Master's thesis in Data Science

Causal Rule Ensemble: interpretable discovery and estimation of Heterogeneous Treatment Effects

School of Computer and Communication Sciences, EPFL
Department of Biostatistics, Harvard T.H. School of Public Health

Author
Riccardo Cadei

Supervisor (Harvard)
Dr. Danielle Braun

Professor (EPFL)
Prof. Dr. Negar Kiyavash

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Boston, MA
March 2023

Abstract

In health and social sciences, it is critically important to identify subgroups of the study population where a treatment has notable heterogeneity in the causal effects with respect to the average treatment effect. Data-driven discovery of heterogeneous treatment effects (HTE) via decision tree methods has been proposed for this task. Despite its high interpretability, the single-tree discovery of HTE tends to be highly unstable and to find an oversimplified representation of treatment heterogeneity. To accommodate these shortcomings, we propose Causal Rule Ensemble (CRE), a new method to discover heterogeneous subgroups through an ensemble-of-trees approach. CRE provides (i) an interpretable representation of the HTE, (ii) via an extensive exploration of complex heterogeneity patterns, while (iii) guaranteeing high stability in the discovery. The discovered subgroups are defined in terms of interpretable decision rules, and we develop a general two-stage approach for subgroup-specific conditional causal effects estimation, providing theoretical guarantees. Via simulations, we show that the CRE method has a strong discovery ability and a competitive estimation performance when compared to state-of-the-art techniques. Finally, we apply CRE to discover subgroups most vulnerable to the effects of exposure to air pollution on mortality for 35.3 million Medicare beneficiaries across the contiguous U.S.

Keywords: *Causal Inference, Heterogeneous Treatment Effects, Interpretability, Machine Learning, Air Pollution Epidemiology*

Acknowledgement

First and foremost, I want to thank Prof. Francesca Dominici for welcoming me into her research group and supporting my research as much as my passion for running. Likewise, I am very grateful to my advisor Falco, for his invaluable guidance throughout this project, and his characteristic positive attitude. A big thanks to everyone else in the NSAPH group, who made me feel the on-campus vibes I have strongly missed during my Master's due to Covid restrictions.

I am grateful to my fellow master students for all the inspiring conversations during these years. To Piersilvio, for introducing me to Machine Learning first (in 2018), and Causal Inference then (in 2021); and for all the challenges, retreats and experiences we shared. To Andrea, for all the wild adventures we faced together with not more than a backpack, a pair of shoes and a ukulele. Keeping me passionate about exploration, fundamental principle in research too. To Raphael, for sharing the challenges and fun of this Master's together, from Lausanne to Boston. To all the academic mentors I encountered during my career.

To all the beautiful people who supported this academic journey indirectly, outside the university campus, filling my life with love and energy. To my great family, which has always supported all my passions, including my studying.

Contribution

This work is based on an idea from Dr. Falco J. Bargagli Stoffi, and it was developed during my 6 months visiting period at the Department of Biostatistics in the T.H. Chan Harvard School of Public Health under the supervision of Dr. Bargagli Stoffi, and as a member of the National Studies on Air Pollution and Health group, directed by Prof. Dr. Francesca Dominici. I joined the project when a preliminary version of the Causal Rule Ensemble had already been proposed. However, its formulation and performances, both in simulation and real-world experiments, still need to be fully explored.

My first contribution to this project is methodological. I revisited and reorganized the algorithm: simplifying where possible (i.e., estimation step) and further developing where necessary (i.e., rules generation, rules selection, individual treatment effect estimation). Based on this new formulation of the Causal Rule Ensemble, I have then introduced the mathematical formulation of the Treatment Effect linear decomposition for interpretable inference of the heterogeneous treatment effect (which was still missing). A new draft of the paper, of which I am a co-author, will be made available on arXiv soon (Bargagli-Stoffi et al.; 2023). My second contribution is to the software. I immediately joined the implementation of the Causal Rule Ensemble in `R`, and in November 2022, we released it on `CRAN` as an `R` official package (Khoshnevis et al.; 2023). Full documentation for the package can be found at `https://nsaph-software.github.io/CRE/`, and in its corresponding Software paper (Cadei et al.; 2023) (under review). As of today, I am the main contributor to this repository, and the package has already been downloaded more than 1,250 times. My last contribution is experimental. I re-designed and enriched the simulation studies, also introducing new evaluation metrics. I then focused on a real-world application of interpretable inference of the heterogeneous causal effect of fine particulate matter ($PM_{2.5}$) exposure on mortality, which is also the main motivation behind the algorithm. Thanks to the above contributions in both the algorithm and its implementation, I could scale the analyses to a national scale (35.3 million observations), extracting meaningful results, partially novel and in agreement with the existing literature.

Together with Dr. Bargagli Stoffi, we are currently submitting this definitive Causal Rule Ensemble formulation to a leading journal in the fields of Statistics and Causal Inference. Based on the positive experience of this visit, I will continue collaborating with this group for at least until the end of the summer, and we plan

to extend this work both methodologically and in the applications, as discussed in the conclusion.

Contents

Chapter 1

Introduction

1.1 Motivating Application

The U.S. Environmental Protection Agency (EPA) has recently set the goal to achieve environmental justice by addressing the disproportionate vulnerabilities in adverse human health effects due to exposure to air pollution (U.S. Environmental Protection Agency; 2022b). According to the EPA, environmental justice is defined as "*no group of people should bear a disproportionate burden of environmental harms and risks*" (see U.S. Environmental Protection Agency; 2022a, page 116). In the effort to promote environmental justice, the EPA has called for scientific studies that would inform the understanding of disproportionate health impacts of air pollution, with particular attention on demographic-specific information (U.S. Environmental Protection Agency; 2022a). Despite strong evidence that exposure to air pollution increases the risk of mortality and morbidity (see, e.g., Schwartz et al.; 2021; Wu, Braun, Schwartz, Kioumourtzoglou and Dominici; 2020; Nethery et al.; 2020; Carone et al.; 2020), little is known about which are the subgroups—i.e., subsets of the population characterized by a given covariate-profile (e.g., female individuals, low-income & male individuals)—who are most vulnerable or resilient to exposure to higher levels of air pollution.

Previous air pollution vulnerability studies are few and limited in their scope. Lee et al. (2021) and Zorzetto et al. (2023) recently proposed to causally assess exposure heterogeneity in air pollution via machine learning and Bayesian non-parametric methodologies, respectively. However, these studies are limited by the scalability of the employed methods, and their coverage is restricted to selected areas of the United States—i.e., New England and California. Di et al. (2017) estimated associations between long-term exposure to air pollution and mortality rates for pre-specified population subgroups defined by age, gender, and race categories. Despite its national coverage, this study has the main limitations of not directly answering a causal question, but an associational one and providing a very limited heterogeneity exploration—i.e., heterogeneous associations are estimated just for a predefined and very limited set of characteristics (e.g., sex, age, race). Thus, in spite of the urgency

of a nationwide study that would extensively explore the heterogeneous health effects of air pollution, a national study on this topic is not yet available.

Furthermore, most analyses on the effects of fine particulate matter ($PM_{2.5}$) on human health are conducted at the ZIP code or at the county level. Nevertheless, such analyses may mask important individual-level sources of heterogeneity and, most importantly, might expose the results to ecological fallacy (Freedman; 1999). The ecological fallacy, also known as the ecological inference fallacy or population fallacy, refers to the incorrect interpretation of results of statistical analyses, where conclusions about individuals are drawn from inferences made about the group to which they belong. Such fallacies in air pollution epidemiology studies have been recently acknowledged (see, e.g., Wu, Netherly, Sabath, Braun and Dominici; 2020). To our knowledge, no study has yet considered the heterogeneous causal effects of exposure to air pollution at an *individual level*.

To goal of our motivating application is to accommodate for this shortcoming and answer the EPA call by providing nationwide data-driven evidence regarding the most vulnerable subgroups to exposure to air pollution via an individual-level analysis. In particular, we aim to develop new methods in causal inference and machine learning with the goal of identifying de novo which subgroups of the Medicare population are most vulnerable or resilient to long-term exposure to $PM_{2.5}$ on mortality.

To do so, we acquired and integrated the data on 35,331,290 Medicare beneficiaries (i.e., individuals 65 years of age or older) across the entire United States for the period 2010-2016. We consider a binary exposure, indicating whether each individual has been exposed to $PM_{2.5}$ greater than 12 $\mu g/m^3$ or not. This exposure is the current National Ambient Air Quality Standard (NAAQS) set by the EPA. We link exposure to two-year annual $PM_{2.5}$ during 2010-2011 at the zip code level to mortality during the 5-years period 2012-2016 and several potential confounders, both at the individual, zip-code, and county level. Our study focuses on exploring the heterogeneity in the causal effects within the four U.S. census geographic regions—namely, Northeast, Midwest, West, and South—that are often utilized in investigations related to the impact of air pollution exposure. More details about the study design and results are illustrated in Chapter 5.

1.2 Contribution and Related Works

The bulk of heterogeneous treatment effect (HTE) literature focuses on two major tasks (Dwivedi et al.; 2020): (i) estimating HTEs by examining the conditional average treatment effect (CATE); (ii) discovering subgroups of a population characterized by HTE.

Seminal works on estimating the CATE rely on nearest-neighbor matching and kernel methods (Crump et al.; 2008; Lee; 2009). Other non-parametric machine learning methods such as the random forest (Breiman; 2001) and Bayesian additive regression tree (BART) (Chipman et al.; 2010) have been extended to estimate heterogeneity in causal effects—see, e.g., Foster et al. (2011), Hill (2011) and Hahn

et al. (2020). Wager and Athey (2018) and Athey et al. (2019) developed forest-based methods for the estimation of HTEs. They also provide an asymptotic theory for the conditional treatment effect estimators and valid statistical inference. Recently, two-stage doubly robust CATE estimators have been proposed first to generate doubly-robust pseudo outcomes and then regress them onto an a priori defined set of effect modifiers (Kennedy; 2020; Semenova and Chernozhukov; 2021).

Various methodologies have also been proposed to identify subgroups characterizing the heterogeneity in treatment effects (Imai et al.; 2013; Qian and Murphy; 2011; Kennedy et al.; 2017; Nie and Wager; 2017). Some methods first estimate the CATE as a function of some set of covariates and then identify heterogeneous subgroups in a second stage (Foster et al.; 2011; Bargagli-Stoffi et al.; 2020; Hahn et al.; 2020; Bargagli-Stoffi, De-Witte and Gnecco; 2022). Another approach is the direct data-driven discovery of heterogeneous subgroups (Wang and Rudin; 2022; Nagpal et al.; 2020). Many of the methodologies in this category are decision tree-based methodologies (see, e.g., Athey and Imbens; 2016; Bargagli-Stoffi and Gnecco; 2020; Lee et al.; 2021; Yang et al.; 2021; Bargagli-Stoffi et al.; 2020; Bargagli-Stoffi, De-Witte and Gnecco; 2022). Tree-based approaches have been widely adopted for treatment effect heterogeneity due to their appealing features. In fact, these methods are based on efficient and easily implementable recursive mathematical programming (e.g., maximization in the heterogeneous treatment effects), they can be easily tweaked and adapted to different scenarios on the basis of the research question of interest, and they guarantee a high degree of interpretability.

Despite their appealing features, single-tree heterogeneity discovery is characterized by two main limitations: (i) instability in the identification of the subgroup, and (ii) reduced exploration of the potential heterogeneity. Firstly, single-tree-based subgroup identification is sensitive to variations in the training sample—e.g., if the data are slightly altered, a completely different set of discovered subgroups might be found (namely, the model variance is high) (Breiman; 1996; Hastie et al.; 2009; Kuhn et al.; 2013). Secondly, it may fail to explore a vast number of potential subgroups (limited subgroup exploration)—e.g., the subgroups discovered are just the ones that can be represented by a single tree (Kuhn et al.; 2013; Spanbauer and Sparapani; 2021). To illustrate, consider a scenario in which two distinct factors are independently contributing to the heterogeneity in treatment effects. In such cases, a single tree algorithm may detect only one of these factors, failing to identify the second. In instances where both factors are identified, they are detected sub-optimally as an interaction between the two variables rather than as distinct drivers of the treatment heterogeneity.

To account for these shortcomings, we propose a novel Causal Rule Ensemble (CRE) method that uses multiple trees rather than a single tree to uncover, in a data-driven way, heterogeneity patterns in the treatment effect via decision rules. CRE provides (i) an interpretable representation of the HTE, (ii) via an extensive exploration of complex heterogeneity patterns, while (iii) guaranteeing high stability in the discovery. We also develop a general two-stage estimation approach for the conditional causal effects of the discovered subgroups and provide theoretical

guarantees.

CRE ensures interpretability providing a linear decomposition of the HTE in terms of *decision rules*. Interpretability is a non-mathematical concept, yet it is often defined as the degree to which a human can understand the cause of a decision (Kim et al.; 2016; Miller; 2019; Lakkaraju et al.; 2016; Wang and Rudin; 2022). *If-then* decision rules are highly interpretable as they resemble human decision-making processes. The discovery of these decision rules is obtained via an extensive exploration of complex heterogeneity patterns. In particular, CRE generates candidate decision rules from an ensemble of decision trees extracting heterogeneity in the treatment effect. Among these candidate decision rules, CRE proposes to extract only a stable set of decision rules characterizing the HTE by a rework of the stability selection algorithm (Meinshausen and Bühlmann; 2010). The stability of statistical results relative to "reasonable" perturbations to data and to the model used is critically important for reproducible research (Yu; 2013). Next to enhanced reproducibility, the stability selection algorithm allows also control for finite sample false discovery error.

Finally, CRE provides a two-stage estimation approach for the estimation of the coefficients in the discovered linear model of the conditional average treatment effect. In the first stage, pseudo-outcomes are produced using any of the available techniques for the estimation of HTE at the individual level. In the second stage, these pseudo-outcomes are regressed onto the discovered rules. Different subsamples are used for rules discovery and estimation in the prevention of overfitting (i.e., honest splitting Athey and Imbens (2016)). We provide theoretical results that guarantee the consistency and asymptotic normality of the estimated model coefficients. We also note that the proposed two-stage estimation is similar in spirit (even if the target estimands are different) to the Double Robust (DR) learner proposed by Kennedy (2020).

The remainder of the paper is organized as follows. In Chapter 2, we introduce the potential output framework and interpretable heterogeneous treatment effect discovery via decision rules. In Chapter 3, we introduce the proposed CRE methodology. In Chapter 4, we validate this methodology by simulated experiments, which are further extended in Appendix A. In Chapter 5, we propose to answer to EPA's call for environmental justice, applying the CRE method to assess vulnerability and resilience from air pollution exposure in the United States. Chapter 6 discusses the strengths and weaknesses of our proposed approach and areas of future research. CRE is implemented in an `R` package available on `CRAN`. Full documentation for the package can be found at `https://nsaph-software.github.io/CRE/`.

Chapter 2

Problem Formulation

2.1 Potential Outcomes Framework

Let $\mathcal{I}$ be a sample of N individuals. For an individual i, with $i = 1, ..., N$, let $\boldsymbol{X}_i \in \mathcal{X} \subseteq \mathbb{R}^P$ be the set of covariates characterizing i, $Z_i \in \{0, 1\}$ be i's observed (binary) treatment, and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ be i's observed outcome. Following the potential outcome framework (Rubin; 1974), for each individual, $i \in \mathcal{I}$, we define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes under treatment and control, respectively; and the Individual Treatment Effect (ITE):

$$\tau_i := Y_i(1) - Y_i(0). \tag{2.1}$$

The Average Treatment Effect is the expected value of the ITE:

$$\bar{\tau} := \mathbb{E}\left[Y_i(1) - Y_i(0)\right]. \tag{2.2}$$

The Conditional Average Treatment Effects (CATE) on $\boldsymbol{x}$ is the expected value of the ITE conditioning over a set of covariates $\boldsymbol{x}$:

$$\tau(\boldsymbol{x}) := \mathbb{E}\left[Y_i(1) - Y_i(0) | \boldsymbol{X}_i = \boldsymbol{x}\right]. \tag{2.3}$$

Both in ATE and CATE, the expected value is computed over the individuals i if $Y_i(Z)$ is deterministic. If $Y_i(Z)$ is random, the average is also computed over any other randomness. The CATE can be specified at different levels of *granularity*. For instance, at the highest level of granularity, one might want to estimate the ITE. At a lower level of granularity, one might want to estimate the average treatment effect for some *subgroups* of the population. This latter estimand can also be referred to as the Group Average Treatment Effect (GATE) (Jacob; 2019). Both the ITE and GATE are special cases of CATE. Throughout this paper, we will simply use the CATE rather than the GATE when referring to the estimated effects in the subgroups detected by the proposed algorithm.

Since only one potential outcome can be observed for each individual, the fundamental problem of causal inference (Holland; 1986), we need to rely on a few assumptions to identify the causal estimands of interest.

Assumption 1 (Stable Unit Treatment Value Assumption (SUTVA))**.**

$$\begin{aligned}(i).\ & Y_i(Z_i) = Y_i, \qquad \forall i \in \mathcal{I} \\ (ii).\ & Y_i(Z_i) = Y_i(Z_1, Z_2, \cdots, Z_i, \cdots, Z_N) \qquad \forall i \in \mathcal{I}.\end{aligned}$$

SUTVA enforces that for each individual i, i's outcome is simply a function of i's treatment. This is a combination of (i) consistency (no different versions of the treatment levels assigned to each unit) and (ii) no interference assumption (among the individuals) (Rubin; 1986).

Assumption 2 (Overlap)**.**

$$0 < e(\boldsymbol{x}) < 1 \ \forall\, \boldsymbol{x} \in \mathcal{X},$$

where $e(\boldsymbol{x}) = \mathbb{E}[Z_i = 1 | \mathbf{X}_i = \boldsymbol{x}]$ *is the propensity score (Rosenbaum and Rubin; 1983).*

The overlap assumption states that, for each unit, the probability of receiving either treatment is bounded away from zero and one.

Assumption 3 (Unconfoundedness)**.**

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i \mid \boldsymbol{X}_i, \qquad \forall i \in \mathcal{I}.$$

The unconfoundedness assumption states that, for each unit i, the two potential outcomes depend on $\boldsymbol{X}_i$, but are independent of Z_i conditioning on $\boldsymbol{X}_i$.

Under Assumptions 1, 2 and 3, the CATE can be identified (i.e., expressed in terms of statistical estimands) as:

$$\tau(\boldsymbol{x}) = \mathbb{E}\left[Y_i | \boldsymbol{X}_i = \boldsymbol{x}, Z_i = 1\right] - \mathbb{E}\left[Y_i | \boldsymbol{X}_i = \boldsymbol{x}, Z_i = 0\right]. \tag{2.4}$$

It is uncertain whether the set of covariates taken into consideration is adequate for establishing unconfoundedness. If it is not the case, the identification results do not hold. Sensitivity analysis provides a useful tool to investigate the impact of unmeasured confounding bias.

