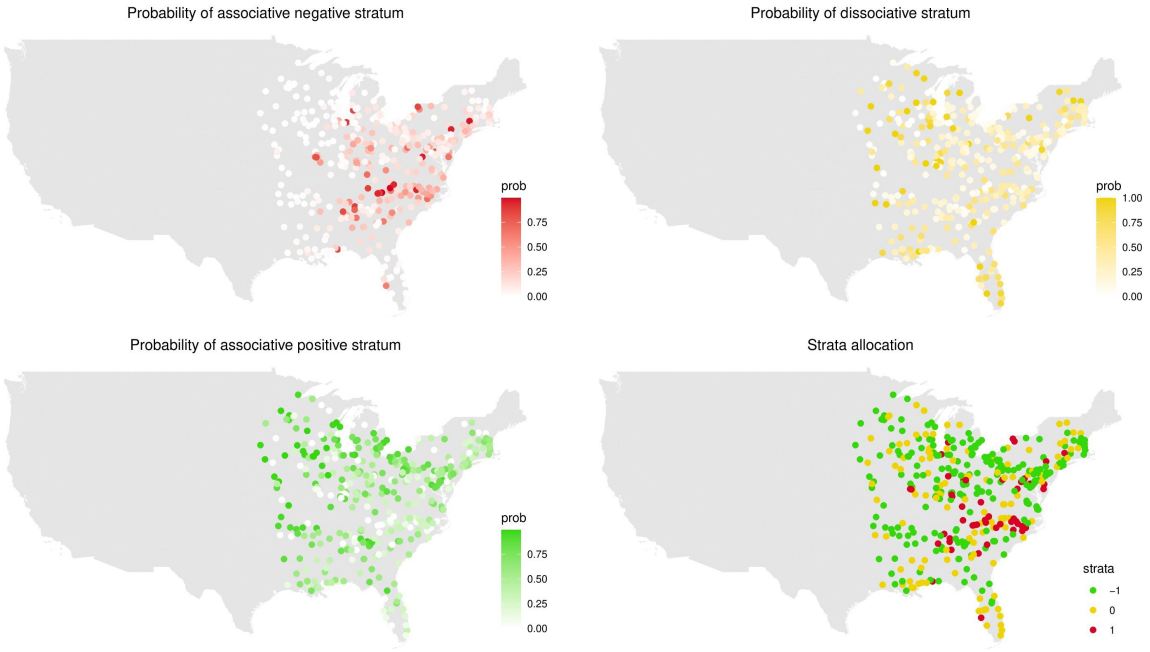

Figure 4.10: Considered counties in the Easter United States. (top left) Probability to be allocated in the associative positive stratum. (top right) Probability to be allocated in the dissociative stratum. (bottom left) Probability to be allocated in the associative negative stratum. (bottom right) Point estimation of strata allocation.

Conclusions

This thesis intertwines three different topics: (i) environmental epidemiology, (ii) the definition of causal estimands in the potential outcomes framework, and (iii) the use of flexible Bayesian nonparametric priors to define novel causal models.

We were initially motivated by the study of the possible causal link between air pollution and public health in its complexity, as the heterogeneity that characterizes the different levels of vulnerability/resilience among the population, with a particular interest in socio-economic disparities in the exposure, and the desire to understand the associative and dissociative causal effect of the air pollution regulation and the implementation of environmental plans on the mortality risk. The definition of causality in statistics entails a clear definition of the relation across the involved variables, making reasonable assumptions, and careful specification of the causal estimands, according to the causal questions. In particular, the estimation of specific estimands requires flexible models. We considered Bayesian nonparametric mixture models since they are well-known for their flexibility, adaptability, and clustering ability. In addition, the imputation of the missing variables, a problem arising in the potential outcome framework, is straightforward with the Bayesian paradigm via missing variable imputation from suitable posterior predictive distributions.

Clustering is the thread that binds the three topics: (i) it is an intrinsic characteristic of the Bayesian nonparametric mixture model, induced by the latent categorical variable that defines the probability of being allocated in the different components of the mixture; (ii) it is fundamental in causal inference framework to characterize the heterogeneity in the causal effect and define estimand as the group average treatment effects or the expected associative/dissociative effects; (iii) it divides the observed population into groups that have different causal effects of air pollution or its regulation on mortality rate between them and similar causal effect in the units inside the group, and allows us to identify the demographic and socio-economic characteristics of each subpopulation.

