Adaptive Estimation of Group Average Policy Effects

Maria Nareklishvili [*]

First draft: August 2022
This draft: December 2023

Abstract

Decision-makers often face the challenge of identifying policy-relevant cohorts from a broad range of subjects. To address this, I investigate variations in policy effects in the presence of noisy data and redundant information. First, I identify data-driven population groups as a preliminary step for examining policy effect heterogeneity. Then I estimate the distribution of policy effects within and across the identified groups. My primary contribution is a multivariate causal forest, a novel, computationally efficient method for personalized policy analysis. I show that the method inherits desirable large sample properties, and enhances the reliability of findings by reducing the variance of policy effect estimates. To illustrate the applicability of the method, I revisit a field experiment, conducted in cooperation with the Norwegian Labor and Welfare Administration. In contrast to prior research, I discover that policy variations highlight the need for personalized interventions to optimize sickness absence management strategies.

[*] Department of Economics, University of Oslo, Oslo, Norway and Booth School of Business, University of Chicago, Illinois, USA (Correspondence to: maria.nareklishvili@frisch.uio.no; nareklishvili@uchicago.edu). This paper has benefitted from the constructive feedback of Knut Røed, Edwin Leuven, Dmitry Arkhangelsky, Stephane Bonhomme, Alexander Torgovitsky, Thomas Wiemann, Phillip Heiler, California Econometrics Conference participants, Seminars at the University of Oslo, University of Aarhus, Ragnar Frisch Center for Economic Research, as well as the Computational Workshop participants at the University of Chicago. All the remaining errors are my own. I acknowledge the funding by the Research Council of Norway, under project number 280350/GE.

1 Introduction

Policymakers and researchers frequently employ randomized experiments to investigate the impact of multiple programs, interventions, or policies on outcomes (Chernozhukov et al., 2017; Belloni et al., 2017; Hadad et al., 2021; Agrawal et al., 2022). For instance, Mayo-Wilson et al. (2017) design randomized clinical trials to investigate the effects of two treatments: gabapentin for neuropathic pain and quetiapine for bipolar depression. Lechner (2001) discusses the identification and estimation of multiple treatment effects under the conditional independence assumption.

The process of estimating and drawing inferences about policy effects encounters several fundamental challenges in the presence of heterogeneity, noise, and multiple outcomes. First, the effects of policies at the individual level remain unobservable. Consequently, distinguishing heterogeneous effects from inherent noise within specific population subgroups is a complex undertaking. Second, when dealing with redundant information and multiple correlated effects, personalized policy analysis can yield inefficient estimates and unreliable findings (Jackson et al., 2011). Traditional techniques such as cross-validation are insufficient for accurately identifying the subgroups that exhibit policy effect heterogeneity (Browning et al., 2007; Athey et al., 2015). Finally, traditional techniques like the causal forest proposed by Athey et al. (2019) demand significant computational resources when making separate predictions for multiple outcomes and parameters.

I propose a multivariate causal forest, a novel, computationally efficient approach for personalized policy analysis in the presence of correlated subjects. The framework sequentially integrates two distinct methodologies: partial least squares (Helland, 1988) and the causal forest algorithm (Athey and Imbens, 2016). Specifically, I identify population groups using the partial least squares algorithm. Subsequently, I use the multivariate causal forest to estimate the means and quantiles of policy effects both within and across the identified groups. Consequently, this article revolves around three fundamental research objectives. Firstly, I assess whether the method efficiently reduces the redundant dimensions and minimizes inherent noise. Secondly, I evaluate the capacity of the multivariate causal forest to capture the most diverse and policy-relevant segments of the population, especially when dealing with multiple correlated outcomes. Lastly, I evaluate the precision and coverage of multi-dimensional confidence intervals when accounting for the covariance of multiple parameters.