2.2 Interpretable Heterogeneity Discovery

Several algorithms have already been proposed for CATE estimation under the above-mentioned assumptions (i.e., Causal Forest, Bayesian Causal Forest, Inverse Probability Weighting, Stabilized Inverse Probability Weighting, Augmented Inverse Probability Weighting, S-Learner, T-Learner, X-Learner, DR-Learner). Although their powerful convergence property (e.g., double-robustness), interpreting the heterogeneity in the function with respect to the covariate space $\mathcal{X}$ can be anything but simple. We define here a new CATE characterization in terms of decision rules in order to enforce the interpretability of its heterogeneity (Lakkaraju et al.; 2016).

2.2.1 Decision Rules

A decision rule r is a general function on the covariates' space $\mathcal{X}$ characterizing a specific subgroup $S \subseteq \mathcal{X}$. We particularly focus on (interpretable) decision rules whose support (i.e., characterized subgroup) decompose as follows:

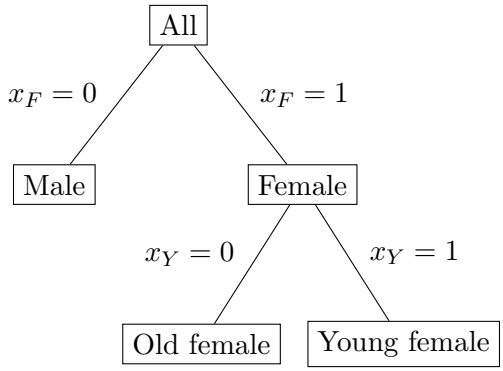
$$S = S_1 \times \cdots \times S_P \tag{2.5}$$

where $S_p \subseteq \mathbb{R}$ for all $p \in \{1, ..., P\}$. In formula:

$$\begin{aligned} r \colon \mathcal{X} &\to \{0, 1\} \\ \boldsymbol{x} &\mapsto r(\boldsymbol{x}) := \prod_{p=1}^{P} \mathbb{1}(x_p \in S_p) \end{aligned} \tag{2.6}$$

In the rest of the paper, we use *decision rule* referring to this specific definition.

In Figure 2.1, we report a dummy (binary) decision tree to provide a few examples of decision rules with two binary covariates x_F (for female) and x_Y (for young). Indeed, each node in a decision tree, combining the conditions of all its ancestors, agrees with the above definition of decision rule. For instance, the young female subgroup is expressed by $r_4(\boldsymbol{x}) = \mathbb{1}(x_F = 1) \cdot \mathbb{1}(x_Y = 1)$; and the male subgroup is expressed by $r_1(\boldsymbol{x}) = \mathbb{1}(x_F = 0)$, where the second term in the product is removed since equal to 1.



	Rule	Subgroup
r_1	$x_F = 0$	Male
r_2	$x_F = 1$	Female
r_3	$x_F = 1$ & $x_Y = 0$	Old female
r_4	$x_F = 1$ & $x_Y = 1$	Young female

Figure 2.1: An example decision tree (left) and the corresponding decision rules (right). Note that, in the Causal Rule Ensemble algorithm, for each decision tree, we consider only the decision rules corresponding to the terminal nodes, in this example r_1, r_3 and r_4.

2.2.2 Treatment Effect Linear Decomposition

We enforce interpretability in heterogeneous treatment effect estimation relying on the following assumption:

Assumption 4 (Treatment Effect Linear Decomposition)**.** *Let $\mathcal{R} = \{r_m\}_{m=1}^M$ a set of decision rules. For each individual $i \in \mathcal{I}$, the (individual) treatment effect can be linearly decomposed as follows:*

$$\tau_i = \bar{\tau} + \sum_{m=1}^{M} \alpha_m(\mathcal{R}) \cdot r_m(\boldsymbol{X}_i) + \nu_i \tag{2.7}$$

where ν is an unobserved and independent additive noise with $\mathbb{E}[\nu_i] = 0$ and $Var[\nu_i] = \sigma_i^2$, and $\alpha_m(\mathcal{R})$ are the model coefficients.

Assumption 4 states that the (individual) treatment effect can be decomposed in (i) average effect (ATE), (ii) additive contributions $\alpha_m(\mathcal{R})$ for all the activated rules (we say that a decision rule is activated if evaluated on its support) characterizing the heterogeneity, and (iii) noise (ν_i).

In terms of the conditional expectation, Equation 2.7 becomes:

$$\begin{aligned} \tau(\boldsymbol{x}) &= \mathbb{E}[\tau_i | X_i = \boldsymbol{x}] \\ &= \bar{\tau} + \sum_{m=1}^{M} \alpha_m(\mathcal{R}) \cdot r_m(\boldsymbol{x}) + \mathbb{E}[\nu_i] \\ &= \bar{\tau} + \sum_{m=1}^{M} \alpha_m(\mathcal{R}) \cdot r_m(\boldsymbol{x}) \end{aligned} \tag{2.8}$$

and it represents a step-wise approximation of the Conditional Average Treatment Effect $\tau(\boldsymbol{x})$. It follows the following result:

Proposition 1. *If the covariate space $\mathcal{X}$ is finite (very common in medical and health-related applications), then the Treatment Effect linear decomposition Assumption holds.*

[See proof in Appendix B]

By definition:

$$\begin{aligned} \alpha_m(\mathcal{R}) := \mathbb{E}[Y_i(1) - Y_i(0) | r_1(X_i) = \rho_1, ..., r_m(X_i) = 1, ..., r_M(X_i) = \rho_M] \\ - \mathbb{E}[Y_i(1) - Y_i(0) | r_1(X_i) = \rho_1, ..., r_m(X_i) = 0, ..., r_M(X_i) = \rho_M] \end{aligned} \tag{2.9}$$

where $\rho_1, ..., \rho_{m-1}, \rho_{m+1}, ..., \rho_M \in \{0, 1\}$. It represents the additive contribution to the Average Treatment Effect of the rule r_m, fixing all the values of all the others decision rules. For example, setting all the other decision rules to 0:

$$\alpha_m(\mathcal{R}) := \mathbb{E}_i \left[Y_i(1) - Y_i(0) | X_i \in \left\{ \boldsymbol{x} \in \mathcal{X} : \left[r_m(\boldsymbol{x}) \cdot \prod_{\substack{k=1 \\ k \neq m}}^{M} (1 - r_k(\boldsymbol{x})) \right] = 1 \right\} \right] - \bar{\tau} \tag{2.10}$$

In the rest of the thesis, we refer to the coefficient $\alpha_m(\mathcal{R})$ as the Additive Average Treatment Effect (AATE) of the m-th rule, and for simplicity of language, we remove its dependency on $\mathcal{R}$.

Assumption 4, can be finally rewritten in a matrix form as:

$$\boldsymbol{\tau} = \bar{\tau} + R\boldsymbol{\alpha} + \boldsymbol{\nu} \tag{2.11}$$

where $\boldsymbol{\tau} \in \mathbb{R}^N$ is the vector of (unknown) Individual Treatment Effects, $R \in \{0,1\}^{NxM}$ is the decision rules matrix which element $R_{i,j} = r_j(X_i)$ for all $i \in \{1,...,N\}$ and $j \in \{1,...,M\}$. For example:

$$R = \begin{array}{c} \\ \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{array} \begin{array}{c} \begin{array}{cccc} r_1(\cdot) & r_2(\cdot) & \dots & r_M(\cdot) \end{array} \\ \begin{pmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} \end{array}.$$

$\bar{\tau} \in \mathbb{R}$ is the Average Treatment Effect, $\boldsymbol{\alpha} \in \mathbb{R}^M$ is the vector of AATEs, and $\boldsymbol{\nu} \in \mathbb{R}^N$ is the vector of heteroscedastic and independent additive noise.

In this framework, interpretable heterogeneous discovery corresponds to finding a robust and minimal set of decision rules $\mathcal{R}$ satisfying Equation 2.7 (either properly piece-wise approximating if Assumption 4 doesn't hold), and heterogeneous causal effect inference corresponds to $\boldsymbol{\alpha}$ (AATEs) estimation.

Chapter 3

Causal Rule Ensemble

In this chapter, we introduce Causal Rule Ensemble (CRE), a new algorithm for interpretable inference of heterogeneous causal effects through the linear treatment effect decomposition by decision rules described in Section 2.2.

Assuming the setup described in Section 2.1, we first divide the observational dataset into two subsamples: a discovery subsample ($\mathcal{I}^d$) and an estimation subsample ($\mathcal{I}^e$). In the discovery step, we use $\mathcal{I}^d$ to select the set of decision rules $\hat{\mathcal{R}}$ robustly describing the heterogeneity in the treatment effect. In the estimation step, we use $\mathcal{I}^e$ to estimate the corresponding linear CATE decomposition in terms of decision rules. The idea of sample splitting is not new in statistics, and the earliest references can be traced back to Stone (1974) and Cox (1975). It is now commonly used also in the HTE literature to prevent overfitting and is referred to as *honest splitting* (Athey and Imbens; 2016; Lee et al.; 2021).

Algorithm 1 illustrates the main steps of the proposed methodology. In the rest of the chapter, we discuss in detail all the steps of the proposed procedure and its theoretical guarantees. In Section 3.1.1, we complement the discussion on ITE estimation, comparing 7 different state-of-the-art causal machine-learning algorithms for the task in Section 3.1.1, noting that the CRE method is agnostic to the choice of this estimator, both at the discovery and estimation steps.

Algorithm 1 Causal Rule Ensemble (CRE)

Inputs: covariates matrix X, (binary) treatment vector $\boldsymbol{z}$, and observed response vector $\boldsymbol{y}$.
Outputs: (i) a set of interpretable decision rules $\hat{\mathcal{R}} = \{\hat{r}_m\}_{m=1}^{M}$,
(ii) ATE $\hat{\bar{\tau}}$ and AATEs $\hat{\boldsymbol{\alpha}}$ estimates with confidence intervals.
Procedure:
$(X^d, \boldsymbol{z}^d, \boldsymbol{y}^d)$, $(X^e, \boldsymbol{z}^e, \boldsymbol{y}^e) \leftarrow$ `HonestSplitting`$(X, \boldsymbol{z}, \boldsymbol{y})$
i. Discovery
$\hat{\boldsymbol{\tau}}^d \leftarrow$ `EstimateITE`$(X^d, \boldsymbol{z}^d, \boldsymbol{y}^d)$ ▷ e.g. AIPW, CF, BCF, BART, S/T/X-Learner
$\hat{\mathcal{R}}' \leftarrow$ `GenerateRules`$(X^d, \hat{\boldsymbol{\tau}}^d)$ ▷ i.e., tree-ensemble method
$\hat{\mathcal{R}} \leftarrow$ `RulesSelection`$(\hat{\mathcal{R}}', X^d, \hat{\boldsymbol{\tau}}^d)$ ▷ Stability Selection
ii. Estimation
$\hat{\boldsymbol{\tau}}^e \leftarrow$ `EstimateITE`$(X^e, \boldsymbol{z}^e, \boldsymbol{y}^e)$ ▷ e.g. AIPW, CF, BCF, BART, S/T/X-Learner
$\hat{\boldsymbol{\alpha}} \leftarrow$ `EstimateAATE`$(\hat{\mathcal{R}}, X^e, \hat{\boldsymbol{\tau}}^e)$ ▷ Linear Decomposition

3.1 Discovery

Discovery is the first step of the Causal Rule Ensemble. It is itself divided into three steps, with the goal of discovering a stable set of decision rules approximating the Conditional Average Treatment Effect via Equation 2.11. First, the Individual Treatment Effect (pseudo-outcome) is estimated by any causal-machine learning algorithm. Secondly, an ensemble of trees algorithm (e.g., random forest) is trained to discover the heterogeneity in the estimated treatment effects (*fit-the-fit*), and a set of candidate decision rules is extracted. Finally, only a robust subset of the proposed decision rules is selected, based on the stability selection algorithm.

3.1.1 Individual Treatment Effect Estimation

For each individual $i \in \mathcal{I}^d$, we estimate the corresponding Individual Treatment Effect $\hat{\tau}_i^d$, relying on Assumption 1-3. Causal Rule Ensemble is model-agnostic with respect to the used ITE estimators, and any algorithm can be used, leading to different convergence properties.

We provide here a brief overview of seven ITE estimators, highlighting and comparing their strengths and weaknesses. An empirical comparison among the different methods is reported in Chapter 4.

T-Learner

The T-Learner (Hansotia and Rukstales; 2002) is a two-step approach where the conditional mean functions:

$$\mu_0(\mathbf{x}) := \mathbb{E}_i[Y_i(0)|\mathbf{X}_i = \mathbf{x}] \tag{3.1}$$
$$\mu_1(\mathbf{x}) := \mathbb{E}_i[Y_i(1)|\mathbf{X}_i = \mathbf{x}] \tag{3.2}$$

are estimated separately with any supervised learning algorithm (e.g., Generalized Linear Model, Tree Ensemble, Neural Network).

In the first step, the conditional mean under control is estimated by all the observations in the control group ($\hat{\mu_0}(\mathbf{x})$), and the conditional mean under treatment is estimated by all the observations in the treated group ($\hat{\mu_0}(\mathbf{x})$). Then, exploiting Equation 2.4, the Treatment Effect is estimated by:

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}) \tag{3.3}$$

S-Learner

The S-learner (Hill; 2011) treats the treatment variable Z_i as if it was just another covariate like those in the vector $\mathbf{X}_i$. Instead of having two models for the response as a function of the covariates, the S-learner has a single model for the response as a function of the covariates and the treatment:

$$\mu(\mathbf{x}, z) := \mathbb{E}_i[Y_i|\mathbf{X}_i = \mathbf{x}, Z_i = z] \tag{3.4}$$

In the first step, all the observations are used to estimate the response function above, $\hat{\mu}(\mathbf{x}, z)$, by any supervised learning algorithm (e.g., Generalized Linear Model, Tree Ensemble, Neural Network). Then, exploiting Equation 2.4, the Treatment Effect is estimated by:

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}(\mathbf{x}, 1) - \hat{\mu}(\mathbf{x}, 0) \tag{3.5}$$

X-Learner

The X-learner (Künzel et al.; 2019) is a three steps approach, estimating a treatment effect separately for the control and the treatment group. In the first step, like in S-Learner, the conditional mean functions:

$$\mu_0(\mathbf{x}) := \mathbb{E}_i[Y_i(0)|\mathbf{X}_i = \mathbf{x}] \tag{3.6}$$

$$\mu_1(\mathbf{x}) := \mathbb{E}_i[Y_i(1)|\mathbf{X}_i = \mathbf{x}] \tag{3.7}$$

are estimated separately by any supervised learning algorithm (e.g., Generalized Linear Model, Tree Ensemble, Neural Network).

The conditional mean under control is estimated by all the observations in the control group ($\hat{\mu_0}(\mathbf{x})$), and the conditional mean under treatment is estimated by all the observations in the treated group ($\hat{\mu_0}(\mathbf{x})$). Secondly, these two estimates are used for predicting the counterfactual outcomes.

$$\hat{\Psi}_1(\mathbf{X}_i) = Y_i - \hat{\mu_0}(\mathbf{X}_i) \qquad \text{for } i : Z_i = 1, \tag{3.8}$$

$$\hat{\Psi}_0(\mathbf{X}_i) = \hat{\mu_1}(\mathbf{X}_i) - Y_i \qquad \text{for } i : Z_i = 0. \tag{3.9}$$

Finally, these imputed effects are regressed individually on the covariates to obtain $\hat{\tau}_0(\mathbf{x})$ (the CATE for the control group) and $\hat{\tau}_1(\mathbf{x})$ (the CATE for the treatment group), and then combined by a weight function $g \in [0, 1]$:

$$\hat{\tau}(\mathbf{x}) = g(\mathbf{x})\hat{\tau}_0(\mathbf{x}) + [1 - g(\mathbf{x})]\hat{\tau}_1(\mathbf{x}) \tag{3.10}$$

A good choice for g is an estimate of the propensity score.

Augmented Inverse Probability Weighting (AIPW)

The Augmented Inverse Probability Weighting estimator (Robins et al.; 1994; Robins and Ritov; 1997) extends the observations balancing of Inverse Probability Weighting methods (Horvitz and Thompson; 1952) with conditional response estimation of the S-Learner, inheriting the benefits of both the approaches.

Firstly, the conditional mean response:

$$\mu(\mathbf{x}, z) := \mathbb{E}_i[Y_i|\mathbf{X}_i = \mathbf{x}, Z_i = z] \tag{3.11}$$

is estimated from all the observations by any supervised learning algorithm, $\hat{\mu}(\mathbf{x}, z)$ (first step S-Learner).

Then, the propensity score:

$$e(\mathbf{x}) := \mathbb{E}_i[Z_i|\boldsymbol{X}_i = \boldsymbol{x}] \tag{3.12}$$

is estimated from all the observations by any supervised learning algorithm, $\hat{e}(\mathbf{x})$.

Finally, the Treatment Effect is computed by:

$$\begin{aligned}\hat{\tau}(\mathbf{x}) = \frac{1}{N} \sum_{i \in \mathcal{I}(\mathbf{x})} \Bigg\{ &\left(\hat{\mu}(\mathbf{X}_i, 1) + \frac{Z_i(Y_i - \hat{\mu}(\mathbf{X}_i, 1))}{\hat{e}(\mathbf{X}_i)} \right) \\ &- \left(\hat{\mu}(\mathbf{X}_i, 0) + \frac{(1 - Z_i)(Y_i - \hat{\mu}(\mathbf{X}_i, 0))}{1 - \hat{e}(\mathbf{X}_i)} \right) \Bigg\}\end{aligned} \tag{3.13}$$

where $\mathcal{I}(\mathbf{x}) = \{i \in \mathcal{I} : \mathbf{X}_i = \mathbf{x}\}$. By construction, it is only required that one estimator among $\hat{e}$ and $\hat{\mu}$ is unbiased in order to get an unbiased estimate of the treatment effect.

Causal Forest (CF)

The Causal Forest is a method from Generalized Random Forests (Athey et al.; 2019). Similarly to Random Forest (Breiman; 2001), Causal Forest attempts to find neighborhoods in the covariate space (recursive partitioning). While a Random Forest is built from Decision Trees, a Causal Forest is built from Causal Trees (Athey and Imbens; 2016), which splitting criterion optimizes for finding splits associated with treatment effect heterogeneity. The goal is to find leaves where the treatment effect is constant but is different from other leaves.

A Causal Forest is simply the average of a large number of Causal Trees, where the trees differ due to subsampling. To create a Causal Forest from Causal Trees, it is necessary to estimate a weighting function and use the resulting weights to solve a local generalized method of moments (GMM) model to estimate the Conditional Average Treatment Effect. To deal with overfitting and biased estimations, Causal Forests, like Causal Rule Ensemble itself, rely on the honesty condition, whereby each training sample i is only used to decide where to place the split (discovery) or to estimate the within-leaf treatment effect (estimation), but not both. Honesty condition also leads to asymptotic normality. T he prediction of treatment effects is the difference in the average outcomes between the treated and the control observations of the estimating subsample in terminal leaves.