Specifically, in this thesis, we have introduced two novel causal models: the confounder-dependent mixture model, which captures the complex density structure of the potential outcome, and the confounders-aware shared-atoms mixture model, which

flexibly defines the distribution of potential post-treatment variable and discovers their strata. The dependent Dirichlet process as prior allows us to exploit rich forms of dependence given the confounders and relationship between the variables with different treatment levels. In fact, we have introduced, in both models, the confounders in the definition of the weights of the stick-breaking representation with the aim, not only of defining the dependence between the confounders and the outcome or post-treatment variable but also of imputing properly the missing data (typical problem in the potential outcome framework) and of characterizing the heterogeneity of the causal effects. Moreover, for the confounders-aware shared-atoms mixture model, we also allow to sharing of information between the two treatment levels of the potential post-treatment variables, defining common atoms between the two mixture distributions.

Exploiting suitable Bayesian nonparametric priors, we have tailored these two models for specific contexts of causal inference. It is common in observational studies to experience heterogeneity in the causal effects and identify different causal structures among the variables. Specifically, we have focused on two scenarios: firstly, in the presence of heterogeneity in the causal effect of the treatment on the outcome, we have theorized that the heterogeneity is induced by the presence of groups of units, characterized by the same causal effect in the group and different effects between them, and defined coherently the group average treatment effects as estimand of interest; secondly, we were interested in estimating the causal effect of a treatment on a primary outcome and to what extent this effect might change across values of a post-treatment variable, i.e. assuming the presence of three principal strata that identify the different effect of the treatment on the post-treatment variable, we define novel estimands for the causal effect of the treatment in the final outcome conditional to the strata.

Through simulation studies, we tested the ability of these two proposed models to estimate the specific estimands, comparing them with the benchmark model. The confounder-dependent mixture model is competitive with Bayesian additive regression trees and Bayesian causal forest in terms of estimating IATEs and is able to correctly identify the groups and estimate the GATEs with a high degree of accuracy. The confounders-aware shared-atoms Bayesian mixture model shows better performance of the model proposed by Schwartz *et al.* (2011) on the estimation of the average causal effect on the post-treatment variable and on the outcome, and additionally, it correctly identifies the strata and the respective expected dissociative and dissociative effects.

In the applications on environmental epidemiology, the models allow us to study, firstly the long-term exposure to $PM_{2.5}$ effects on the mortality rate in the Medicare enrollees in Texas, discovering six distinct groups that characterize the different levels

of vulnerability/resilience, and secondly the causal effect on the air pollution regulations on the age-adjusted mortality rate in the population of the Easter United State, conditional to the three discovered strata that are identified by the different variation of the $PM_{2.5}$ with and without the enforce of air pollution plans. In both applications, the characterization of the different groups or strata has identified the different ethnic composition of the population and the level of poverty as key distinction factors of different levels of vulnerability/resilience.

The proposed confounder-dependent mixture model and the confounders-aware shared-atoms mixture model are examples of how the flexibility and adaptability of Bayesian nonparametric mixtures can address specific questions that arise in various contexts of causal inference and real-world applications. However, similar Bayesian nonparametric models can be explored for different and numerous settings of causal inference framework, with careful attention to the causal question that we want to achieve and consequently the specific estimand that needs to be estimated. Moreover, real-world applications are flourishing stirring of new challenging contests and research questions in causal inference framework.

Bibliography

Albert, J. H. and Chib, S. (2001) Sequential ordinal modeling with applications to survival data. *Biometrics* **57**(3), 829–836.

Arellano-Valle, R. B. and Azzalini, A. (2006) On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**(3), 561–574.

Austin, P. C. (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46**(3), 399–424.

Baker, S. G., Lindeman, K. S. and Kramer, B. S. (2011) Clarifying the role of principal stratification in the paired availability design. *The International Journal of Biostatistics* **7**(1), 000010220215574679 1338.

Bargagli-Stoffi, F. J., Cadei, R., Lee, K. and Dominici, F. (2020) Causal rule ensemble: Interpretable discovery and inference of heterogeneous causal effects. *arXiv preprint arXiv:2009.09036* .

Bargagli-Stoffi, F. J., Cadei, R., Lee, K. and Dominici, F. (2023) Causal rule ensemble: Interpretable discovery and inference of heterogeneous causal effects. *arXiv preprint arXiv:2009.09036* .

Bargagli-Stoffi, F. J., De-Witte, K. and Gnecco, G. (2022) Heterogeneous causal effects with imperfect compliance: a Bayesian machine learning approach. *The Annals of Applied Statistics* (3), 1986–2009.

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**(6), 1173.

Barrientos, A. F., Jara, A. and Quintana, F. A. (2012) On the support of maceachern's dependent dirichlet processes and extensions. *Bayesian Analysis* **7**(2), 277–310.