The multivariate causal forest is a natural extension of the causal forest algorithm (Athey and Imbens, 2016). The proposed approach offers distinct advantages compared to conventional methods that estimate policy effects within arbitrarily defined covariate subgroups. Specifically, it is adaptable to handle multiple, potentially correlated policy or treatment effects. This property makes it highly advantageous for situations in which policymakers are dealing with multiple interventions, and outcomes, or possess prior insights into the correlation structure across subjects. The method automatically selects covariates and their corresponding values for partitioning. This process incorporates honest splitting and cross-validation techniques, effectively reducing the potential for overfitting bias and thereby enhancing the reliability of the results. Additionally, the multivariate causal forest is highly computationally efficient and allows for the joint hypothesis testing.

The strategy for identifying data-driven population groups, inspired by factor models, aligns with the approach proposed by Nareklishvili et al. (2022). In this framework, the segments are characterized as continuous combinations of explanatory variables. Unlike discrete clusters, these continuous segments span the entire spectrum of the population subgroups, facilitating a flexible and comprehensive analysis of policy effect heterogeneity. In addition, they are designed to accommodate multiple correlated outcomes, and the resulting subgroups possess meaningful economic interpretations.

I contribute to the literature by offering both theoretical and empirical insights into personalized policy analysis. Building upon the work of Athey and Wager (2019) and Athey et al. (2019), my theoretical contribution is to establish the asymptotic normality of the multivariate causal forest estimator. By various simulated experiments, I show that in comparison to benchmark methods such as causal forest (Athey and Wager, 2019) and sorted group average policy effects (Sorted GATE by Chernozhukov et al., 2018), the multivariate causal forest improves the efficiency of personalized policy effects. The improvement is visible through several distinct mechanisms: it produces highly accurate policy effect estimates, and it significantly reduces the variance of personalized policy effects relative to the benchmark methods, even when dealing with a limited number of subjects.

My empirical contribution is to investigate the randomized field experiment conducted by Alpino et al. (2022) in cooperation with the Norwegian Labour and Welfare Administration (NAV). The primary objective of this experiment is to examine the effects of summoning sick-listed individuals to dialogue meetings, as compared to those who are not summoned but still have the option to request one. Specifically, I estimate the impact of dialogue meetings on two variables: total days of sickness absence and the number of sick leave days experienced within the given sick leave spell. In this paper, my emphasis is on a single policy and two outcomes. However, it's worth noting that the multivariate causal forest can be applied to predict multiple outcomes and policy effects simultaneously.

Alpino et al. (2022) shows that the conventional causal forest methodology fails to discover statistically significant heterogeneity in the effect of the dialogue meetings on sickness absence. Interestingly, I find that linear combinations of these characteristics forming policy-relevant groups might reflect statistically significant heterogeneity. In particular, I find distinct patterns among never-married females and married sick-listed workers concerning the effect of dialogue meetings on sickness absence. For never-married females with a low percentage of sickness absence, I observe a significant reduction in the total number of days on sick leave. On average, never-married females, comprising more than 15% of the sample population, experience a reduction of up to 10 days of sick leave. Conversely, the results for married sick-listed workers paint a different picture. Dialogue meetings lead to a statistically significant prolongation of sick leave, with an extension of up to 10 days. The policy effect for married individuals exhibits a higher variance compared to never-married females. The variations in policy effects highlight the need for tailored interventions that consider individual characteristics, such as marital status, to optimize sickness absence management strategies.

2 Related Literature

Despite the widespread practical application of random forests, the early theoretical research on this method primarily focuses on stylized or simplified versions of the original algorithm. Breiman (2001) establishes an upper bound on the generalization error of forests in terms of correlation and the strength of individual trees. This is followed by Breiman (2004), which concentrates on a stylized version of the original algorithm. Lin and Jeon (2006) draw attention to a connection between random forests and a specific class of nearest neighbor predictors. They also establish the lower bound of the expected mean squared error for nonadaptive forests, which is independent of the training set. Meinshausen and Ridgeway (2006) demonstrate the consistency of random forests in the context of conditional quantile prediction. Moving forward, Biau (2012) proves the consistency of random forests, provided independence of the candidate splits and the predicted leaf outcome. Denil et al. (2014), Wager (2014), and Scornet et al. (2015) propose consistent random forests that closely resemble the original algorithm, particularly in sparse feature spaces.