Causal Bayesian Additive Regression Trees (Causal BART)

The Bayesian Additive Regression Trees (BART) approach (Chipman et al.; 2010) combines gradient-boosting trees in a Bayesian framing using Markov Chain Monte Carlo (MCMC) sampling for back fitting (using additive and generalized additive models for posterior sampling). The Causal Bayesian Additive Regression Trees (Causal BART) approach (Hill; 2011) relies on such non-parametric Bayesian models to estimate treatment effects via S-Learner (see Equation 3.5). The method is specially designed to estimate the Treatment Effect from observational studies with small effect sizes and heterogeneous effects.

Bayesian Causal Forest (BCF)

The Bayesian Causal Forest (Hahn et al.; 2020) is a state-of-the-art model for causal inference that builds on Bayesian Additive Regression Trees. BCF combines Bayesian regularization with regression trees to provide a highly flexible response surface that, thanks to regularization from prior distributions, does not overfit the training data. In particular, BCF model the response as a function of the covariates and the treatment, adding the following priors:

$$\mu(\mathbf{x}, z) := \mathbb{E}_i[Y_i|\mathbf{X}_i = \mathbf{x}, Z_i = z] \tag{3.14}$$

$$= \mu(\mathbf{x}, \hat{e}(z)) + \tau(\mathbf{x})z \tag{3.15}$$

where $\hat{e}(\mathbf{x})$ is the estimated propensity score and the functions $\mu(\mathbf{x})$ and $\tau(\mathbf{x})$ are independent BART priors. The inclusion of the estimated propensity score can be seen as a covariate-dependent prior to controlling for confounding bias. The treatment effect is then computed as a S-Learner estimator by Equation 3.5.

3.1.2 Rules Generation

We detect the heterogeneity in the treatment effect by a $fit - the - fit$ approach. Once the Individual Treatment Effect estimates on the discovery sample are obtained, we fit these estimates ($\hat{\tau}_i^d$) from the observed covariates ($\hat{\boldsymbol{X}}_i^d$) by a tree-ensemble method (i.e., Random Forest (Breiman; 2001), Gradient Boosting Machine (Friedman; 2001)). In formula:

$$\hat{\tau}_i^d = \texttt{aggregate}(\mathcal{T}_1(\mathbf{X}_i^d), ..., \mathcal{T}_T(\mathbf{X}_i^d)) \qquad \forall i \in \mathcal{I}^d, \tag{3.16}$$

where $\mathcal{T}_t$ represents t-th distinct tree (its structure, its internal and terminal nodes) and $\texttt{aggregate}(\cdot)$ is an aggregating function (i.e., the mean for regression, the mode for classification).

Several variants of tree-ensemble methods can be considered, for example modifying the splitting criteria in the forest generation to enforce heterogeneity discovery. Once the forest is generated, we test, a posteriori, the predictive performance of each terminal node, comparing an error metric for the model with and without that leaf.

If the performances drop less than a certain threshold (t_{decay}), the node is discarded (pruned) as not significant for prediction (Deng; 2019).

We associate each leaf (terminal node) in the resulting forest with the corresponding decision rule obtained by combining the conditions of all its ancestors. Then, we collect all the distinct decision rules associated with the leaves in the generated forest as candidate decision rules for Conditional Average Treatment Effect linear decomposition ($\hat{\mathcal{R}}''$). Finally, we discard all the extreme or redundant decision rules based on the following two criteria:

i. **Extreme**: A (candidate) decision rule $r_m \in \hat{\mathcal{R}}''$ is said extreme if either too rare:

$$\sum_{i \in \mathcal{I}^d} r_m(\mathbf{X}_i) \leq t_{ext} N^d, \tag{3.17}$$

or too common:

$$\sum_{i \in \mathcal{I}^d} r_m(\mathbf{X}_i) \geq (1 - t_{ext}) N^d, \tag{3.18}$$

where t_{ext} is the threshold parameter defining the limit ratio, and $N^d = |\mathcal{I}^d|$.

ii. **Redundant**: A (candidate) decision rule $r_a \in \hat{\mathcal{R}}''$ is said redundant if exists at least another (candidate) decision rule $r_b \in \hat{\mathcal{R}}''$ such that their correlation is greater than a fixed threshold (t_{corr}).

The filtered set of candidate decision rules $\hat{\mathcal{R}}' \subseteq \hat{\mathcal{R}}''$ is then given in input to the rules selection step. By design, the maximal complexity of the candidate decision rules can be controlled a priori by the maximal length parameter (L) and the other stopping criteria in the tree-ensemble method. The filtering criteria also results very useful in practice to preliminary filter irrelevant decision rules and speed up the rules selection step. In Figure 3.1, we present an example of a tree-ensemble estimate to visualize the above-described procedure. The generated forest is composed of $T = 5$ trees and 12 total leaves. They correspond to 11 distinct decision rules ($x_2 \geq 0.6$ is double). Among these, we discard all the not significant rules (in light blue) based on the filtering described above. The remaining 8 leaves in dark blue are the candidate decision rules given in input to the rules selection.

By default, we propose to combine both Random Forest and Gradient Boosting Machine (GBM) for rules generation, following the parameters setting described by Friedman and Popescu (2008) and Nalenz and Villani (2018).

3.1.3 Rules Selection

The number of candidate decision rules M' extracted by the rules generation procedure grows exponentially with the maximal length and linearly with respect to the number of trees (before filtering). Although the filtering criteria already discard the not-significant rules, we are not controlling anywhere on the joint stability of these decision rules (i.e., given variations in the discovery set, these rules might be replaced with different ones). In order to enforce joint stability in the discovery, we

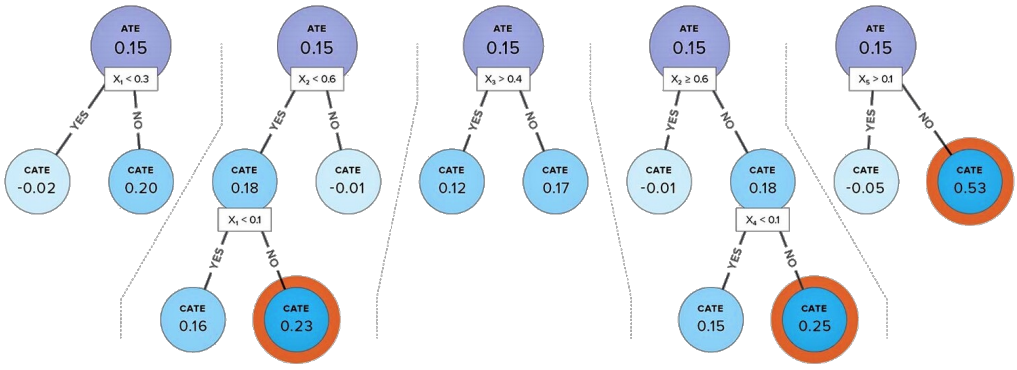


Figure 3.1: Visual representation of an example of (simple) rule generation and selection procedure. Among the 12 leaves, only the decision rules associated with the 8 leaves in dark blue are considered as candidate decision rules for the rule selection. The 3 leaves highlighted in red represent the decision rules that are finally selected by the stability selection procedure.

apply a stability selection regularization procedure to extract only the set of robust and predictive decision rules in terms of heterogeneity characterization.

We rely on the following penalized regression for rule selection:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\tau} - (\bar{\tau} + R\boldsymbol{\alpha})\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_l \tag{3.19}$$

where R is the decision rules matrix, λ is the regularization parameter, and $\|\cdot\|_l$ is a given norm. We select only the rules whose corresponding AATE estimation is different from zero. The least absolute shrinkage and selection operator (LASSO) estimator (Tibshirani; 1996) has been popular and widely used over the past two decades in order to solve the problem in (3.19) with $l = 1$.

The usefulness of this estimator among other penalization regression methods is demonstrated in various applications (see,e.g., Su et al.; 2016; Belloni et al.; 2016; Chernozhukov et al.; 2016, 2017). In practice, the true individual (and average) treatment effects are not observed, and we replace them with the corresponding estimates $\hat{\boldsymbol{\tau}}^d$ already computed. We also propose enforcing the discovery of shorter, thus less complex, decision rules, by weighting the columns of the rules matrix by the complexity (length) of the corresponding rule:

$$\tilde{R}_{i,m} = \frac{R_{i,m}}{\text{length}(r_m)} \qquad \forall i \in \mathcal{I}^d, \forall m \in \{1, ..., M'\}. \tag{3.20}$$

Indeed, the same heterogeneity in the treatment effect can be expressed by different sets of decision rules with different complexity. Weighting the columns of the rules matrix by the inverse of the length of the rules, we enforce the discovery of shorter, yet simpler (Bargagli Stoffi, Cevolani and Gnecco; 2022), characterizations.

When the data is high-dimensional, selecting λ can be challenging and generally unstable, depending on the unknown level noise of the observations. Stability Selection Meinshausen and Bühlmann (2010) extends this regularization, providing a procedure to stably extract the set of decision rules characterizing the heterogeneity of treatment effect, even controlling for the false discovery. Here, we propose to rework the stability selection procedure as follows.

Let $\mathcal{D}^d = (\hat{\boldsymbol{\tau}}^d, \hat{\mathcal{R}}')$ the discovery subsample individual treatment effect estimations and candidate decision rule estimates. For each value of $\lambda \in \Lambda$, we sample with replacement B different subsample $\mathcal{D}_{(b)}$ of $\mathcal{D}$ of size $\lfloor \frac{N^d}{2} \rfloor$ (bootstrapping). For each subsample $\mathcal{D}_{(b)}$ a selection algorithm, e.g., the model in (3.19), is run on $\boldsymbol{D}_{(b)}$ to obtain a selection set $\hat{\mathcal{R}}^{\lambda}_{(b)} \subseteq \hat{\mathcal{R}}'$ of decision rules. For each candidate decision rule $r_m \in \hat{\mathcal{R}}'$, let π^{λ}_m its probability of being selected by a certain selection algorithm characterized by :

$$\pi^{\lambda}_m = P\{r_m \in \hat{\mathcal{R}}^{\lambda}\}, \tag{3.21}$$

estimated by:

$$\hat{\pi}^{\lambda}_m = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\{r_m \in \hat{\mathcal{R}}^{\lambda}_{(b)}\}. \tag{3.22}$$

Given an estimate of the selection probabilities for each discovered rule and for each value of λ, we select a stable set of decision rules characterizing heterogeneity in the treatment effect, selecting all the rules which reached a selection probability bigger than a certain threshold π_{thr} for at least one value of λ:

$$\hat{\mathcal{R}} = \{r_m \ : \max_{\lambda \in \Lambda} \hat{\pi}^{\lambda}_m \geq \pi_{\text{thr}}\} \tag{3.23}$$

Meinshausen and Bühlmann (2010) discussed that the solution of stability selection is not sensitive to the initial regularization chosen, which is a desirable feature when selecting decision rules. The authors also state that the empirical results vary little for threshold values $\pi_{\text{thr}} \in (0.6, 0.9)$. The choice of the set of regularization parameters Λ is slightly more challenging than the choice of π_{thr}, but it can be explicitly controlled by an upper bound on the (allowed) per-family error rate (PFER). In addition to this, controlling for false discoveries is been shown to be of critical importance in the field of heterogeneous treatment effect discovery (Johnson et al.; 2022; Bargagli-Stoffi, De-Witte and Gnecco; 2022). By bounding the finite sample probability of making a Type I error, i.e., the probability of discovering a false positive decision rule, stability selection allows to control for the discovery of subgroups that are not likely to substantially contribute to the heterogeneity in the causal effects (we refer to Meinshausen and Bühlmann; 2010, for further details on the finite sample properties of the stability selection methodology). In practice, setting π_{thr} together with an upper bound on the PFER can be very data-dependent, easily leading to unaccepted combinations. Bodinier et al. (2021) proposes an automated selection of these parameters coming from the maximization of a stability measure.

In Figure 2.1, we presented an example of an ensemble of trees for rules selection. Among the eight decision rules associated with the terminal nodes in dark blue, the

three terminal nodes highlighted in red represent the selected decision rules by the stability selection procedure.

3.2 Estimation

Once a set of (robust) decision rules $\hat{\mathcal{R}}$ is estimated from the discovery subsample, we estimate the coefficient of the corresponding treatment effect linear decomposition on the inference subsample $\mathcal{I}^e$ through a two-stage estimation.

3.2.1 Individual Treatment Effect Estimation

First, for each individual $i \in \mathcal{I}^e$, we estimate the corresponding Individual Treatment Effect $\hat{\tau}_i^e$, relying on Assumption 1-3. Causal Rule Ensemble is model-agnostic with respect to the used ITE estimators, and any algorithm can be used, leading to different convergence properties. Let's observe that it is not required to use the same estimator selected in the discovery step, and certain methods could be preferred to others for one or the other task. For an overview of different successful causal machine learning algorithms for the task, see Section 3.1.1.

3.2.2 Additive Average Treatment Effect Estimation

Then, relying on Assumption 4:

$$\boldsymbol{\tau} = \bar{\tau} + R\boldsymbol{\alpha} + \boldsymbol{\nu}, \tag{3.24}$$

we replace the true individual (and average) treatment effects with estimates just computed from a consistent estimator:

$$\hat{\boldsymbol{\tau}} = \hat{\bar{\tau}} + R\boldsymbol{\alpha} + \boldsymbol{\nu}, \tag{3.25}$$

where R is the rules matrix over the estimation subsample using the decision rules $\hat{\mathcal{R}}$ retrieved in the discovery step, and we removed all the indexing referring to the estimation subsample for simplicity of language (but we still consider only this subsample in the whole Section, unless otherwise specified).

We can fit the model described in Equation 3.25 over $\mathcal{I}^e$, and compute an estimate of the Additive Average Treatment Effects (AATEs) by Ordinary Least Square:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T(\hat{\boldsymbol{\tau}} - \hat{\bar{\tau}}). \tag{3.26}$$

The two-stage Conditional Average Treatment Effect estimate is then given by:

$$\hat{\tau}^{\text{CRE}}(\mathbf{x}) = \hat{\bar{\tau}} + \sum_{m=1}^{M} \hat{\alpha}_m \cdot r_m(\mathbf{x}) \tag{3.27}$$

Equation 3.27 can be used both to compute the CRE estimate of the (individual) treatment effect over the whole population and also to characterize vulnerable and resilient subgroups based on the retrieved decision rules and the corresponding AATEs magnitude and sign.

Under a few additional assumptions, we prove here the consistency and asymptotic properties of the estimator $\hat{\boldsymbol{\alpha}}$.

Proposition 2. *Let $\hat{\boldsymbol{\tau}}$ a consistent estimator for $\boldsymbol{\tau}$ (i.e., AIPW). Under the Treatment Effect linear decomposition Assumption (Condition 1) and assuming $\mathbb{E}(\boldsymbol{R}^T\boldsymbol{R}) = Q$ is a positive definite matrix (Condition 2), the Additive Average Treatment Effects estimator $\hat{\boldsymbol{\alpha}} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T(\hat{\boldsymbol{\tau}} - \hat{\bar{\tau}})$ is a consistent estimator for $\boldsymbol{\alpha}$.*

[See proof in Appendix B]

Three additional assumptions are required to prove the asymptotic normality of the estimator $\hat{\boldsymbol{\alpha}}$:

3. $\mathbb{E}(\boldsymbol{R}_{ij}^4) < \infty \qquad \forall i \in \mathcal{I}^e$ and $\forall j \in \{1, ..., M\}$;
4. $\mathbb{E}(\nu_i^4) < \infty \qquad \forall i \in \mathcal{I}^e$;
5. $\mathbb{E}(\nu_i^2 \boldsymbol{R}_i^T \boldsymbol{R}_i) = \Omega \succ 0$ (positive definite) $\forall i \in \mathcal{I}^e$.

where $\boldsymbol{R}_i$ represents the i-th row of the rules matrix R. Since R is a binary matrix, Condition (3) is satisfied by design. The following theorem represents the asymptotic distribution of $\hat{\boldsymbol{\alpha}}$.

Proposition 3. *If Conditions (1)-(5) hold, then*

$$\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, V) \quad as \quad N \to \infty \tag{3.28}$$

where $V = Q^{-1}\Omega Q^{-1}$.

[See proof in Appendix B]

A variance-covariance matrix estimator $\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}$ can be obtained by the sandwich formula where:

$$\hat{Q} = \frac{R^T R}{N}, \tag{3.29}$$

$$\hat{\Omega} = \frac{\boldsymbol{\nu}^T R R^T \boldsymbol{\nu}}{N} \tag{3.30}$$

$$\hat{\boldsymbol{\nu}} = \hat{\boldsymbol{\tau}} - (\hat{\bar{\tau}} + R\boldsymbol{\alpha}) \tag{3.31}$$

This estimator is robust and often referred to as White's estimator (White; 1980). There are other approaches to obtain a heteroscedasticity consistent covariance matrix as discussed in Long and Ervin (2000). For small samples, Efron's estimator

(Efron; 1982), known as the HC3 estimator, can be considered alternatively. Also, if the variance σ_i^2 is known from the large sample properties of existing methods for obtaining $\hat{\boldsymbol{\tau}}_i^e$, then feasible generalized least squares estimators (Lewis and Linzer; 2005) can be considered. Given an estimate of the covariance-variance matrix, we also provide an (asymptotic) confidence interval for each Additive Average Treatments Effect α_m.

Chapter 4

Simulations

To assess the relative performance of CRE, we carried out three simulation studies. In the first simulation study, we assess the algorithm's performance in heterogeneity characterization, retrieving the correct effect modifiers and decision rules. We compare different variants of CRE using different ITE estimators, and we evaluate them with different magnitudes of the causal effect.

We benchmark their ATE and ITE estimation accuracy in the second simulation study, comparing them with the same 'standalone' treatment effect estimators. We also empirically verify the consistent estimation of the AATEs (Proposition 2). Both these studies rely completely on the assumptions of CATE identifiability (Assumptions 1-3) and linear decomposition (Assumption 4). Several data-generating processes are considered, varying the confounding mechanism, sample size and the number and complexity of the rules.

In this chapter, we report the main results of both the analyses; for a complete overview of the results with all the variant data-generating processes, see Appendix A. In the last simulation study, we test and discuss the estimation performances in case the treatment effect linear decomposition assumption doesn't hold.