Bell, M. L., Zanobetti, A. and Dominici, F. (2013) Evidence on vulnerability and susceptibility to health risks associated with short-term exposure to particulate matter: a systematic review and meta-analysis. *American Journal of Epidemiology* **178**(6), 865–876.

Binder, D. A. (1978) Bayesian cluster analysis. *Biometrika* **65**(1), 31–38.

Breiman, L. (2001) Random forests. *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Cart: Classification and regression trees. *Wadsworth and Brooks/Cole Monterey, CA, USA* .

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2007) Bayesian inference for linear dynamic models with dirichlet process mixtures. *IEEE Transactions on Signal Processing* **56**(1), 71–84.

Chen, J. and Hoek, G. (2020) Long-term exposure to PM and all-cause and cause-specific mortality: a systematic review and meta-analysis. *Environment International* **143**, 105974.

Chen, Y., Ebenstein, A., Greenstone, M. and Li, H. (2013) Evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. *Proceedings of the National Academy of Sciences* **110**(32), 12936–12941.

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.

Chung, Y. and Dunson, D. B. (2009) Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104**(488), 1646–1660.

Cochran, W. G. and Chambers, S. P. (1965) The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)* **128**(2), 234–266.

Cochran, W. G. and Rubin, D. B. (1973) Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A* pp. 417–446.

Cook, D. I., Gebski, V. J. and Keech, A. C. (2004) Subgroup analysis in clinical trials. *Medical Journal of Australia* **180**(6), 289–291.

Crabbé, J., Curth, A., Bica, I. and van der Schaar, M. (2022) Benchmarking heterogeneous treatment effect models through the lens of interpretability. *Advances in Neural Information Processing Systems* **35**, 12295–12309.

De la Cruz-Mesía, R., Quintana, F. A. and Müller, P. (2007) Semiparametric bayesian classification with longitudinal markers. *Journal of the Royal Statistical Society Series C: Applied Statistics* **56**(2), 119–137.

Dahabreh, I. J., Hayward, R. and Kent, D. M. (2016) Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International Journal of Epidemiology* **45**(6), 2184–2193.

Daniels, M. J., Linero, A. and Roy, J. (2023) *Bayesian Nonparametrics for Causal Inference and Missing Data*. Volume 124. CRC Press.

Daniels, M. J., Roy, J. A., Kim, C., Hogan, J. W. and Perri, M. G. (2012) Bayesian inference for the causal effect of mediation. *Biometrics* **68**(4), 1028–1036.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2013) Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence* **37**(2), 212–229.

De Iorio, M., Johnson, W. O., Müller, P. and Rosner, G. L. (2009) Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65**(3), 762–771.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An anova model for dependent random measures. *Journal of the American Statistical Association* **99**(465), 205–215.

Denti, F., Camerlenghi, F., Guindani, M. and Mira, A. (2021) A common atom model for the bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association* (just-accepted), 1–22.

Denti, F., Camerlenghi, F., Guindani, M. and Mira, A. (2023) A common atoms model for the bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association* **118**(541), 405–416.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F. and Schwartz, J. D. (2017) Air pollution and mortality in the medicare population. *New England Journal of Medicine* **376**(26), 2513–2522.

Ding, P. (2023) A first course in causal inference. *arXiv preprint arXiv:2305.18793* .

Ding, P. and Li, F. (2018) Causal inference: a missing data perspective. *Statistical Science* **33**(2), 214–237.

Ding, P., Lu, J. *et al.* (2017) Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B* **79**(3), 757–777.

Dominici, F., Bargagli-Stoffi, F. J. and Mealli, F. (2021) From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework. *Harvard Data Science Review* .

Dominici, F., Greenstone, M. and Sunstein, C. R. (2014) Particulate matter matters. *Science* **344**(6181), 257–259.

Dominici, F., Zanobetti, A., Schwartz, J., Braun, D., Sabath, B. and Wu, X. (2022) Assessing adverse health effects of long-term exposure to low levels of ambient air pollution: Implementation of causal inference methods. *Research Report (Health Effects Institute)* (211), 1–56.

Dorn, H. F. (1953) Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health* **43**(6_Pt_1), 677–683.

Dunson, D. B. and Park, J.-H. (2008) Kernel stick-breaking processes. *Biometrika* **95**(2), 307–323.

Durante, D. (2019) Conjugate bayes for probit regression via unified skew-normal distributions. *Biometrika* **106**(4), 765–779.

Dwivedi, R., Tan, Y. S., Park, B., Wei, M., Horgan, K., Madigan, D. and Yu, B. (2020) Stable discovery of interpretable subgroups via calibration in causal studies. *International Statistical Review* **88**, S135–S178.