4.1 Heterogeneity Discovery

Let $\mathcal{I}$ a sample of $N = 2,000$ individuals. For each individual $i \in \mathcal{I}$, let's define:

$$X_i^1, ..X_i^p \overset{\text{iid}}{\sim} Bernoulli(0.5) \qquad \text{and } \boldsymbol{X}_i = (X_i^1, ..., X_i^p) \tag{4.1}$$

$$Z_i \sim Bernoulli(\pi_i) \qquad \text{with } \pi_i = \frac{1}{1 + e^{+1 - X_i^1 + X_i^2 - X_i^3}} \tag{4.2}$$

where $\boldsymbol{X}_i$ is the vector of the $p = 10$ observed (binary) covariates of individual i, and Z_i represents its assigned (binary) treatment. Let us further define the potential outcomes:

$$Y_i(0) \sim \mathcal{N}(\mu_i^0, 1) \qquad \text{with } \mu_i^0 = X_i^1 + X_i^3 + X_i^4 + k \cdot 1_{\{x_1=1; x_2=0\}}(\boldsymbol{X}_i) \tag{4.3}$$

$$Y_i(1) \sim \mathcal{N}(\mu_i^1, 1) \qquad \text{with } \mu_i^1 = X_i^1 + X_i^3 + X_i^4 + k \cdot 1_{\{x_5=1; x_6=0\}}(\boldsymbol{X}_i) \tag{4.4}$$

where $k \in \mathbb{R}$ represents the magnitude of the causal effect. It follows that the (unobserved) Treatment Effect of individual i is equal to:

$$\tau_i = Y_i(1) - Y_i(0) = -k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{X}_i) + k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{X}_i) + \nu_i \tag{4.5}$$

where:

$$\nu_i \sim \mathcal{N}(0, 2) \tag{4.6}$$

is an additive zero-mean noise. The linear decomposition Assumption (Assumption 4) holds, with $M = 2$ decision rules, and:

$$\begin{aligned}
\bar{\tau} &= 0, && (4.7)\\
r_1(\boldsymbol{x}) &= 1_{\{x_1=1;x_2=0\}}(\boldsymbol{x}), && (4.8)\\
r_2(\boldsymbol{x}) &= 1_{\{x_5=1;x_6=0\}}(\boldsymbol{x}), && (4.9)\\
\alpha_1 &= \alpha_2 = k, && (4.10)\\
\tau(\boldsymbol{x}) &= -k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{x}) + k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{x}) && (4.11)\\
&= \sum_{m=1}^{M} \alpha_m \cdot r_m(\boldsymbol{x}).
\end{aligned}$$

We measure the CRE's capability in retrieving both the true effect modifiers (i.e., x_1, x_2, x_5, x_6) and the exact decision rules (i.e., r_1, r_2) varying the magnitude of the causal effect (i.e., k). Let S be the set of true effect modifiers (or decision rules) and $\hat{S}$ the set discovered by CRE. We define first:

$$\begin{aligned}
TP &= |\hat{\mathcal{S}} \cap \mathcal{S}|, && (4.12)\\
FP &= |\hat{\mathcal{S}} - \mathcal{S}|, && (4.13)\\
FN &= |\mathcal{S} - \hat{\mathcal{S}}|, && (4.14)
\end{aligned}$$

where TP represents the number of elements properly retrieved, FP represents the number of elements wrongly retrieved, and FN represents the number of right elements not retrieved. We can then define:

$$\begin{aligned}
Recall &= \frac{TP}{TP + FN}, && (4.15)\\
Precision &= \frac{TP}{TP + FP}, && (4.16)\\
F1 - score &= 2\frac{Recall \cdot Precision}{Recall + Precision}, && (4.17)
\end{aligned}$$

where *Recall* is the ratio of true elements properly retrieved (quantitative performance), *Precision* is the ratio of correct elements retrieved (qualitative performance), and the *F1-score* combines these 2 measures in a harmonic mean. We consider 7 variants of CRE with the ITE estimators described in Section 3.1.1: Augmented Inverse Probability Weighting (AIPW), Causal Forest (CF), Bayesian

Causal Forest (BCF), Causal Bayesian Additive Regression Trees (Causal BART), S-Learner, T-Learner and X-Learner. For each variant of CRE and causal effect size k, we compare the mean *Recall*, *Precision* and $F1 - score$ and their corresponding 95% confidence intervals over 250 Monte Carlo experiments. In Table 4.1, we summarize the Causal Rule Ensemble's method parameters and the hyperparameters used for this analysis, where 'XGboost' stands for the scalable end-to-end tree-boosting system algorithm by Chen and Guestrin (2016).

	Parameter	Value
Honest Splitting	Ratio	0.5
Discovery		
ITE Estimation	Propensity Score estimator ($\hat{e}$)	XGboost
	Outcome estimator ($\hat{\mu}$)	XGBoost
Rules Generation	N. Trees (Random Forest)	40
	N. Trees (GBM)	40
	Replace	True
	Max node (subgroup) size	20
	Max depth (L)	3
	Max number of nodes (per tree)	$2^3 = 8$
Filtering	t_{decay}	0.025
	t_{ext}	0.01
	t_{corr}	1
Rules Selection	Upper Bound PFER	$\frac{L}{k+1}$
	π_{thr}	0.8
Estimation		
ITE Estimation	Propensity Score estimator ($\hat{e}$)	XGboost
	Outcome estimator ($\hat{\mu}$)	XGBoost
CATE Estimation	$t_{p-\text{value}}$	0.05

Table 4.1: List of CRE method parameters and hyperparameters used for the simulations.

The results are reported in Figure 4.1.

As expected, both effect modifiers retrieval and decision rules retrieval metrics increase monotonically with respect to the causal effect size k. All the method variants perform similarly for effect modifiers retrieval, and all the variants reach almost perfect discovery by $k = 3$. Decision rule discovery is more challenging due to the larger hypothesis space. Indeed, in the case of binary covariates (current setting),

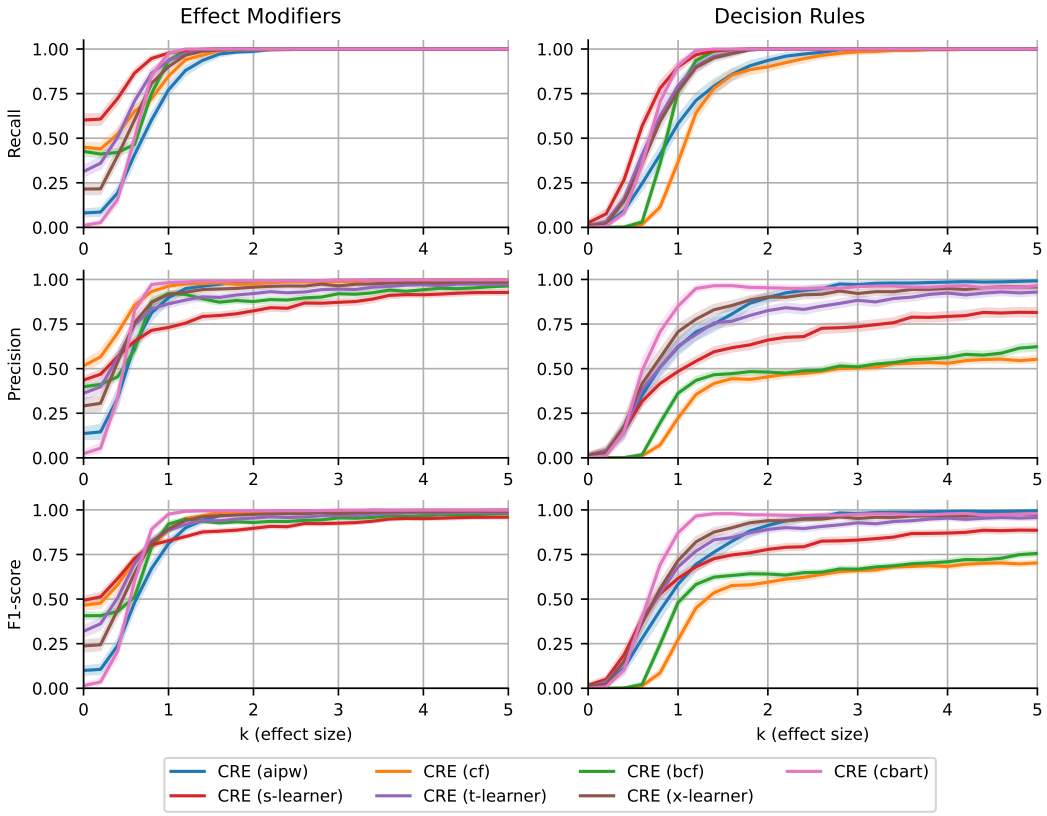


Figure 4.1: Simulation study for heterogeneity discovery results with 2 rules, linear confounders and 2,000 observations. Mean *Precision*, *Recall* and *F1-score* (lines) with the corresponding 95% confidence intervals (bands) over 250 Monte Carlo experiments are reported for each method and causal effect size k. For each CRE variant, the heterogeneity characterization discovery converges (with respect to effect size) to the true heterogeneity characterization.

the number of possible decision rules is equal to:

$$\sum_{l=1}^{L} \binom{p}{l} \cdot 2^l \tag{4.18}$$

growing exponentially with maximum rules' length L. and even more in the setting of discrete/continuous covariates, depending on the discretization criteria in the rules generation.

All the method variants perfectly retrieve the true decision rules ($Recall = 1$) by $k = 3$, but a few variants (i.e., CF, BCF, and S-Learner) also keep retrieving additional redundant rules ($Precision \ll 1$) even for $k > 3$. This drawback can be addressed by fine-tuning the strength in the Rules Selection step (e.g., increasing cutoff π_{thr}, or reducing the PFER in the Stability Selection), which we kept constant

for all the methods for a fair comparison. Causal Rule Ensemble based on Causal BART, AIPW, T-Learner, and X-Learner for ITE estimation leads to a more stable and precise rules discovery. In agreement with the literature (Hill; 2011), Causal Rule Ensemble based on Causal BART works better than any other variants for small effect sizes.

Consistent results are obtained in varying:

i. the sample size (1,000, 2,000, 5,000);

ii. the number of the decision rules (2, 4);

iii. the complexity of the decision rules (1, 2, 3);

iv. the type of confounding (none, linear, non-linear).

A comprehensive analysis of these additional simulations is reported in Appendix A.

4.2 Heterogeneous Treatment Effect Estimation

The simulation study on the heterogeneous effect estimation follows the same data-generating process described in Section 4.1. We fix the causal effect size $k = 5$, large enough to allow (almost) perfect discovery of all the CRE variants, and we compare the accuracy in both ITE and ATE estimation. In particular, for each method, we evaluate the mean and standard deviation Root-Mean-Square Error (RMSE) and Bias on ITE estimation over 250 Monte Carlo experiments per method, where:

$$\text{RMSE} = \sqrt{\frac{\sum_{i\in\mathcal{I}}(\tau_i - \hat{\tau}_i)^2}{N}}, \tag{4.19}$$

$$\text{Bias} = \frac{1}{N}\sum_{i\in\mathcal{I}}(\tau_i - \hat{\tau}_i) \xrightarrow{P} \bar{\tau} - \frac{1}{N}\sum_{i\in\mathcal{I}}\hat{\tau}_i, \tag{4.20}$$

and the (ITE) Bias is also representing the bias in the corresponding ATE estimation. We consider 7 variants of CRE with the following ITE estimators: Causal Forest (CF), Bayesian Causal Forest (BCF), Causal Bayesian Additive Regression Trees (Casual BART), S-Learner, T-Learner, and X-Learner; and we compare them with the same 'standalone' ITE estimators, which are commonly recognized among the best-performing algorithms for heterogeneous treatment effect estimation. We report the results in Table 4.2.

Overall, CRE outperforms the corresponding 'standalone' ITE estimators for both ITE and ATE estimation. In particular, CRE (AIPW), CRE (S-Learner), CRE (T-Learner), CRE (X-Learner), and CRE (Causal BART) significantly outperform the corresponding AIPW, S-Learner, T-Learner, X-Learner, Causal BART estimators for ITE estimation, and CRE (BCF) and BCF lead to comparable performances. CRE (CF) is the unique method worsening the performance of the corresponding CF estimator. Among the 'standalone' ITE estimators, CF and BCF are the two

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	**0.1336**	0.0603	0.0016	0.0891
CRE (CF)	0.6269	0.1404	0.0303	0.0957
CRE (BCF)	0.1482	0.0558	0.0047	0.0795
CRE (S-Learner)	0.1494	0.0589	0.0017	0.0860
CRE (T-Learner)	0.1495	0.0649	0.0011	0.0937
CRE (X-Learner)	0.1466	0.0659	0.0010	0.0937
CRE (Causal BART)	0.1398	0.0625	**0.0009**	0.0816
AIPW	2.0807	0.1919	0.0032	0.0562
CF	0.2955	0.0868	0.0051	0.0541
BCF	0.1339	0.0373	0.0042	0.0522
S-Learner	0.4837	0.0334	0.0020	0.0532
T-Learner	0.8065	0.0373	0.0035	0.0573
X-Learner	1.1878	0.0291	0.0035	0.0573
Causal BART	0.9925	0.0163	0.0020	0.0520

Table 4.2: Simulation study for (heterogeneous) treatment effect estimation, with $M = 2$ rules, linear confounder, 2,000 individuals and under CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

methods with the largest Bias (> 0.04). Our hypothesis is that their corresponding systematic errors in ITE estimation in the CRE discovery step lead to incorrect heterogeneity characterization, propagating the error at the estimation time.

Among the 'standalone' ITE estimators, AIPW is the one with the worst performance in ITE estimation. This result was somehow expected since the AIPW estimator was designed for (doubly-robust) (G)ATE estimation, while here we are extending it for ITE estimation (also known as pseudo-outcome in Doubly Robust literature (Kennedy; 2020)). Pseudo-outcome estimation by AIPW is unstable in presence of extreme propensity score estimation ($\hat{e}_i \approx 0$ or $\hat{e}_i \approx 1$) due to its dependence on $1/\hat{e}_i$ and $1/(1-\hat{e}_i)$. The presence of a few extreme predictions is robustly compensated in ATE estimation (see Bias= $0.0032 \approx 0$ in agreement with its double robustness) but leads to significantly high RMSE (which is more sensitive to outliers) for ITE estimation. On the other hand, CRE (AIPW) leads to the best performances for ITE estimation (RMSE= 0.1336) among all the considered methods, confirming that AIPW ITE estimation is still properly capturing the heterogeneity in the treatment effect. The best performances in ATE estimation (Bias= 0.0009) are obtained by CRE (Causal BART).

We then report in Figure 4.2 a boxplot on the AATEs ($\boldsymbol{\alpha}$) estimation bias:

$$\text{Bias}(r_m) = \alpha_m - \hat{\alpha}_m \qquad \forall r_m \in \mathcal{R}, \tag{4.21}$$

for a comparison among the different CRE variants over the same 250 Monte Carlo experiments. We remove CRE (CF) from this comparison due to its consistently incorrect rules discovery with redundant rules, leading to systematically biased AATEs estimations. Indeed, AATEs estimations strictly depend on the retrieved set of decision rules $\hat{\mathcal{R}}$.

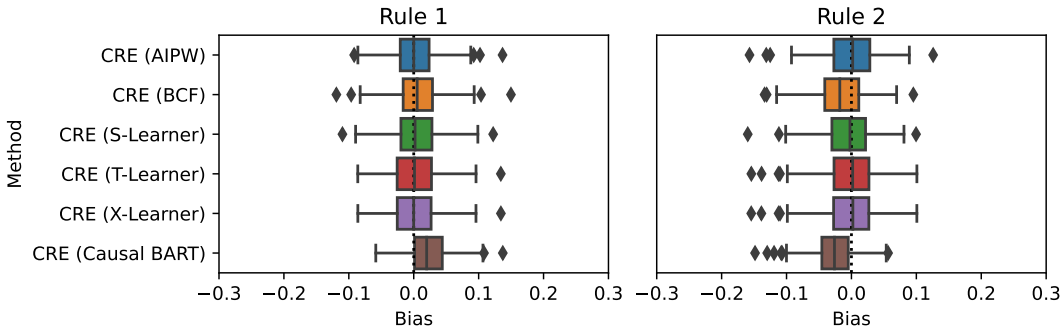


Figure 4.2: Simulation study for (heterogeneous) treatment effect estimation, with $M = 2$ rules, linear confounding and 2,000 individuals. For all the CRE variants, for each rule, the AATE's bias over 250 Monte Carlo experiments is reported in a boxplot.

The results for each method are obtained considering all of the 250 Monte Carlo experiments with $Recall = 1$ (retrieving both the true decision rules), also considering the cases of incorrect discovery ($Precision < 1$). As expected from Proposition 2, all the CRE variants lead to consistent AATEs estimation (median centered in 0), even without assuming perfect rules discovery. Consistent results for ITE, ATE, and AATE estimation are obtained in varying:

i. the sample size (1,000, 2,000, 5,000);

ii. the number of the decision rules (2, 4);

iii. the complexity of the decision rules (1, 2, 3);

iv. the type of confounding (none, linear, non-linear).

A comprehensive analysis of these additional simulations is reported in Appendix A.

4.3 Beyond Treatment Effect Linear Decomposition

All the encouraging results in the last two sections about CRE on both heterogeneity discovery and estimation are based on the Treatment Effect linear decomposition assumption. As discussed in Section 2, this assumption trivially holds if the covariates space $\mathcal{X}$ is finite. This is often the case in medical applications, where the majority of variables are binary or discretized to preserve interpretability.

However, there are also several scenarios where the heterogeneity in the Conditional Average Treatment Effect doesn't linearly decompose in terms of decision rules, and neither a step-wise approximation of the CATE is sufficient. In this section, we present a third simulation study, revisiting the data-generating process described in Section 4.1 breaking the Treatment Effect linear decomposition assumption.

Let $\mathcal{I}$ a sample of $N = 2000$ individuals. For each individual $i \in \mathcal{I}$, let's define:

$$X_i^1, ..X_i^p \overset{\text{iid}}{\sim} \mathcal{U}_{[0,1]} \qquad \text{and } \boldsymbol{X}_i = (X_i^1, ..., X_i^p) \tag{4.22}$$

$$Z_i \sim Bernoulli(\pi_i) \qquad \text{with } \pi_i = \frac{1}{1 + e^{+1 - X_i^1 + X_i^2 - X_i^3}} \tag{4.23}$$

where $\boldsymbol{X}_i$ is the vector of the $p = 10$ observed (binary) covariates of individual i, and Z_i represents its assigned treatment. Let us further define the potential outcomes:

$$Y_i(0) \sim \mathcal{N}(\mu_i^0, 1) \qquad \text{with } \mu_i^0 = X_i^1 + X_i^3 + X_i^4 + k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{X}_i) \tag{4.24}$$

$$Y_i(1) \sim \mathcal{N}(\mu_i^1, 1) \qquad \text{with } \mu_i^1 = X_i^1 + X_i^3 + k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{X}_i) \tag{4.25}$$

where $k \in \mathbb{R}$ represents the magnitude of the causal effect. It follows that the (unobserved) Treatment Effect of individual i is equal to:

$$\tau_i = Y_i(1) - Y_i(0) = -k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{X}_i) + k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{X}_i) + k \cdot X_i^4 + \nu_i \tag{4.26}$$

where:

$$\nu_i \sim \mathcal{N}(0, 2) \tag{4.27}$$

is an additive zero-mean noise, and

$$\tau(\boldsymbol{x}) = k \cdot (1_{\{x_5=1;x_6=0\}}(\boldsymbol{x}) - 1_{\{x_1=1;x_2=0\}}(\boldsymbol{x}) + x_4) \tag{4.28}$$

is the corresponding Conditional Average Treatment Effect.

By definition, Equation 4.28 doesn't satisfy Assumption 4, and the CRE consistency results described in Chapter 3 don't hold. In this framework, the decision rules discovery task is not even defined, but we still propose to evaluate the estimation performances, as proposed in Section 4.2. Causal Rule Ensemble cannot capture the continuous dependence of the CATE on the fourth covariate, but it can still try to approximate it by a step-wise function.

We consider the usual 7 variants of CRE with the corresponding 7 'standalone' ITE estimators, and for each method, we report the mean and standard deviation RMSE and Bias in ITE estimation over 250 Monte Carlo experiments. We consider the same method parameters and hyperparameters reported in Table 4.1, with exception of two discovery parameters we modify due to the new covariate space. In particular, we fix now: $t_{corr} = 0.7$ and $PFER = 0.5$. We report the results in Table 4.3.

Before discussing the results, it is important to mention that they are not comparable with the analysis in the previous Section, not just because of the modified

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	1.5438	0.2291	-0.1430	0.2705
CRE (CF)	3.7773	0.6114	1.2915	1.4688
CRE (BCF)	2.1611	0.7145	-0.5800	1.0673
CRE (S-Learner)	1.5496	0.2412	-0.1105	0.2772
CRE (T-Learner)	1.4785	0.2723	-0.1939	0.2970
CRE (X-Learner)	1.4839	0.2671	-0.1947	0.2996
CRE (Causal BART)	2.7969	0.1737	-1.4769	0.2866
AIPW	0.5907	0.0365	-0.0305	0.0497
CF	3.3537	0.3618	0.6974	1.1117
BCF	0.9507	0.3409	0.0834	0.4185
S-Learner	**0.5105**	0.0439	-0.0311	0.0494
T-Learner	1.1149	0.0257	**-0.0133**	0.0487
X-Learner	1.1626	0.0258	**-0.0133**	0.0487
Causal BART	1.9984	0.2841	-0.5242	0.6673

Table 4.3: Simulation study for (heterogeneous) treatment effect estimation, with $M = 2$ rules, linear confounder, 2,000 individuals, and violating CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

CATE function, but also because we are now considering different covariates space, and distribution (uniform).

Overall the 'standalone' ITE estimators outperform the corresponding CRE variants in both ITE and ATE estimation. In particular, the S-Learner gets the best performances in ITE estimation (RMSE= 0.5105) and both T-Learner and X-Learner get the best performances in ATE estimation (Bias=-0.0133). Causal Forest and Causal BART provide a biased estimation, which drastically propagates in the corresponding CRE variants. Similar results, but milder, are observed for Bayesian Causal Forest. Empirical evidence about the Causal Forest slow convergence rate was already observed in papers analyzing the empirical results of these methods (Hahn et al.; 2019; Wendling et al.; 2018). All the remaining CRE variants lead to the (almost) comparable ITE estimation without systematic biases.

With this final simulation study, we proposed to test the CRE estimation performances beyond its main assumption of Treatment Effect linear decomposition. Despite a few CRE variants (CF, BCF, Causal BART) wrongly propagating the systematic errors of the corresponding ITE estimators, several CRE variants (AIPW, S-Learner, T-Learner, X-Learner) still reasonably approximate the heterogeneity in the treatment effect and preserve interpretability. However, classic ITE estimators outperform CRE in this setting. The treatment effect linear decomposition in terms of decision rules is not flexible enough to capture complex heterogeneity. There is

no free lunch, and the price for more interpretability is paid in the form of a reduced model flexibility (degrees of freedom).

Chapter 5

Heterogeneous Effects of Air Pollution Exposure on Mortality

The literature indicates that long-term exposure to lower levels of $PM_{2.5}$ is associated with a significant decrease in mortality (see, e.g., Dockery et al.; 1993; Di et al.; 2017; Liu et al.; 2019; Pappin et al.; 2019; Wu, Braun, Schwartz, Kioumourtzoglou and Dominici; 2020). While previous research has contributed to understanding the average treatment effect of long-term $PM_{2.5}$ exposure, it has largely neglected to explore potential heterogeneity in the causal effects. However, it is essential to investigate how the causal effect may differ across different groups of individuals in health studies to develop more effective health policies.

In this context, our focus is on identifying vulnerability or resilience in the causal effects with respect to the average effect of exposure to air pollution on mortality. In particular, we examine the heterogeneous effects of long-term $PM_{2.5}$ exposure to high levels of air pollution among individuals aged 65 and above who were enrolled in Medicare in the years 2010-2016. By utilizing our CRE methodology, we showcase how our approach can identify distinct groups, estimate the heterogeneity in the effects of long-term $PM_{2.5}$ exposure on mortality, and identify the social-economical characteristics that distinguish the different heterogeneous subgroups.

5.1 Data

We collected data from 35,331,290 Medicare beneficiaries across the contiguous U.S. For each beneficiary, we have information on age, sex, race (specifically categorized as Hispanic, black, white, and other race), eligibility for Medicaid (this variable is a proxy of low social-economic status), and whether or not they died in the 5 follow-up years (2012-2016). We integrated these data with average $PM_{2.5}$ levels in 2010 and 2011. Figure 5.1 depicts the average levels $PM_{2.5}$ for the biennium 2010-2011 across the contiguous U.S.

Furthermore, we integrated census variables at the ZIP code level and county-level variables. At the ZIP code level, we have information on the average household

income, average home value, the proportion of residents in poverty, the proportion of residents without a high school diploma, the population density, the proportion of residents who own their houses and the proportion of the black and Hispanic population. Furthermore, we considered meteorological data such as the average maximum daily temperatures and relative humidity during summer (June to September) and winter (December to February). At the county level, we considered variables on the average body mass index and the average smoking rate.

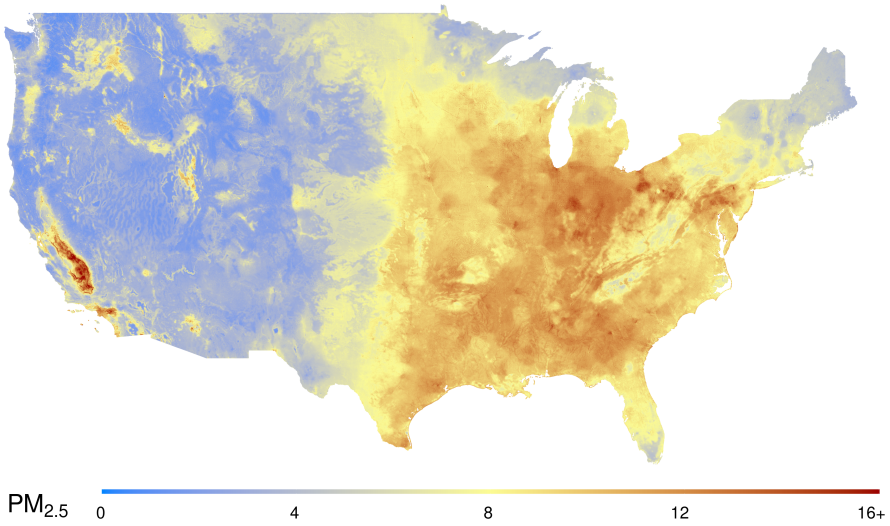


Figure 5.1: Average levels $PM_{2.5}$ for the biennium 2010-2011 in the contiguous U.S.

5.2 Study Design

We define the treatment variable as $Z = 1$ if the average $PM_{2.5}$ in 2010 and 2011 is above the threshold of $12\mu g/m^3$ and $Z = 0$ otherwise. The choice of $12\mu g/m^3$ as a threshold aligns with the current National Ambient Air Quality Standard (NAAQS) established by the Environmental Protection Agency (EPA). All the covariates at the individual level (except for age) are already binary, and we keep them as such. In order to enforce interpretability, we also binarize all the other covariates—i.e., age, zip code level variables, and county level variables—using the median as a threshold. Different discretization criteria and thresholds can be considered for more detailed heterogeneity characterization (even not discretization at all). For each individual, the observed factual outcome Y is equal to 1 if the person died in the five follow-up years (2012-2016) and 0 otherwise.

We investigate the heterogeneity in the effects of air pollution on mortality in the four different geographical regions defined by the U.S. Census Bureau separately (see Figure 5.2). It is crucial to investigate the effects of air pollution on health across

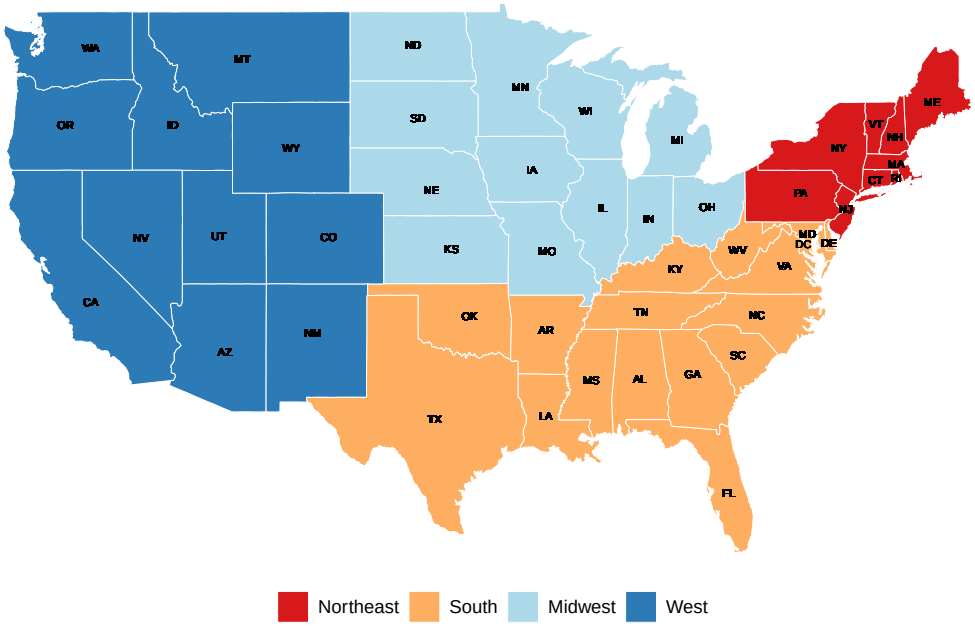


Figure 5.2: Map of the four census geographical regions for the contiguous U.S.

the different geographical regions of the U.S. for several reasons.

Firstly, the U.S. is a vast country with a diverse climate and environmental conditions, leading to substantial differences in air quality and exposure to air pollution across different regions (see Figure 5.1). As highlighted by Baxter et al. (2013), it is utterly important to assess the differential risks of air pollution on mortality at a regional level as they could be suggestive heterogeneous health responses driven by variations in the $PM_{2.5}$ composition and the concentration of gaseous pollutants.

Secondly, people living in different regions may have different susceptibilities to the health effects of air pollution due to various factors such as genetics, lifestyle, and pre-existing health conditions (Dominici et al.; 2006; Kloog et al.; 2013; Zanobetti et al.; 2009). Therefore, understanding the heterogeneity in the health effects of air pollution across different regions can help identify vulnerable populations and design targeted interventions to mitigate the adverse health effects.

Thirdly, Dedoussi et al. (2020) found that 41 to 53% of air-quality-related premature mortality resulting from a state's emissions occurs outside that state. Hence, regional-level analyses—factoring in the potential out-of-state sources of emission—directly map into region-wide policies that may be more effective in reducing the mortality burdens from exposure to air pollution.

All this considered, investigating the effects of air pollution on health across different regions of the U.S. is essential for identifying the specific risks associated with exposure to pollutants, understanding the heterogeneity in the health effects across different populations, and informing public health policies and interventions.

The list of CRE methods and hyper-parameters used in these analyses is reported in Table 5.1.

	Parameter	Value
Honest Splitting	Ratio	0.5
Discovery		
ITE Estimation	Estimator	X-Learner
	Outcome estimator ($\hat{\mu}$)	XGBoost
Rules Generation	N. Trees (Random Forest)	100
	N. Trees (GBM)	100
	Replace	True
	Max node (subgroup) size	20
	Max depth (L)	2
	Max number of nodes (per tree)	4
Filtering	t_{decay}	0.002
	t_{ext}	0.005
	t_{corr}	1
Rules Selection	Upper Bound PFER	1
	π_{thr}	0.8
Estimation		
ITE Estimation	Estimator	X-Learner
	Outcome estimator ($\hat{\mu}$)	XGBoost
CATE Estimation	$t_{p-\text{value}}$	0.05

Table 5.1: List of CRE method parameters and hyper-parameters used for the discovery and estimation of HTE of air pollution exposure on mortality.

5.3 Results

Consistently with the literature, we found that being exposed to higher levels of air pollution with respect to the NAAQS of $12\mu g/m^3$ leads to an increase in mortality in each of the four regions of the contiguous U.S. considered. The greatest increase was found in the Northeast, where individuals exposed, in the biennium 2010-2011, to levels of $PM_{2.5}$ higher than the NAAQS were found to be 16.2% more likely to die in the five following years, compared to those exposed to levels lower than the NAAQS. We found 14.9%, 7.1%, and 2.3% increases in mortality in the West, Midwest, and South, respectively.

Using CRE, next to the average treatment effects, we were also able to discover notable heterogeneities with respect to the average treatment effect in each of the

four regions. Figure 5.3 depicts the results of our analyses.

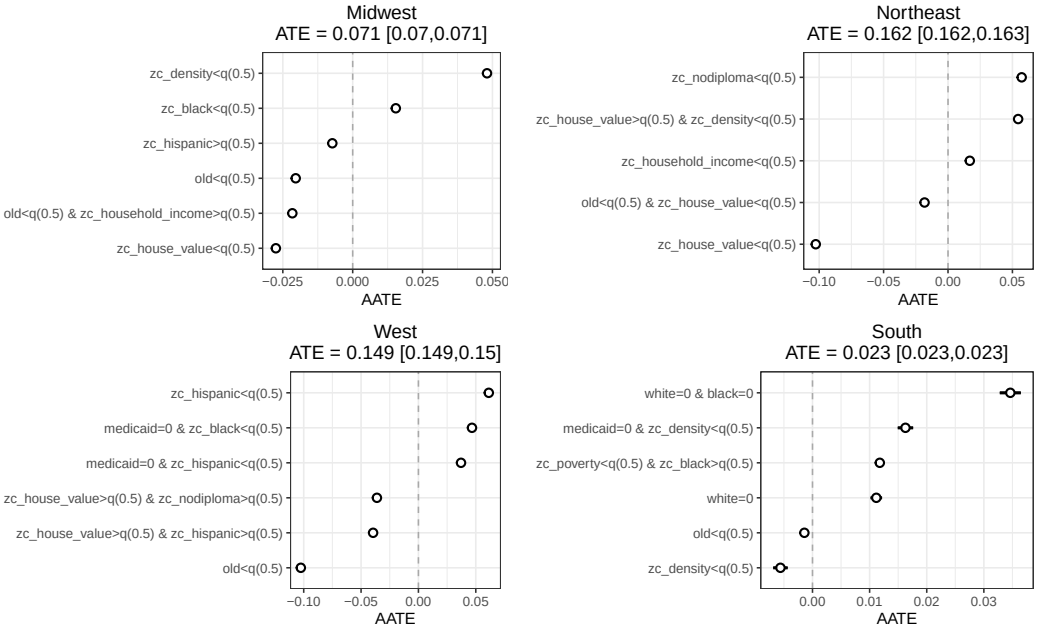


Figure 5.3: Results obtained from CRE for the discovery and estimation of HTE of air pollution exposure on mortality. For each U.S. Census geographical region, we report the ATE and the various AATEs for the discovered decision rules (with the corresponding 95% confidence interval). The prefix `zc` in the variable names stands for zip-code level variable, and `county` stands for county level. All the other variables are at the individual level. The threshold $q(\alpha)$ stands for the α-quantile of that variable in that region—e.g., $q(0.5)$ is the median.

For each of the four regions, we found both positive and negative additive average treatment effects. When an AATE is positive, it means that, under that decision rule, it was estimated a treatment effect greater than without (fixing all the others' contributions). Conversely, when an AATE is lower than zero, it means that, under that decision rule, it was estimated a treatment effect smaller than without. Notably, for all the regions, the effects of the AATEs never completely offset the ATE, indicating an overall detrimental impact of exposure to higher levels of air pollution on health both at a population and a subgroup level.

We find high fragmentation in the heterogeneity characterizing the groups where exposure to higher levels of $PM_{2.5}$ increases the mortality rate—i.e., a positive AATE. A clear trend is provided by an increased risk for individuals living in low-density areas (i.e., `zc_density`$< q(0.5)$) in the Northeast, Midwest, and South. Low-density areas, such as rural areas, can be characterized by decreased access to medical care, and this, paired with exposure to higher levels of air pollution, may be a possible driver of higher vulnerability. The second notable heterogeneity driver comes from individuals with a low socio-economic status (i.e., `medicaid`$= 0$) or living in low-

income areas (i.e., `zc_household_income`$< q(0.5)$) in the West and South being more vulnerable. The last one is for individuals living in black (i.e., `zc_black`$> q(0.5)$), yet less poor (i.e., `zc_poverty`$< q(0.5)$), areas in the South being more vulnerable. Other vulnerabilities are found in individuals living in areas with a minor black population (in the Midwest), a smaller rate of people without a diploma (in the Northeast), a minor Hispanic population (in the West), and people neither white nor black (in the South).

In juxtaposition, the groups where exposure to higher levels of $PM_{2.5}$ decreases the ATE on the mortality rate are mainly composed of a population with fewer old individuals (in particular in the Midwest, West, and South). Consistently, with what was found for the vulnerability, individuals living in areas with a higher density of minority groups (i.e., Hispanic in the Midwest and West) were found to be at lower risk. While surprising this has already been documented in the literature (Liu et al.; 2021; Jbaily et al.; 2022), and it finds a possible explanation in potential survival bias (see, e.g., Mayeda et al.; 2018; Shaw et al.; 2021). Survival bias happens in cohort studies that start at a late stage in people's lives, 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 lower mortality risk, with respect to the average risk, even when exposed to higher levels of pollutants (Liu et al.; 2021). This is likely to be the case for these groups. We know, from previous literature, that Hispanic and black populations 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 the most fragile individual dying before entering the Medicare cohort (see, e.g., Pope III et al.; 2019; Liu et al.; 2021; Jbaily et al.; 2022).

To conclude, CRE was able to identify the key factors in the population characteristics that explain different degrees of vulnerability to exposure to air pollution. This application demonstrates the ability of CRE to retrieve non-trivial yet interpretable characterization of the heterogeneous subgroups.

Chapter 6

Conclusion

In this thesis, we introduced a new method for interpretable discovery and estimation of heterogeneous treatment effects. The proposed CRE methodology accommodates the well-known shortcomings in the flexibility of (individual) causal trees and interpretability of ensembles of causal trees (i.e., Causal Forest) relying on the linear decomposition of the treatment effect in terms of decision rules. By design (see Stability Selection), its heterogeneity discovery is stable to sample-to-sample variations, and under the assumptions of identifiability and linear decomposition, CRE leads to consistent estimates.