Egleston, B. L. (2011) Response to pearl's comments on principal stratification. *The International Journal of Biostatistics* **7**(1), 000010220215574679133O.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**(430), 577–588.

Fasano, A., Durante, D. *et al.* (2022) A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research* **23**(30).

Feller, A., Mealli, F. and Miratrix, L. (2017) Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics* **42**(6), 726–758.

Ferguson, T. S. (1973) A bayesian analysis of some nonparametric problems. *The Annals of Statistics* pp. 209–230.

Ferguson, T. S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics* **2**(4), 615–629.

Fisher, R. A. (1919) The causes of human variability. *The Eugenics Review* **10**(4), 213.

Fisher, R. A. *et al.* (1936) Statistical methods for research workers. *Statistical Methods for Research Workers.* (6th Ed).

Foster, J. C., Taylor, J. M. and Ruberg, S. J. (2011) Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**(24), 2867–2880.

Frangakis, C. E. and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics* **58**(1), 21–29.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005) Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association* **100**(471), 1021–1035.

Gelman, A. and Vehtari, A. (2021) What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association* **116**(536), 2087–2097.

Gilbert, P. B., Hudgens, M. G. and Wolfson, J. (2011) Commentary on" principal stratification–a goal or a tool?" by judea pearl. *The International Journal of Biostatistics* **7**(1), 000010220215574679134 1.

Green, P. J. and Richardson, S. (2001) Modelling heterogeneity with and without the dirichlet process. *Scandinavian Journal of Statistics* **28**(2), 355–375.

Gutiérrez, L., Barrientos, A. F., González, J. and Taylor-Rodríguez, D. (2019) A bayesian nonparametric multiple testing procedure for comparing several treatments against a control .

Hahn, P. R. and Carvalho, C. M. (2015) Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* **110**(509), 435–448.

Hahn, P. R., Carvalho, C. M., Puelz, D. and He, J. (2018) Regularization and confounding in linear regression for treatment effect estimation .

Hahn, P. R., Murray, J. S. and Carvalho, C. M. (2020) Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis* **15**(3), 965–1056.

HEI Accountability Working Group (2003) Assessing the health impact of air quality regulations: Concepts and methods for accountability research. Technical report, Health Effects Institute.

Hernández, B., Raftery, A. E., Pennington, S. R. and Parnell, A. C. (2018) Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing* **28**(4), 869–890.

Hill, J. L. (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010) *Bayesian Nonparametrics*. Volume 28. Cambridge University Press.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**(3), 199–236.

Holland, P. W. (1986) Statistics and causal inference. *Journal of the American Statistical Association* **81**(396), 945–960.

Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. *Psychological Methods* **15**(4), 309.

Imai, K. and Ratkovic, M. (2013) Estimating treatment effect heterogeneity in randomized program evaluation .

Imbens, G. W. and Rubin, D. B. (1997) Bayesian inference for causal effects in randomized experiments with noncompliance. *The annals of statistics* pp. 305–327.

Ishwaran, H. and Rao, J. S. (2005) Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics* **33**(2), 730–773.

Jara, A., Lesaffre, E., De Iorio, M. and Quintana, F. (2010) Bayesian semiparametric inference for multivariate doubly-interval-censored data .

Jbaily, A., Zhou, X., Liu, J., Lee, T.-H., Kamareddine, L., Verguet, S. and Dominici, F. (2022) Air pollution exposure disparities across us population and income groups. *Nature* **601**(7892), 228–233.

Joffe, M. (2011) Principal stratification and attribution prohibition: good ideas taken too far. *The International Journal of Biostatistics* **7**(1), 00001022021557467913 67.

Kennedy, E. H. (2020) Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* .

Kennedy, E. H., Ma, Z., McHugh, M. D. and Small, D. S. (2017) Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(4), 1229–1245.

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.

Kioumourtzoglou, M.-A., Schwartz, J., James, P., Dominici, F. and Zanobetti, A. (2016) Pm2. 5 and mortality in 207 us cities: modification by temperature and city characteristics. *Epidemiology (Cambridge, Mass.)* **27**(2), 221.

Lee, K., Small, D. S. and Dominici, F. (2021) Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *Journal of the American Statistical Association* **116**(534), 569–580.

Li, F., Ding, P. and Mealli, F. (2023) Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A* **381**(2247), 20220153.

Li, Z., Konisky, D. M. and Zirogiannis, N. (2019) Racial, ethnic, and income disparities in air pollution: A study of excess emissions in Texas. *PloS One* **14**(8), e0220696.

Lijoi, A., Mena, R. H. and Prünster, I. (2007a) Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**(4), 769–786.