The decision rules characterizing the heterogeneity are estimated by a $fit - the - fit$ procedure, relying on a preliminary individual treatment effect (pseudo-outcome) estimation by any existing treatment effect estimator. Similarly, the final linear model relies on an analogous preliminary individual treatment effect (pseudo-outcome) estimation. Therefore, the CRE method can be thought of as a *refinement* process of the outputs produced by existing methods. Different properties characterize different estimators, and the performance of CRE varies with respect to them. If an estimator properly estimates the heterogeneity in the treatment effect, the CRE method discovers the underlying treatment effect structure with higher probability and represents this structure in an easy-to-interpret form.

The maximal number and complexity of the rules can be set by researchers or practitioners. Indeed, a few simple (i.e., not lengthy) rules are utterly important for public policy implications, where policy guidelines need to be as simple and as general as possible. However, when it comes to precision medicine, discovering a possibly lengthy rule that is specific to a patient could be of interest. Also, the choice of how many causal rules to discover in the discovery step may depend on the questions that practitioners want to answer. For example, policymakers generally want to discover a short list of risk factors. A few important subgroups defined by the risk factors are usually easy-to-understand and further foster focused discussions about the assessments of potential risks and benefits of policy actions. Due to the restriction of resources, public health can be promoted efficiently when prioritized subgroups are available. Conversely, in precision medicine, a comparatively larger set of decision

rules can be chosen. Indeed, an important goal is to identify patient subgroups that respond to treatment at a much higher (or lower) rate than the average (Loh et al.; 2019). Also, identifying a subgroup that must avoid the treatment due to its excessive side effects can be valuable information. However, discovering only a few subgroups is likely to miss this extreme subgroup.

From simulations, we exhibited that CRE has competitive performances both in the discovery and estimation of the treatment effect. We showed first, as a proof of concept, that under the linear decomposition assumption and significant causal effect, CRE perfectly retrieves the correct treatment effect decomposition. Then we compared CRE performances in estimation with several of the most successful causal machine-learning methods for heterogeneous treatment effect estimation. Under linear decomposition assumption, CRE significantly outperforms all the other estimators, correctly capturing the heterogeneity in the treatment effect in terms of decision rules. Opposite results are observed in absence of this assumption, where the heterogeneity is too complex to be captured by decision rules, and more flexible but less interpretable methods are required. All the simulation studies are repeated with different data-generating processes, leading to consistent results; and numerous seeds to enforce reproducibility.

The use of CRE allowed for the identification of crucial factors in characteristics that help explain the varying levels of susceptibility to air pollution exposure in the elderly population in the U.S. By employing CRE, a non-trivial and comprehensible characterization of distinct subgroups has been retrieved. This application not only showcased the efficacy of CRE in deciphering complex patterns in data but also highlighted the significance of understanding the heterogeneous nature of populations in relation to environmental hazards. Such insights should be paired with extensive analyses of vulnerability to air pollution in the younger population. The results of our analyses, found indeed, that there may be room for possible survival bias when analyzing the vulnerability to air pollution due to structural differences in air pollution exposure across the U.S.

A number of extensions of the CRE method can be possible. CRE deals with the exploration of heterogeneous treatment effects in the simple case of a binary treatment in a cross-sectional setting. It would be of great interest to extend the CRE setting to continuous treatment effects and time-series studies as these dimensions might be critically important for a number of applications in social and health sciences. Furthermore, starting from CRE to develop interpretable methods for optimal policies or targeted treatment assignment is also crucial. Optimal policies involve assigning treatments to individuals in a manner that maximizes the desired outcome while taking into account their unique characteristics. Such an approach could lead to more effective and efficient treatment outcomes, reduce unnecessary treatment and improve patient outcomes. Additionally, by targeting treatments to the appropriate individuals, optimal policies can help reduce health disparities and ensure that interventions are more equitably distributed. Therefore, developing methods for optimal policies is an important future extension of CRE in the direction of moving one step closer to achieving personalized and effective healthcare.

Appendix A

Additional Simulations

In this Appendix, we present a more extensive analysis of the heterogeneity discovery and treatment effect estimation simulation studies under different variants of the data-generating process, varying the sample size, the number of decision rules, the complexity of the decision rules and the type of confounding. In particular, for both the simulation studies, we consider the following five variants to the data generating process described in Section 4.1 (where all the definitions are kept equal if not otherwise specified):

i. **Large Sample**: $N = 5,000$ individuals, $M = 2$ rules (r_1, r_2), linear confounding;

ii. **Small Sample**: $N = 1,000$ individuals, $M = 2$ rules (r_1, r_2), linear confounding;

iii. **More Rules**: $N = 2,000$ individuals, $M = 4$ rules (r_1, r_2, r_3, r_4), linear confounding, where:

$$\begin{aligned}\mu_i^0 &= X_i^1 + X_i^3 + X_i^4 + k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{X}_i) + \frac{k}{2} \cdot 1_{\{x_4=0\}}(\boldsymbol{X}_i) \\ &= X_i^1 + X_i^3 + X_i^4 + k \cdot r_1(\boldsymbol{X}_i) + \frac{k}{2} \cdot r_3(\boldsymbol{X}_i),\end{aligned} \tag{A.1}$$

$$\begin{aligned}\mu_i^1 &= X_i^1 + X_i^3 + X_i^4 + k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{X}_i) + 2k \cdot 1_{\{x_5=0;x_7=1;x_8=0\}}(\boldsymbol{X}_i) \\ &= X_i^1 + X_i^3 + X_i^4 + k \cdot r_2(\boldsymbol{X}_i) + 2k \cdot r_4(\boldsymbol{X}_i),\end{aligned} \tag{A.2}$$

and then:

$$\tau(\boldsymbol{x}) = -k \cdot r_1(\boldsymbol{x}) + k \cdot r_2(\boldsymbol{x}) - \frac{k}{2} \cdot r_3(\boldsymbol{x}) + 2k \cdot r_4(\boldsymbol{x}); \tag{A.3}$$

iv. (Pseudo) **Randomized Controlled Trial**: $N = 2,000$ individuals, $M = 2$ rules (r_1, r_2), only confounding by decision rules, i.e.:

$$\mu_i^0 = k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{X}_i) = k \cdot r_1(\boldsymbol{X}_i), \tag{A.4}$$

$$\mu_i^1 = k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{X}_i) = k \cdot r_2(\boldsymbol{X}_i); \tag{A.5}$$

and then (as the original data-generating process):

$$\tau(\boldsymbol{x}) = -k \cdot r_1(\boldsymbol{x}) + k \cdot r_2(\boldsymbol{x}); \tag{A.6}$$

v. **Non-Linear Confounding**: $N = 2,000$ individuals, $M = 2$ rules (r_1, r_2), non-linear confounding, i.e.:

$$\mu_i^0 = X_i^1 + \sin(x_i^3 \cdot x_i^4) + k \cdot 1_{\{x_1=1;x_2=0\}}(\boldsymbol{X}_i) = k \cdot r_1(\boldsymbol{X}_i), \tag{A.7}$$
$$\mu_i^1 = X_i^1 + \sin(x_i^3 \cdot x_i^4) + k \cdot 1_{\{x_5=1;x_6=0\}}(\boldsymbol{X}_i) = k \cdot r_2(\boldsymbol{X}_i). \tag{A.8}$$

and then (as the original data-generating process):

$$\tau(\boldsymbol{x}) = -k \cdot r_1(\boldsymbol{x}) + k \cdot r_2(\boldsymbol{x}); \tag{A.9}$$

Each of the described data-generating process vary the original design for a specific characteristic, which we desire to test our methodology on. In the Section A.1 we report and discuss the results of the heterogeneity discovery simulation study over these different data generating processes, and in the Section A.1 we report and discuss the results of the heterogeneous treatment effect estimation simulation study over the same instances.

A.1 Discovery

In this Section, we discuss, one by one, the results of the simulations study on heterogeneity discovery presented in Section 4.1 on the five variant data generating processes described above.

Large Sample

In Figure A.1 we report the results for heterogeneity discovery increasing the sample size to $N = 5,000$ individuals. As expected, all the methods follow the same trends

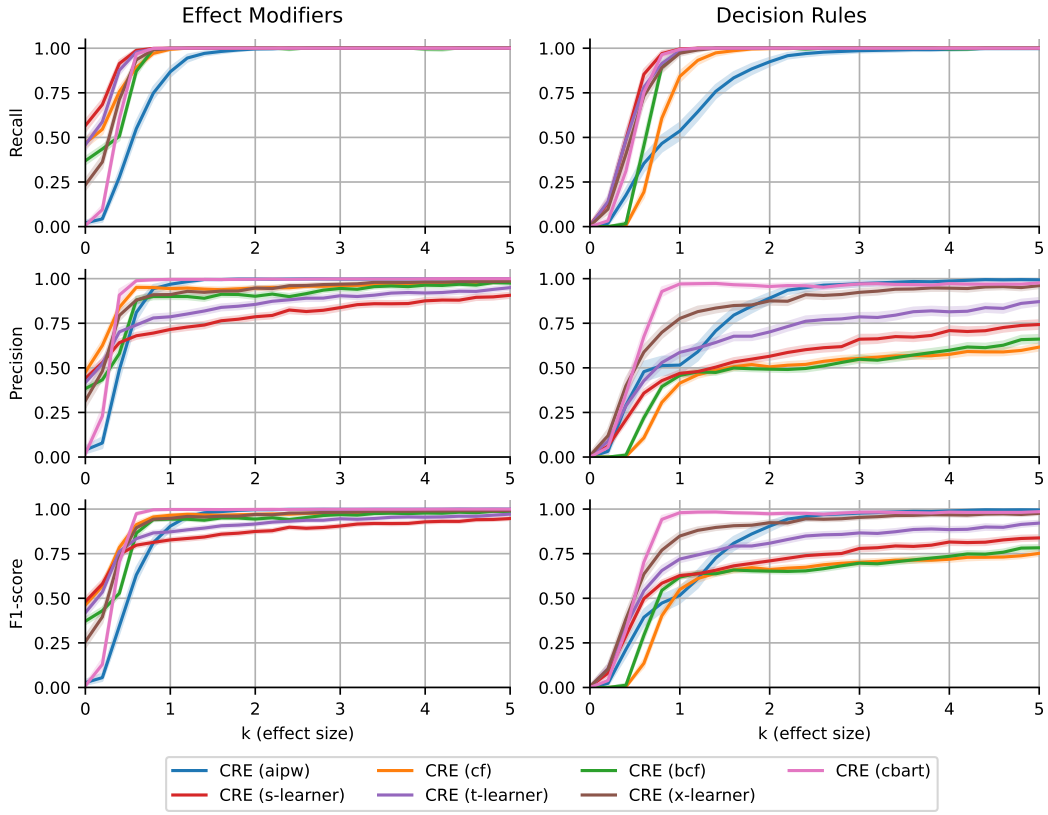


Figure A.1: Simulation study for heterogeneity discovery results with 2 rules, linear confounders and 5,000 observations. Mean *Precision*, *Recall* and $F1-score$ (lines) with the corresponding 95% confidence intervals (bands) over 250 Monte Carlo experiments are reported for each method and causal effect size k. For each CRE variant, the heterogeneity characterization discovery converges (with respect to effect size) to the true heterogeneity characterization.

described for the original data generating process, but significantly increasing the convergence rate, in particular for the *Recall* in both Estimation and Decision Rules retrieval. CRE (AIPW) is the unique method not speeding up the convergence rate

to perfect recovery, probably due to its instability issued already discussed in Section 4.2.

Small Sample

In Figure A.2 we report the results for heterogeneity discovery decreasing the sample size to $N = 1,000$ individuals. As expected, all the methods follow the same trends

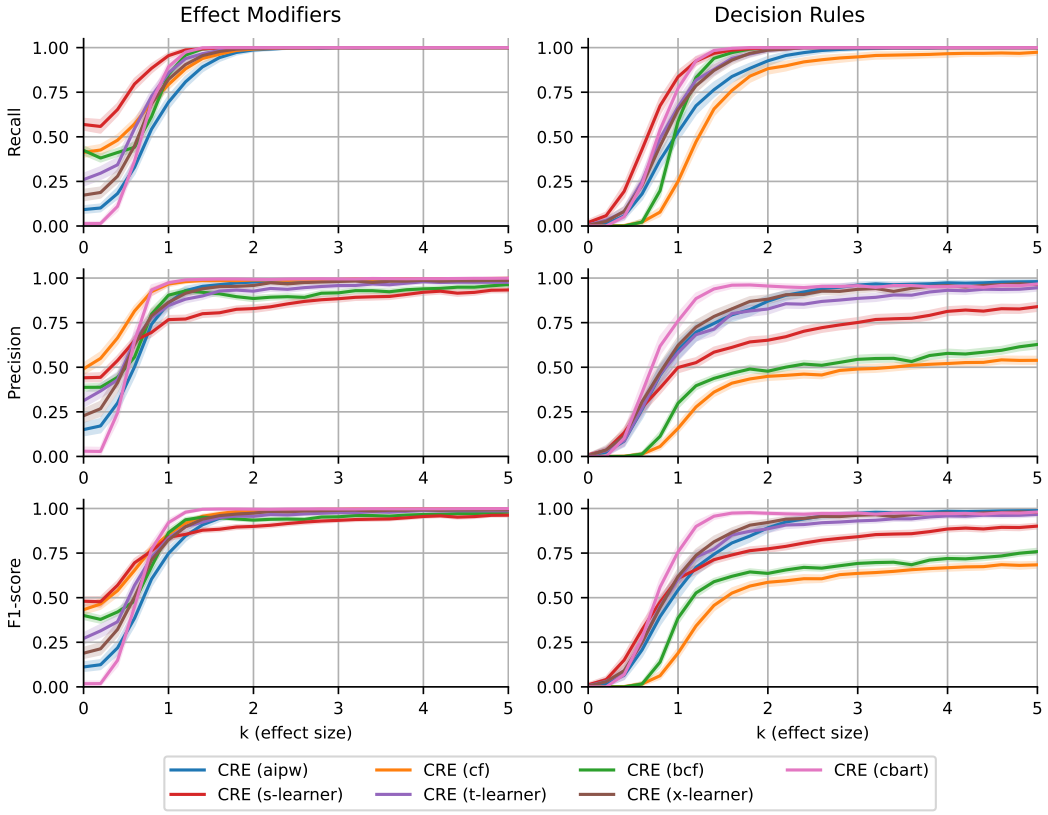


Figure A.2: Simulation study for heterogeneity discovery results with 2 rules, linear confounders and 1,000 observations. Mean $Precision$, $Recall$ and $F1 - score$ (lines) with the corresponding 95% confidence intervals (bands) over 250 Monte Carlo experiments are reported for each method and causal effect size k. For each CRE variant, the heterogeneity characterization discovery converges (with respect to effect size) to the true heterogeneity characterization.

described for the original data generating process, without significantly decreasing the convergence rate toward perfect discovery. These results strongly encourage the use of CRE even in a small sample regime.

More Rules

In Figure A.3, we report the results for heterogeneity discovery, increasing the number of decision rules to $M = 4$. As expected, (almost) all the methods increase their

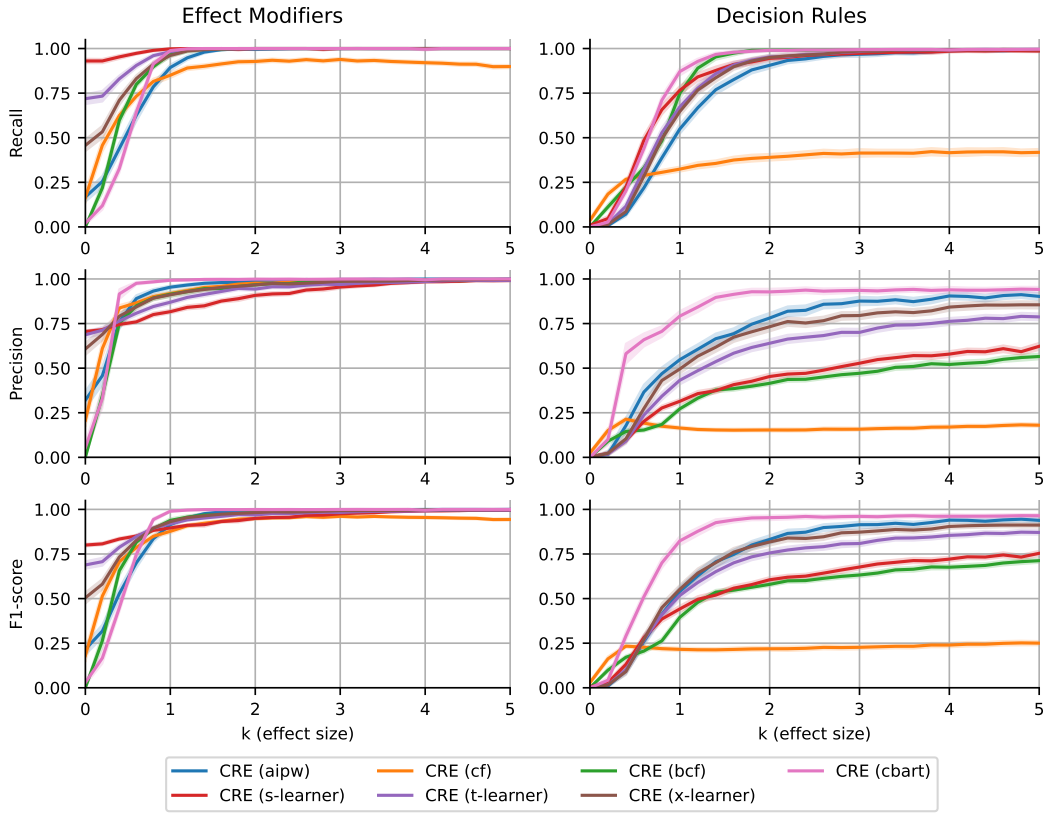


Figure A.3: Simulation study for heterogeneity discovery results with 4 rules, linear confounders and 2,000 observations. Mean $Precision$, $Recall$ and $F1-score$ (lines) with the corresponding 95% confidence intervals (bands) over 250 Monte Carlo experiments are reported for each method and causal effect size k. For each CRE variant, the heterogeneity characterization discovery converges (with respect to effect size) to the true heterogeneity characterization.

discovery performances increasing the causal effect (k). There are now seven effect modifiers out of $p = 10$ covariates, which leads to easier effect modifiers retrieval. The decision rules discovery is instead more challenging due to the higher and heterogeneous/more complex number of rules to retrieve. All the methods, with except to CRE (CF), still perfectly retrieve all the true decision rules with $k > 3$ ($Recall = 1$), but again, they often retrieve also wrong or redundant rules ($Precision < 1$). CRE (CF) is the unique method that does not showing a significant dependence on the causal effect for $k > 1$. Our hypothesis is that during the ITE estimation, it struggles more than the other methods in trying to express the heterogeneity in the longest

rules (i.e., r_3) through its causal trees.

Randomized Controlled Trial

In Figure A.4, we report the results for heterogeneity discovery with no confounding (if not in terms of decision rules). As expected, all the methods follow the same

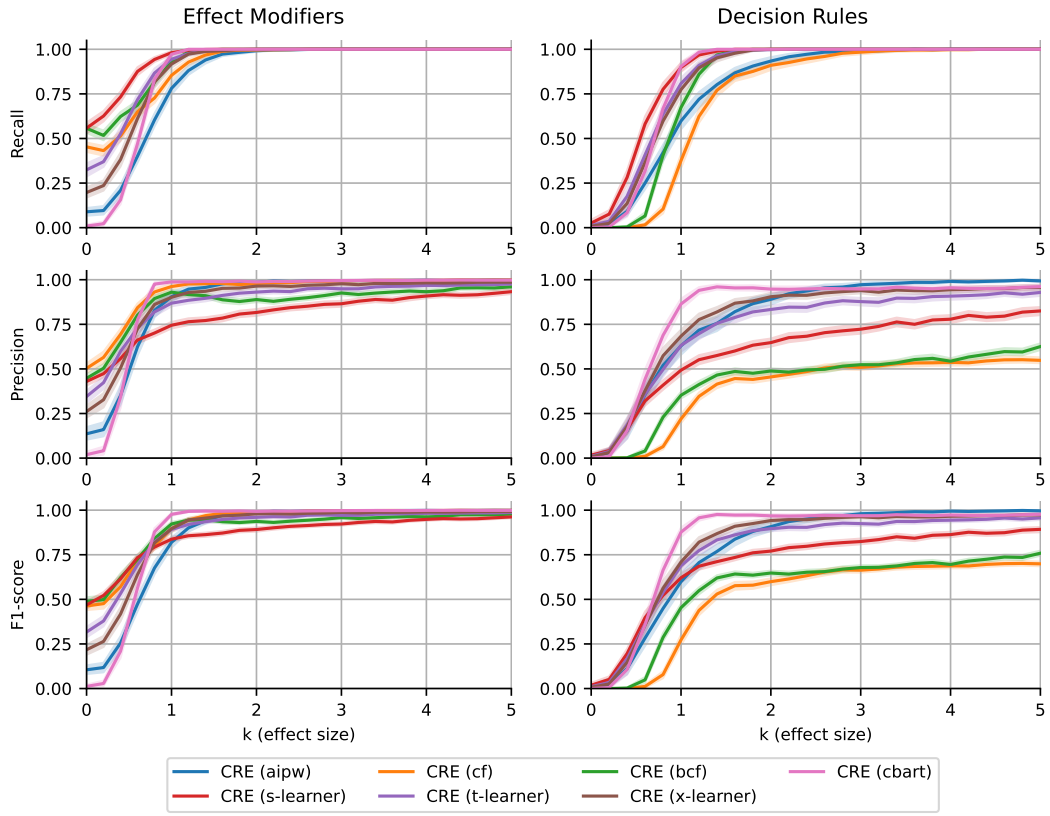


Figure A.4: Simulation study for heterogeneity discovery results with 2 rules, no confounders (randomized controlled trial) and 2,000 observations. Mean $Precision$, $Recall$ and $F1-score$ (lines) with the corresponding 95% confidence intervals (bands) over 250 Monte Carlo experiments are reported for each method and causal effect size k. For each CRE variant, the heterogeneity characterization discovery converges (with respect to effect size) to the true heterogeneity characterization.

trends described for the original data generating process, but significantly and equally increasing the convergence rate, in particular for the *Recall* in both Estimation and Decision Rules retrieval. Indeed, removing additional confounding mechanism facilitate all the estimation steps.

Non-Linear Confounding

In Figure A.5, we report the results for heterogeneity discovery with non-linear confounding. As expected, all the methods follow the same trends described for the

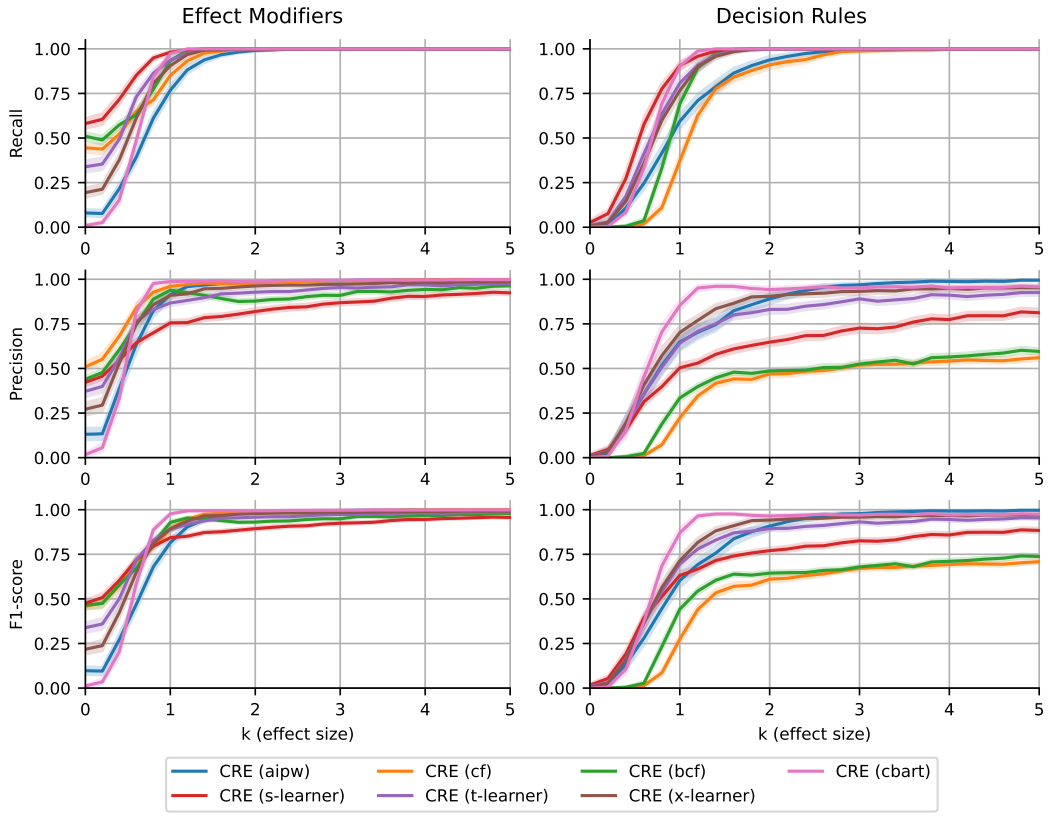


Figure A.5: Simulation study for heterogeneity discovery results with 2 rules, non-linear confounders and 2,000 observations. Mean $Precision$, $Recall$ and $F1-score$ (lines) with the corresponding 95% confidence intervals (bands) over 250 Monte Carlo experiments are reported for each method and causal effect size k. For each CRE variant, the heterogeneity characterization discovery converges (with respect to effect size) to the true heterogeneity characterization.

original data generating process, without significantly decreasing the convergence rate towards perfect discovery, although the more complex confounding mechanism.

A.2 Estimation

In this Section, we discuss, one by one, the results of the simulations study on heterogeneous treatment effect estimation presented in Section 4.2 on the five variant data generating processes described above.

Large Sample

In Table A.1, we report the results for heterogeneous treatment effect estimation, increasing the sample size to $N = 5,000$ individuals. As discussed in Section 4.2,

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	0.0804	0.0357	0.0011	0.0527
CRE (CF)	0.1598	0.0714	-0.0007	0.0505
CRE (BCF)	0.0891	0.0324	0.0022	0.0488
CRE (S-Learner)	0.0918	0.0332	0.0010	0.0497
CRE (T-Learner)	0.0905	0.0343	0.0034	0.0522
CRE (X-Learner)	0.0850	0.0337	0.0034	0.0522
CRE (Causal BART)	**0.0781**	0.0306	**0.0002**	0.0490
AIPW	2.2526	0.0910	0.0016	0.0342
CF	0.2070	0.0606	-0.0043	0.0330
BCF	0.0814	0.0234	0.0020	0.0319
S-Learner	0.3110	0.0218	0.0012	0.0333
T-Learner	0.5090	0.0214	0.0024	0.0333
X-Learner	1.0756	0.0140	0.0024	0.0333
Causal BART	0.9977	0.0099	0.0003	0.0315

Table A.1: Simulation study for HTE estimation, with $M = 2$ rules, linear confounder, 5,000 individuals and under CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

all the CRE variants significantly outperform the corresponding 'standalone' ITE estimators in both ITE and ATE estimation. Bayesian Causal Forest is the unique ITE estimator in getting similar performances to the corresponding CRE variant. AIPW estimator still suffers from not stabilized ITE prediction. Causal Forest, and similarly CRE (CF), significantly improve their estimation performances with respect to the original data generating process, leading to unbiased estimation, enforcing our hypothesis of empirically slower Causal Forest consistency convergence rate.

In Figure A.6, we report the results for AATEs estimation, increasing the sample size to $N = 5,000$ individuals. As expected from Proposition B, all the methods lead to consistent AATEs estimation, with a confidence interval even smaller than

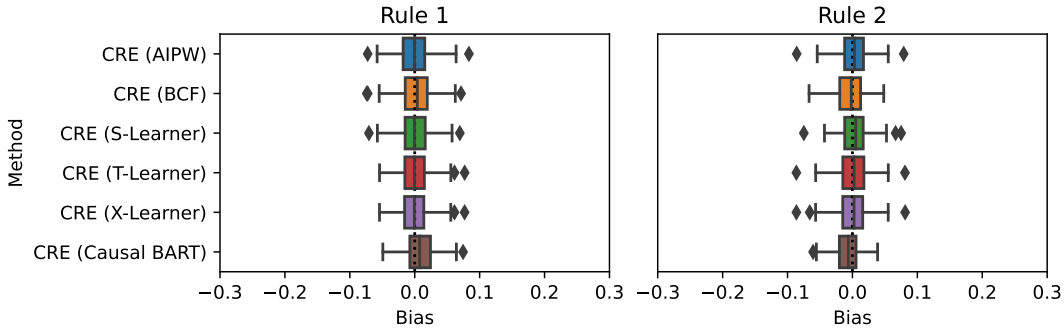


Figure A.6: Simulation study for HTE estimation, with $M = 2$ rules, linear confounding and 5,000 individuals. For all the CRE variants, for each rule, the AATE's bias over 250 Monte Carlo experiments is reported in a boxplot.

the original data-generating process.

Small Sample

In Table A.2, we report the results for heterogeneous treatment effect estimation, decreasing the sample size to $N = 1,000$ individuals. As discussed in Section 4.2,

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	0.2195	0.0882	-0.0136	0.1343
CRE (CF)	1.4953	0.2715	0.1043	0.1695
CRE (BCF)	0.2215	0.0758	0.0055	0.1182
CRE (S-Learner)	0.2312	0.0880	-0.0132	0.1321
CRE (T-Learner)	0.1976	0.0859	-0.0265	0.1194
CRE (X-Learner)	**0.1961**	0.0854	-0.0265	0.1194
CRE (Causal BART)	0.2225	0.0821	**-0.0003**	0.1187
AIPW	1.7045	0.1320	-0.0037	0.0843
CF	0.7172	0.1290	0.0338	0.0945
BCF	0.2056	0.0570	0.0022	0.0847
S-Learner	0.6844	0.0590	-0.0030	0.0808
T-Learner	1.1078	0.0590	-0.0064	0.0858
X-Learner	1.3403	0.0527	-0.0064	0.0858
Causal BART	0.9862	0.0248	-0.0007	0.0850

Table A.2: Simulation study for HTE estimation, with $M = 2$ rules, linear confounder, 1,000 individuals and under CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

(almost) all the CRE variants significantly outperform the corresponding 'standalone' ITE estimators in both ITE and ATE estimation without significantly worsening the performances from the original data-generating process (with larger sample size), with exceptions of CF and BCF. Indeed, Causal Forest in a small sample regime leads to even more systematic errors in estimation, which drastically propagate in the corresponding CRE variant.

In Figure A.7, we report the results for AATEs estimation, increasing the sample size to $N = 1,000$ individuals. As expected from Proposition B, all the methods lead

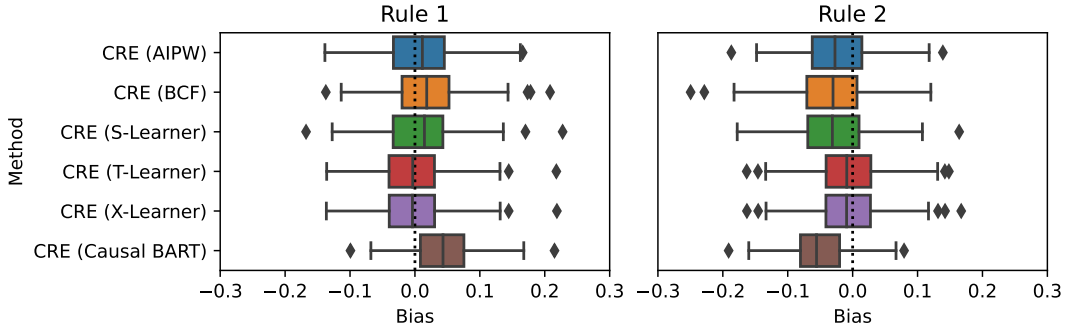


Figure A.7: Simulation study for HTE estimation, with $M = 2$ rules, linear confounding and 1,000 individuals. For all the CRE variants, for each rule, the AATE's bias over 250 Monte Carlo experiments is reported in a boxplot.

to consistent AATEs estimation, with not significantly larger confidence intervals with respect to the original data generating process, although the sample size.

More Rules

In Table A.3, we report the results for heterogeneous treatment effect estimation, increasing the number of decision rules to $M = 4$. As discussed in Section 4.2, (almost) all the CRE variants significantly outperform the corresponding 'standalone' ITE estimators in both ITE and ATE estimation without significantly worsening the performances from the original data-generating process (with simpler CATE decomposition), with exceptions of CF and BCF. Indeed, Causal Forest in a more complex CATE characterization regime leads to even more systematic errors in estimation, which drastically propagate in the corresponding CRE variant.

In Figure A.8, we report the results for AATEs estimation, increasing the number of decision rules to $M = 4$. As expected from Proposition B, all the methods lead to consistent AATE estimation for (almost) all the rules. Only the fourth and longest rule is slightly underestimated by almost all the methods, probably due to the redundant recovery of other similar decision rules ($Precision < 1$).

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	0.2505	0.1897	-0.0029	0.0936
CRE (CF)	3.3730	0.1470	0.2048	0.1780
CRE (BCF)	0.1901	0.0884	0.0061	0.0804
CRE (S-Learner)	0.2967	0.1394	-0.0044	0.0900
CRE (T-Learner)	0.2349	0.0943	**0.0003**	0.0944
CRE (X-Learner)	0.2356	0.1416	**0.0003**	0.0948
CRE (Causal BART)	0.1757	0.0729	-0.0017	0.0810
AIPW	2.1139	0.2077	0.0007	0.0561
CF	2.2405	0.1421	0.1331	0.0932
BCF	**0.1698**	0.0353	0.0045	0.0519
S-Learner	0.5410	0.0394	-0.0006	0.0532
T-Learner	0.8075	0.0371	0.0026	0.0567
X-Learner	1.1883	0.0293	0.0026	0.0567
Causal BART	0.9994	0.0161	0.0012	0.0517

Table A.3: Simulation study for HTE estimation, with $M = 4$ rules, linear confounder, 2,000 individuals and under CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

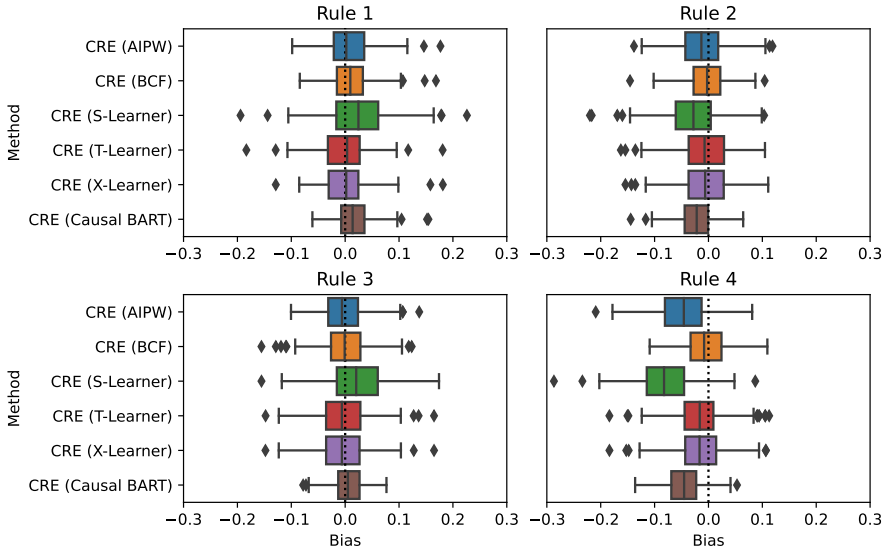


Figure A.8: Simulation study for HTE estimation, with $M = 4$ rules, linear confounding and 2,000 individuals. For all the CRE variants, for each rule, the AATE's bias over 250 Monte Carlo experiments is reported in a boxplot.

Randomized Controlled Trial

In Table A.4, we report the results for heterogeneous treatment effect estimation, with no confounding (if not in terms of decision rules). As discussed in Section 4.2,

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	**0.1342**	0.0602	0.0031	0.0884
CRE (CF)	0.6379	0.1389	0.0196	0.0946
CRE (BCF)	0.1492	0.0552	0.0047	0.0796
CRE (S-Learner)	0.1465	0.0576	0.0025	0.0848
CRE (T-Learner)	0.1489	0.0659	0.0062	0.0931
CRE (X-Learner)	0.1443	0.0655	0.0062	0.0931
CRE (Causal BART)	0.1457	0.0621	**0.0010**	0.0810
AIPW	2.0757	0.1897	0.0039	0.0560
CF	0.3022	0.0884	0.0019	0.0542
BCF	0.1351	0.0371	0.0045	0.0517
S-Learner	0.4690	0.0342	0.0031	0.0527
T-Learner	0.8042	0.0367	0.0042	0.0563
X-Learner	1.1863	0.0286	0.0042	0.0563
Causal BART	0.9924	0.0165	0.0021	0.0523

Table A.4: Simulation study for HTE estimation, with $M = 2$ rules, no-confounder (randomized controlled trial), 2,000 individuals and under CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

(almost) all the CRE variants significantly outperform the corresponding 'standalone' ITE estimators in both ITE and ATE estimation without significantly worsening the performances from the original data-generating process (with linear confounding), with exceptions of CF and BCF. Given the similarity of the results with the original data-generating process, we empirically observe that the under the assumption of unconfoundness (Assumption 3) CRE algorithm is robust with respect to the confounfounding mechanism. CRE (AIPW) is the best-performing method in ITE estimation (although the unstable AIPW pseudo-outcome estimation) and CRE (Causal BART) leads to the most consistent estimate.

In Figure A.9, we report the results for AATEs estimation, with no confounding (if not in terms of decision rules). As expected from Proposition B, all the methods lead to consistent AATEs estimation.