Lijoi, A., Mena, R. H. and Prünster, I. (2007b) Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**(4), 715–740.

Linero, A. R. (2018) Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association* **113**(522), 626–636.

Linero, A. R. and Antonelli, J. L. (2023) The how and why of Bayesian nonparametric causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics* **15**(1), e1583.

Linero, A. R. and Yang, Y. (2018) Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(5), 1087–1110.

Liu, M., Saari, R. K., Zhou, G., Li, J., Han, L. and Liu, X. (2021) Recent trends in premature mortality and health disparities attributable to ambient pm2. 5 exposure in china: 2005–2017. *Environmental Pollution* **279**, 116882.

Lo, A. Y. (1984) On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics* pp. 351–357.

Logan, B. R., Sparapani, R., McCulloch, R. E. and Laud, P. W. (2019) Decision making and uncertainty quantification for individualized treatments using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research* **28**(4), 1079–1093.

MacEachern, S. N. (2000) Dependent dirichlet processes. technical report. *Department of Statistics, The Ohio State University, Columbus, OH.* .

MacKinnon, D. (2012) *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, D. P. and Dwyer, J. H. (1993) Estimating mediated effects in prevention studies. *Evaluation Review* **17**(2), 144–158.

Mayeda, E. R., Filshtein, T. J., Tripodis, Y., Glymour, M. M. and Gross, A. L. (2018) Does selective survival before study enrolment attenuate estimated effects of education on rate of cognitive decline in older adults? a simulation approach for quantifying survival bias in life course epidemiology. *International Journal of Epidemiology* **47**(5), 1507–1517.

Mealli, F. and Mattei, A. (2012) A refreshing account of principal stratification. *The International Journal of Biostatistics* **8**(1).

Meilă, M. (2007) Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* **98**(5), 873–895.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**(404), 1023–1032.

Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66**(3), 735–749.

Müller, P. and Walker, S. (2010) Bayesian nonparametrics: Principles and practice.

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

Nethery, R. C., Mealli, F. and Dominici, F. (2019) Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The Annals of Applied Statistics* **13**(2), 1242.

Nethery, R. C., Mealli, F., Sacks, J. D. and Dominici, F. (2020) Evaluation of the health impacts of the 1990 clean air act amendments using causal inference and machine learning. *Journal of the American Statistical Association* pp. 1–12.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences* pp. 1–51.

Nie, X. and Wager, S. (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**(2), 299–319.

Oganisian, A., Mitra, N. and Roy, J. (2020a) Bayesian nonparametric cost-effectiveness analyses: Causal estimation and adaptive subgroup discovery. *arXiv preprint arXiv:2002.04706* .

Oganisian, A., Mitra, N. and Roy, J. (2020b) Hierarchical bayesian bootstrap for heterogeneous treatment effect estimation. *arXiv preprint arXiv:2009.10839* .

Oganisian, A., Mitra, N. and Roy, J. A. (2021) A bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *Biometrics* **77**(1), 125–135.

O'Hagan, A. (1978) Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)* **40**(1), 1–24.

Pearl, J. (2009) *Causality*. Cambridge university press.

Pearl, J. (2011) Principal stratification–a goal or a tool? *The International Journal of Biostatistics* **7**(1), 1–13.

Pearl, J. *et al.* (2000) Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* **19**(2), 3.

Perman, M., Pitman, J. and Yor, M. (1992) Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields* **92**(1), 21–39.

Pitman, J. and Yor, M. (1997) The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability* pp. 855–900.

Pope III, C. A., Lefler, J. S., Ezzati, M., Higbee, J. D., Marshall, J. D., Kim, S.-Y., Bechle, M., Gilliat, K. S., Vernon, S. E., Robinson, A. L. *et al.* (2019) Mortality risk and fine particulate air pollution in a large, representative cohort of us adults. *Environmental Health Perspectives* **127**(7), 077007.

Prentice, R. (2011) Invited commentary on pearl and principal stratification. *The International Journal of Biostatistics* **7**(1), 000010220215574679 1359.

Qian, M. and Murphy, S. A. (2011) Performance guarantees for individualized treatment rules. *Annals of Statistics* **39**(2), 1180.

Quintana, F. A. (2006) A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference* **136**(8), 2407–2429.

Quintana, F. A., Mueller, P., Jara, A. and MacEachern, S. N. (2020) The dependent dirichlet process and related models. *arXiv preprint arXiv:2007.06129* .

Ren, L., Du, L., Carin, L. and Dunson, D. B. (2011) Logistic stick-breaking process. *Journal of Machine Learning Research* **12**(1).

Rodriguez, A. and Dunson, D. B. (2011) Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis (Online)* **6**(1).