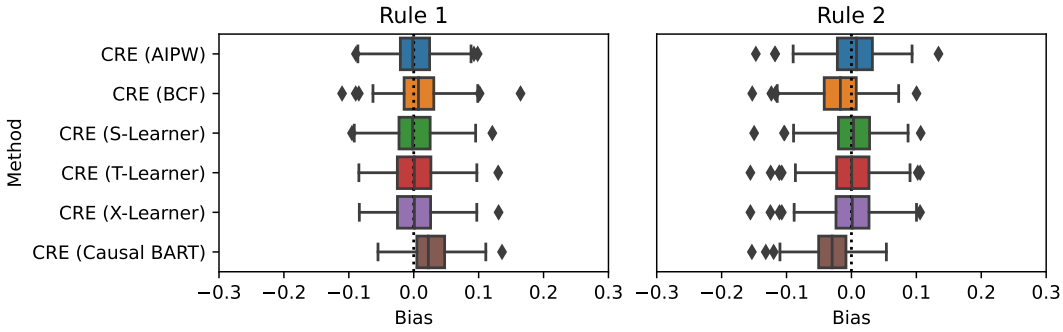


Figure A.9: Simulation study for HTE estimation, with $M = 2$ rules, no-confounding (randomized controlled trial) and 2,000 individuals. For all the CRE variants, for each rule, the AATE's bias over 250 Monte Carlo experiments is reported in a box-plot.

Non-Linear Confounding

In Table A.5, we report the results for heterogeneous treatment effect estimation with non-linear confounding. The confounding mechanism doesn't seem to significantly

Method	RMSE		Bias	
	μ	σ	μ	σ
CRE (AIPW)	**0.1343**	0.0601	0.0037	0.0881
CRE (CF)	0.6374	0.1400	-0.0059	0.0949
CRE (BCF)	0.1493	0.0555	0.0050	0.0790
CRE (S-Learner)	0.1492	0.0599	0.0035	0.0849
CRE (T-Learner)	0.1490	0.0664	0.0074	0.0937
CRE (X-Learner)	0.1460	0.0651	0.0073	0.0938
CRE (Causal BART)	0.1421	0.0628	**0.0002**	0.0817
AIPW	2.0767	0.1945	0.0043	0.0560
CF	0.3053	0.0896	-0.0074	0.0549
BCF	0.1347	0.0370	0.0045	0.0524
S-Learner	0.4721	0.0333	0.0033	0.0529
T-Learner	0.8052	0.0373	0.0048	0.0568
X-Learner	1.1870	0.0292	0.0048	0.0568
Causal BART	0.9925	0.0164	0.0019	0.0517

Table A.5: Simulation study for HTE estimation, with $M = 2$ rules, non-linear confounder, 2,000 individuals and under CATE linear decomposition assumption. For all the methods, the mean (μ) and standard deviation (σ) treatment effect root mean squared error (RMSE) and bias (Bias) over 250 Monte Carlo experiments are reported.

impact the estimation performances, and the results obtained look very similar to

the ones from the original data-generating process and the randomized controlled experiment variant. As discussed in Section 4.2, (almost) all the CRE variants significantly outperform the corresponding 'standalone' ITE estimators in both ITE and ATE estimation without significantly worsening the performances from the original data-generating process (with linear confounding), with exceptions of CF and BCF.

In Figure A.10, we report the results for AATEs estimation, with non-linear confounding. As expected from Proposition B, all the methods lead to consistent AATEs estimation.

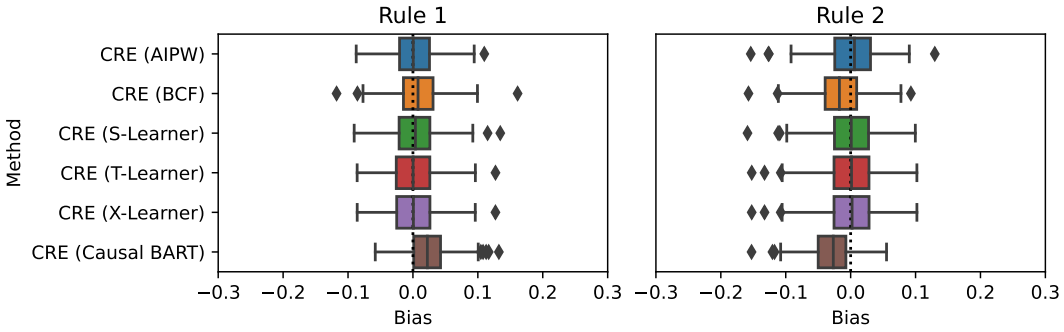


Figure A.10: Simulation study for HTE estimation, with $M = 2$ rules, non-linear confounding and 2,000 individuals. For all the CRE variants, for each rule, the AATE's bias over 250 Monte Carlo experiments is reported in a boxplot.

Appendix B

Deferred Proofs

In this Appendix, we report the proofs of the Propositions presented in Chapter 2 and 3.

Proposition 1. (Linear Decomposition in finite covariate space)

If the covariate space $\mathcal{X}$ is finite, then the Treatment Effect linear decomposition Assumption holds.

Proof. Without loss of generality, we assume that all the covariates are binary: $\mathcal{X} = \{0,1\}^p$ (the same argument can be used in the discrete case). For all $m \in \{0,1,2,3,...2^p-1\}$ let $\boldsymbol{x}_m = (b_1, b_2, ..., b_p) \in \mathcal{X}$, where $b_1 b_2 ... b_p$ is the representation of m in base 2, adding zeros on the left if less then p digits are required. For example $\boldsymbol{x}_2 = (0,0,...,0,1,0)$, $\boldsymbol{x}_6 = (0,0,...,0,1,1,0)$ and $\boldsymbol{x}_{2^p-1} = (1,1,...,1,1,1,1)$.

Since, by construction, $\mathcal{X} = \cup_{m=0}^{2^p-1}\{\boldsymbol{x}_m\}$, the (centered) Conditional Average Treatment Effect can be dummy decomposed in 2^p point-wise contributions:

$$\begin{aligned} \tau(\boldsymbol{x}) - \bar{\tau} &= \sum_{m=0}^{2^p-1} \tau(\boldsymbol{x}_m) \cdot \mathbf{1}_{\boldsymbol{x}_m}(\boldsymbol{x}) \\ &= \sum_{m=0}^{2^p-1} \alpha_m \cdot r_m(\boldsymbol{x}) \end{aligned} \tag{B.1}$$

□

Proposition 2. (Consistency of the AATE estimator)

Let $\hat{\boldsymbol{\tau}}$ a consistent estimator for $\boldsymbol{\tau}$ (i.e., AIPW). Under the Treatment Effect linear decomposition Assumption (Condition 1) and assuming $\mathbb{E}(\boldsymbol{R}^T\boldsymbol{R}) = Q$ is a positive definite matrix (Condition 2), the Additive Average Treatment Effects estimator $\hat{\boldsymbol{\alpha}} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T(\hat{\boldsymbol{\tau}} - \hat{\bar{\tau}})$ is a consistent estimator for $\boldsymbol{\alpha}$.

Proof. Multiplying Equation 3.25 (Condition 1) on the both sides by $(R^T R)^{-1} R^T$, we get:

$$(R^T R)^{-1} R^T (\hat{\boldsymbol{\tau}} - \hat{\bar{\tau}}) = (R^T R)^{-1} R^T R \boldsymbol{\alpha} + (R^T R)^{-1} R^T \boldsymbol{\nu}. \tag{B.2}$$

Using Equation 3.26, and simplifying the right member:

$$\rightarrow \quad \hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + (R^T R)^{-1} R^T \boldsymbol{\nu}. \tag{B.3}$$

Observing that by Condition 2 and the Law of large numbers:

$$\frac{R^T R}{N} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{R}_i^T \cdot \boldsymbol{R}_i \xrightarrow{d} Q \succ 0, \tag{B.4}$$

where $\boldsymbol{R}_i$ represents the i-th row of the rules matrix, and:

$$\frac{R^T \boldsymbol{\nu}}{N} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{R}_i^T \cdot \nu_i \xrightarrow{d} \mathbf{0}; \tag{B.5}$$

combining them in Equation B.3 (simplifying N), by Slutsky's theorem:

$$\hat{\boldsymbol{\alpha}} \xrightarrow{d} \boldsymbol{\alpha} \tag{B.6}$$

□

Proposition 3. (Asymptotic Normality of the AATE estimator)

If Conditions (1)-(5) hold, then

$$\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, V) \quad \text{as} \quad N \to \infty \tag{B.7}$$

where $V = Q^{-1} \Omega Q^{-1}$.

Proof. Similarly to the proof of Proposition 2, multiplying Equation 3.25 (Condition 1) on the both sides by $(R^T R)^{-1} R^T$, and inserting Equation 3.26, we get:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + (R^T R)^{-1} R^T \boldsymbol{\nu}. \tag{B.8}$$

Multiplying both sides by $\sqrt{N}$ and rearranging we get:

$$\begin{aligned} \sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) &= \left(\frac{R^T R}{N} \right)^{-1} \frac{R^T \boldsymbol{\nu}}{\sqrt{N}} \\ &= \left(\frac{\sum_{i=1}^{N} \boldsymbol{R}_i^T \boldsymbol{R}_i}{N} \right)^{-1} \frac{\sum_{i=1}^{N} \boldsymbol{R}_i^T \nu_i}{\sqrt{N}} \end{aligned}$$

By hypothesis:

$$\mathbb{E}[\boldsymbol{R}_i \nu_i] = \mathbf{0} \qquad \forall i \in \mathcal{I}^e \tag{B.9}$$

and:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{R}_i\nu_i) &= \mathbb{E}[\nu_i^2\boldsymbol{R}_i^T\boldsymbol{R}_i] - \mathbb{E}[\boldsymbol{R}_i\nu_i]^2 \\ \& = \mathbb{E}[\nu_i^2\boldsymbol{R}_i^T\boldsymbol{R}_i] = \Omega \succ 0 \qquad & \forall i \in \mathcal{I}^e \end{aligned} \tag{B.10}$$

Then, by the Central Limit Theorem:

$$\frac{\sum_{i=1}^N \boldsymbol{R}_i^T\nu_i}{N} \xrightarrow{d} \mathcal{N}(0,\Omega). \tag{B.11}$$

We have already discussed in the proof of Proposition 2 that:

$$\frac{R^TR}{N} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{R}_i^T \cdot \boldsymbol{R}_i \xrightarrow{d} Q \succ 0. \tag{B.12}$$

Then, combining these results in Equation B.9, by Slutsky' theorem and Cramer-Wold theorem:

$$\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, Q^{-1}\Omega Q^{-1}) \;\text{ as }\; N \to \infty \tag{B.13}$$

□

Bibliography

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences* **113**(27): 7353–7360.

Athey, S., Tibshirani, J., Wager, S. et al. (2019). Generalized random forests, *The Annals of Statistics* **47**(2): 1148–1178.

Bargagli-Stoffi, F. J., Cadei, R., Lee, K. and Dominici, F. (2023). Causal rule ensemble: Interpretable Discovery and Inference of Heterogeneous Treatment Effects, *arXiv preprint arXiv:2009.09036* .

Bargagli Stoffi, F. J., Cevolani, G. and Gnecco, G. (2022). Simple models in complex worlds: Occam's razor and statistical learning theory, *Minds and Machines* **32**(1): 13–42.

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

Bargagli-Stoffi, F. J. and Gnecco, G. (2020). Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms, *International Journal of Data Science and Analytics* **9**: 315–337.

Bargagli-Stoffi, F. J., Tortù, C. and Forastiere, L. (2020). Heterogeneous treatment and spillover effects under clustered network interference, *arXiv preprint arXiv:2008.00707* .

Baxter, L. K., Duvall, R. M. and Sacks, J. (2013). Examining the effects of air pollution composition on within region differences in pm2. 5 mortality risk estimates, *Journal of Exposure Science & Environmental Epidemiology* **23**(5): 457–465.

Belloni, A., Chernozhukov, V., Hansen, C. and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control, *Journal of Business & Economic Statistics* **34**(4): 590–605.

Bodinier, B., Filippi, S., Nost, T. H., Chiquet, J. and Chadeau-Hyam, M. (2021). Automated calibration for stability selection in penalised regression and graphical models: a multi-omics network application exploring the molecular response to tobacco smoking, *arXiv preprint arXiv:2106.02521* .

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The annals of statistics* **24**(6): 2350–2383.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Cadei, R., Khoshnevis, N. and Bargagli-Stoffi, F. J. (2023). Cre: an r package for interpretable discovery and estimation of heterogeneous treatment effect, *Working paper* .

Carone, M., Dominici, F. and Sheppard, L. (2020). In pursuit of evidence in air pollution epidemiology: the role of causally driven data science, *Epidemiology* **31**(1): 1–6.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects, *American Economic Review* **107**(5): 261–65.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2016). Locally robust semiparametric estimation, *arXiv preprint arXiv:1608.00033* .

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.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels, *Biometrika* **62**(2): 441–444.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity, *The Review of Economics and Statistics* **90**(3): 389–405.

Dedoussi, I. C., Eastham, S. D., Monier, E. and Barrett, S. R. (2020). Premature mortality related to united states cross-state air pollution, *Nature* **578**(7794): 261–265.

Deng, H. (2019). Interpreting tree ensembles with intrees, *International Journal of Data Science and Analytics* **7**(4): 277–287.

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.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G. and Speizer, F. E. (1993). An association between air pollution

and mortality in six us cities, *New England Journal of Medicine* **329**(24): 1753–1759.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L. and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases, *JAMA* **295**(10): 1127–1134.

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**: 135–178.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, Vol. 38, Philadelphia, PA: Society for Industrial and Applied Mathematics.

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.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy, *International Encyclopedia of the social & Behavioral sciences* **6**(4027-4030): 1–7.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *The Annals of Statistics* **29**(9): 1189–1232.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles, *The Annals of Applied Statistics* **2**(3): 916–954.

Hahn, P. R., Dorie, V. and Murray, J. S. (2019). Atlantic causal inference conference (acic) data analysis challenge 2017, *arXiv preprint arXiv:1905.09515* .

Hahn, P. R., Murray, J. S., Carvalho, C. M. et al. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, *Bayesian Analysis* .

Hansotia, B. and Rukstales, B. (2002). Incremental value modeling, *Journal of Interactive Marketing* **16**(3): 35–46.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

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

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

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American statistical Association* **47**(260): 663–685.

Imai, K., Ratkovic, M. et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation, *The Annals of Applied Statistics* **7**(1): 443–470.

Jacob, D. (2019). Group average treatment effects for observational studies, *arXiv preprint arXiv:1911.02688* .

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.

Johnson, M., Cao, J. and Kang, H. (2022). Detecting heterogeneous treatment effects with instrumental variables and application to the oregon health insurance experiment, *The Annals of Applied Statistics* **16**(2): 1111–1129.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects, *arXiv preprint arXiv:2004.14497* .
URL: *https://arxiv.org/abs/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**: 1229–1245.

Khoshnevis, N., Garcia, D. M., Cadei, R., Lee, K. and Bargagli-Stoffi, F. J. (2023). *CRE: Interpretable Subgroups Identification Through Ensemble Learning of Causal Rules*. R package version 0.2.0.9000.
URL: *https://github.com/NSAPH-Software/CRE*

Kim, B., Khanna, R. and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability, *Advances in Neural Information Processing Systems*, pp. 2280–2288.

Kloog, I., Ridgway, B., Koutrakis, P., Coull, B. A. and Schwartz, J. D. (2013). Long- and short-term exposure to pm2. 5 and mortality: using novel exposure models, *Epidemiology (Cambridge, Mass.)* **24**(4): 555.

Kuhn, M., Johnson, K. et al. (2013). *Applied predictive modeling*, Vol. 26, Springer.

Künzel, S. R., Sekhon, J. S., Bickel, P. J. and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning, *Proceedings of the National Academy of Sciences* **116**(10): 4156–4165.

Lakkaraju, H., Bach, S. H. and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684.

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* pp. 1–33.

Lee, M.-j. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(1): 243–264.

Lewis, J. B. and Linzer, D. A. (2005). Estimating regression models in which the dependent variable is based on estimates, *Political Analysis* **13**(4): 345–364.

Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Guo, Y., Tong, S., Coelho, M. S., Saldiva, P. H., Lavigne, E., Matus, P. et al. (2019). Ambient particulate air pollution and daily mortality in 652 cities, *New England Journal of Medicine* **381**(8): 705–715.

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.

Loh, W.-Y., Cao, L. and Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(5): e1326.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model, *The American Statistician* **54**(3): 217–224.

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.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4): 417–473.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267**: 1–38.

Nagpal, C., Wei, D., Vinzamuri, B., Shekhar, M., Berger, S. E., Das, S. and Varshney, K. R. (2020). Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines, *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 19–29.

Nalenz, M. and Villani, M. (2018). Tree ensembles with rule structured horseshoe regularization, *The Annals of Applied Statistics* **12**(4): 2379–2408.

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.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects, *Biometrika* **108**: 299–319.
URL: *https://arxiv.org/abs/1712.04912v4*

Pappin, A. J., Christidis, T., Pinault, L. L., Crouse, D. L., Brook, J. R., Erickson, A., Hystad, P., Li, C., Martin, R. V., Meng, J. et al. (2019). Examining the shape of the association between low levels of fine particulate matter and mortality across three cycles of the canadian census health and environment cohort, *Environmental Health Perspectives* **127**(10): 107008.

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.

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

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models, *Statistics in medicine* **16**(3): 285–319.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American statistical Association* **89**(427): 846–866.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies., *Journal of Educational Psychology* **66**(5): 688–701.

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

Schwartz, J., Wei, Y., Di, Q., Dominici, F., Zanobetti, A. et al. (2021). A national difference in differences analysis of the effect of pm2. 5 on annual death rates, *Environmental Research* **194**: 110649.

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.

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.

Spanbauer, C. and Sparapani, R. (2021). Nonparametric machine learning for precision medicine with longitudinal clinical trials and bayesian additive regression trees with mixed models, *Statistics in Medicine* **40**(11): 2665–2691.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2): 111–133.

Su, L., Shi, Z. and Phillips, P. C. (2016). Identifying latent structures in panel data, *Econometrica* **84**(6): 2215–2264.

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

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

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association* **113**(523): 1228–1242.

Wang, T. and Rudin, C. (2022). Causal rule sets for identifying subgroups with enhanced treatment effects, *INFORMS Journal on Computing* .

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases, *Statistics in Medicine* **37**(23): 3309–3324.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.

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.

Wu, X., Netherly, R., Sabath, M., Braun, D. and Dominici, F. (2020). Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis, *Science advances* **6**(45): eabd4049.

Yang, J., Dahabreh, I. J. and Steingrimsson, J. A. (2021). Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data, *Biometrics* .
URL: *http://dx.doi.org/10.1111/biom.13432*

Yu, B. (2013). Stability, *Bernoulli* **19**(4): 1484–1500.

Zanobetti, A., Franklin, M., Koutrakis, P. and Schwartz, J. (2009). Fine particulate air pollution and its components in association with cause-specific emergency admissions, *Environmental Health* **8**(1): 1–12.

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