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008) The nested dirichlet process. *Journal of the American Statistical Association* **103**(483), 1131–1154.

Rodriguez, A. and Ter Horst, E. (2008) Dynamic density estimation with financial applications. *Bayesian Analysis* **3**, 339–366.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55.

Roy, J., Lum, K. J., Zeldow, B., Dworkin, J. D., Re III, V. L. and Daniels, M. J. (2018) Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics* **74**(4), 1193–1202.

Roy, S., Daniels, M. J., Kelly, B. J. and Roy, J. (2022) A Bayesian nonparametric approach for causal inference with multiple mediators. *arXiv preprint arXiv:2208.13382* .

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**(5), 688.

Rubin, D. B. (1978) Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* pp. 34–58.

Rubin, D. B. (1980) Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association* **75**(371), 591–593.

Rubin, D. B. (1981) The bayesian bootstrap. *The Annals of Statistics* pp. 130–134.

Rubin, D. B. (1986) Comment: Which ifs have causal answers. *Journal of the American Statistical Association* **81**(396), 961–962.

Rückerl, R., Schneider, A., Breitner, S., Cyrys, J. and Peters, A. (2011) Health effects of particulate air pollution: a review of epidemiological evidence. *Inhalation Toxicology* **23**(10), 555–592.

Samet, J. M. (2011) The clean air act and health—a clearer view from 2011. *New England Journal of Medicine* **365**(3), 198–201.

Schlesinger, R. B. (2007) The health impact of common inorganic components of fine particulate matter (pm2. 5) in ambient air: a critical review. *Inhalation toxicology* **19**(10), 811–832.

Schwartz, J. (2006) Long-term effects of exposure to particulate air pollution. *Clinics in occupational and environmental medicine* **5**(4), 837–848. `https://doi.org/10.1016/j.coem.2006.07.008`.

Schwartz, J., Bind, M.-A. and Koutrakis, P. (2017) Estimating causal effects of local air pollution on daily deaths: effect of low levels. *Environmental Health Perspectives* **125**(1), 23–29.

Schwartz, S. L., Li, F. and Mealli, F. (2011) A bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association* **106**(496), 1331–1344.

Semenova, V. and Chernozhukov, V. (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* **24**(2), 264–289.

Sethuraman, J. (1994) A constructive definition of dirichlet priors. *Statistica Sinica* pp. 639–650.

Shaw, C., Hayes-Larson, E., Glymour, M. M., Dufouil, C., Hohman, T. J., Whitmer, R. A., Kobayashi, L. C., Brookmeyer, R. and Mayeda, E. R. (2021) Evaluation of selective survival and sex/gender differences in dementia incidence using a simulation model. *JAMA network open* **4**(3), e211001–e211001.

Sivaganesan, S., Müller, P. and Huang, B. (2017) Subgroup finding via Bayesian additive regression trees. *Statistics in Medicine* **36**(15), 2391–2403.

Stephens, D. A., Nobre, W. S., Moodie, E. E. and Schmidt, A. M. (2022) Causal inference under mis-specification: adjustment based on the propensity score. *arXiv preprint arXiv:2201.12831* .

Teh, Y., Jordan, M., Beal, M. and Blei, D. (2004) Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems* **17**.

Teh, Y. W. and Jordan, M. I. (2009) Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics* **28**(158), 42.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical dirichlet processes. *Journal of the American Statistical Association* **101**(476), 1566–1581.

U.S. Census Bureau (2020) QuickFacts: Texas. `https://www.census.gov/quickfacts/TX`.

U.S. Environmental Protection Agency (2022a) Regulatory impact analysis for the proposed reconsideration of the national ambient air quality standards for particulate matter. *Technical Report: EPA-452/P-22-001* .

U.S. Environmental Protection Agency (2022b) Regulatory impact analysis for the proposed reconsideration of the national ambient air quality standards for particulate matter. *Technical Report: EPA-452/P-22-001* .

U.S. Environmental Protection Agency (2022c) Reconsideration of the national ambient air quality standards for particulate matter. *Technical Report: EPA-452/P-22-001* .

U.S. Environmental Protection Agency (2022d) Supplement to the 2019 integrated science assessment for particulate matter. *Technical Report: EPA/635/R-22/028* .

VanderWeele, T. J. (2011) Principal stratification–uses and limitations. *The International Journal of Biostatistics* **7**(1), 1–14.

Wade, S., Dunson, D. B., Petrone, S. and Trippa, L. (2014) Improving prediction from Dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research* **15**(1), 1041–1071.

Wade, S. and Ghahramani, Z. (2018) Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis* **13**(2), 559–626.

Wade, S., Mongelluzzo, S. and Petrone, S. (2011) An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis* **6**(3), 359–385.

Wang, Y., Kloog, I., Coull, B. A., Kosheleva, A., Zanobetti, A. and Schwartz, J. D. (2016) Estimating causal effects of long-term PM2.5 exposure on mortality in New Jersey. *Environmental Health Perspectives* **124**(8), 1182–1188.

Wu, X., Braun, D., Kioumourtzoglou, M.-A., Choirat, C., Di, Q. and Dominici, F. (2019) Causal inference in the context of an error prone exposure: air pollution and mortality. *The Annals of Applied Statistics* **13**(1), 520.

Wu, X., Braun, D., Schwartz, J., Kioumourtzoglou, M. and Dominici, F. (2020) Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science Advances* **6**(29), eaba5692.

Zeng, J. and Wang, R. (2022) A survey of causal inference frameworks. *arXiv preprint arXiv:2209.00869* .

Zigler, C. M., Choirat, C. and Dominici, F. (2018) Impact of National Ambient Air Quality Standards nonattainment designations on particulate pollution and health. *Epidemiology (Cambridge, Mass.)* **29**(2), 165.

Zigler, C. M. and Dominici, F. (2014a) Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology. *American Journal of Epidemiology* **180**(12), 1133–1140.

Zigler, C. M. and Dominici, F. (2014b) Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association* **109**(505), 95–107.

Zigler, C. M., Dominici, F. and Wang, Y. (2012) Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13**(2), 289–302.

Zigler, C. M., Kim, C., Choirat, C., Hansen, J. B., Wang, Y., Hund, L., Samet, J., King, G., Dominici, F. *et al.* (2016) Causal inference methods for estimating long-term health effects of air quality regulations. *Research Report (Health Effects Institute)* (187), 5–49.

Zorzetto, D., Bargagli-Stoffi, F. J., Canale, A. and Dominici, F. (2023a) Confounder-dependent bayesian mixture model: Characterizing heterogeneity of causal effects in air pollution epidemiology. *arXiv preprint arXiv:2302.11656* .

Zorzetto, D., Bargagli-Stoffi, F. J., Canale, A., Mealli, F. and Dominici, F. (2023b) Bayesian nonparametrics for principal stratification: an application on environmental policies effects on health. *Work in progress* .

Dafne Zorzetto

Department of Statistical Science, Università degli Studi di Padova
Via C. Battisti 241, 35121, Padova, Italy
dafne.zorzetto@phd.unipd.it

Updated: September 2023

RESEARCH INTERESTS

Bayesian Nonparametrics: Models and Computational Aspects, Dependent Dirichlet Mixture Models.
Causal Inference: Heterogeneity of Causal Effects, Principal Stratification, and Negative Controls for Unmeasured Confounding.

CURRENT POSITION

Ph.D. in Statistics *at Università degli Studi di Padova*
Visiting Ph.D. Scholar *at Harvard University*
working with Antonio Canale (Università degli Studi di Padova), Francesca Dominici (Harvard University), and Falco J. Bargagli Stoffi (Harvard University) on Bayesian Nonparametrics in the context of Causal Inference.

EDUCATION

Ph.D. in Statistics *October 2020 - Present*
Department of Statistical Science, Università degli Studi di Padova.
Supervisor: Antonio Canale. Co-supervisor: Francesca Dominici.

Visiting Ph.D. Scholar (2 years) *January 2022 - Present*
Department of Biostatistics, Harvard T.H. Chan School of Public Health.

M.Sc. in Statistical Sciences *2018 - 2020*
Department of Statistical Science, Università degli Studi di Padova.
Thesis title: Hierarchical Bayesian models for extreme values in the cylinder.
Supervisor: Antonio Canale.
Honors: 110/110 cum laude.

B.Sc. in Statistics for Economics and Business *2015 - 2018*
Department of Statistical Science, Università degli Studi di Padova.
Thesis title: Inflation forecasting with GARCH models.
Supervisor: Luisa Bisaglia.

PUBLICATIONS

Publications & Manuscripts

- Hu J.*,**Zorzetto D.***, Dominici F. *A Bayesian Nonparametric Method to Adjust for Unmeasured Confounding with Negative Controls.* `https://arxiv.org/abs/2309.02631`[pdf]
- **Zorzetto D.**, Bargagli-Stoffi F.J., Canale A., Dominici F. *Confounder-Dependent Bayesian Mixture Model: Characterizing Heterogeneity of Causal Effects in Air Pollution Epidemiology.* `http://arxiv.org/abs/2302.11656` [pdf]
- **Zorzetto D.**, Bargagli-Stoffi F.J., Canale A., Dominici F. (2022). *Dependent Dirichlet Mixture Processes for Causal Inference.* Proceedings of the 36th International Workshop on Statistical Modelling. (pp. 618 - 623)

* the authors contributed equally to the work, alphabetically ordered by surnames.

Manuscripts in preparation

- **Zorzetto D.**, Bargagli-Stoffi F.J., Canale A., Dominici F., Mealli F. *Bayesian Nonparametric Mixture Model to Identify Principal Strata based on the Heterogeneous Causal Effects of Treatment on Post-Treatment Variable.*
- Alfonzetti G.*, Rossi L.*, **Zorzetto D.***, Mealli F., *Model-free estimation of causal effects of different stimuli on neuron activities.*
- **Zorzetto D.**, Canale A., Marani M. *Intensity of extreme epidemics.*
- Vanciu L., **Zorzetto D.**, Dominici F. *Bayesian Spatial Analysis of Mortality Disparities across the United States*

* the authors contributed equally to the work, alphabetically ordered by surnames.

AWARDS

- Young researcher travel award, the 2022 ISBA world meeting.

CONFERENCES PRESENTATIONS

Invited talks

- Confounder Dependent Bayesian Mixture Model: Application in Environmental Epidemiology.
  GRASPA 2023.
  Palermo (Italy), July 2023.
- Dependent nonparametric priors for causal inference problems.
  BNP-ISBA webseminar. Joint presentation with Antonio Canale.
  Online, June 2023
- Confounder Dependent Bayesian Mixture Model: Application in Environmental Epidemiology.
  NESS - 36th New England Statistics Symposium: Statistics and Data science.
  Boston (Massachusetts, USA), June 2023.

Contributed talks

- Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida
  SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation.
  Ancona (Italy), June 2023
- Bayesian Nonparametric for Causal Inference.
  BNP13 - 13th International Conference on Bayesian Nonparametrics.
  Puerto Varas (Chile), October 2022.
- Dependent Dirichlet mixture processes for Causal Inference.
  BaYSM 2022. Centre de recherches mathématiques at Université de Montréal.
  Montréal (Canada). June 2022

Poster presentations

- Bayesian Nonparametrics for Principal Stratification: an Application on Environmental Policies Effects on Health
  Bayesian Causal Inference Summer school.
  Firenze (Italy), July 2023.
- Bayesian Nonparametric for Heterogeneity in Treatment Effect.
  Atlantic Causal Inference Conference
  Austin (Texas, USA), May 2023
- Probit Stick-Breaking Process for Causal Inference.
  36th International Workshop on Statistical Modelling.
  Trieste (Italy), July 2022.
- Probit Stick-Breaking Process for Causal Inference.
  The 2022 ISBA World Meeting.
  Montréal (Canada), July 2022.

WORKSHOPS & SUMMERSCHOOL

- Data Research Camp
  San Servolo island, Venice, Italy. July 2022.
  4-day meeting where small research groups of young scholars, advised by senior researchers with a well-established experience in different areas of Statistics. We developed innovative methods and models to analyze a dataset—recorder neuron activities—, with the goal of answering scientific questions.
- Bayesian Causal Inference
  Florence Center for Data Science, University of Florence, Italy. June 2023.
  A week summer school about the fundamental concepts and the state-of-the-art methods for causal inference under the potential outcomes framework, with an emphasis on the Bayesian inferential paradigm.

MENTORING EXPERIENCES

- *Statistics.* Academic Tutor, Department of Biology.
  Università degli Studi di Padova, 2020.

ADVISING & CO-ADVISING

- Leo Vanciu *Summer 2023*
  Bachelor's Student, Harvard College
- Francesco Martella *Spring 2023*
  Master's student, Università degli Studi di Padova

SERVICES TO PROFESSION

- Organizer of *Explain like I'm an Undergrad*
  Padova, Italy. Spring 2023
  Weekly seminars that want to foster connections among PhD students and postdocs in the statistics department at Padova, as well as beyond, and provide with the opportunity to deliver engaging talks using a lighthearted, concise, and accessible presentation style, like to explaining complex concepts to undergraduate students in Statistics.
- Volunteer for *StatisticAll*
  Treviso, Italy. 2016
  Statistical games and activities to show the magic of statistics to kids and adults, in collaboration with ISTAT (Italian Statistics Institute)

SKILLS

Programming

Programming languages: R ; Python; Matlab.
Other statistical Software: Excel; SAS; MySQL.
Markup Languages: HTML; LaTex; Markdown.
Other Software: Git/GitHub; Windows and relative software.

Languages

Italian: native;
English: full professional proficiency.