Athey and Imbens (2016) and Wager and Athey (2018) take a step forward in forest exploration by introducing causal forests for heterogeneous treatment effect analysis. They develop a univariate treatment effect estimator and prove the asymptotic normality of the estimated treatment effects using the Hajek projection of a U-statistic (Korolyuk and Borovskich, 2013). Athey et al. (2019) build upon this foundation by introducing moment conditions for the outcome variable, thus generalizing the method to a broader class of parameters. They show the consistency of a random forest tailored for correlated policy effects. Nekipelov et al. (2018) provide an interesting study demonstrating the uniform consistency of random forests with multiple correlated parameters for classification problems. Li (2020) extend the asymptotic theory of the random forest to accommodate correlated coefficients across multiple subsets (leaves) of a given tree. Ćevid et al. (2020) propose a novel forest construction for multivariate responses based on their joint conditional distribution, independent of the estimation target and the data model. Unlike this paper, the authors focus on outcomes rather than parameters, and prediction performance rather than the performance of joint confidence ellipses. Additionally, Wang et al. (2022) introduce random forests for the instrumental variable approach.

Chernozhukov and Hansen (2006) propose quantile instrumental variable regression for heterogeneous treatment effect analysis. A closely related article by Chernozhukov et al. (2018) introduces a generic machine learning approach to estimate and infer key features of heterogeneous treatment effects in randomized experiments. They proxy conditional average treatment effects by a given machine learning approach and post-process them for inferring treatment effects. Their approach is also valid for high-dimensional data. Additionally, Belloni et al. (2017) discuss inference on heterogeneous treatment effects based on high-performing machine learning methods.

The multivariate causal forest complements the existing theoretical work by extending the large sample theory to a multivariate setup. The multivariate causal forest has the ability to jointly predict the means and quantiles of policy parameters. In this paper, the multivariate random forest relies on local moment equations that are generalizations of Chernozhukov et al. (2018); Chernozhukov and Hansen (2006) and Athey and Imbens (2016) to group average policy effects. Specifically, we assume, each group policy effect is the sum of its' expectation and the corresponding error. The goal is to minimize the deviation between the unobserved personalized policy effects, and the observed group-level means of these effects. The moment function in this paper is also similar to generative adversarial networks, where we minimize the deviation of the covariates with respect to their expected value (Creswell et al., 2018; Liu and Yu, 2021).

The multivariate causal forest sets itself apart from the causal forest introduced by Athey and Imbens (2016) due to its multivariate nature, while still inheriting a similar local moment function. In contrast, it differs from the generalized random forest proposed by Athey et al. (2019) in terms of the estimation procedure. The multivariate causal forest jointly estimates parameters and outcome variables, whereas the generalized random forest estimates coefficients separately. Later in this article, I demonstrate that different estimation procedures can yield similar results. However, Wang and Lin (2005) and Kolenikov and Bollen (2012) discuss the efficiency of coefficients under misspecified variance. They find that, in the context of longitudinal analysis, correctly specifying the variance function can improve estimation efficiency. Additionally, the group average policy effects in this paper differ from the sorted group average policy effects (Chernozhukov et al., 2018) by the nature of the identified groups. Specifically, the groups here are allowed to be continuous, unordered, and a priori unknown. The identification of the latent groups in this article relies on the underlying data structure, rather than proxy estimates of conditional average policy effects (CATE).

3 Problem Formulation

In this section, I describe three main concerns: incorrectly specified distribution of the outcomes, computational complexity, and the issues in the presence of redundant characteristics. These concerns serve as the drivers for introducing the multivariate causal forest and the partial least squares tailored to multiple policies and outcomes.

Consider a policy-maker who can choose between two different policies: one involving summoning sick-listed workers to dialogue meetings (P_1) and the other using email notifications to inform them about the consequences of long-term sick leave (P_2). She is interested in two specific outcomes: the total number of days an individual spends on sick leave (Y_1); and the number of days specifically within a single sickness episode (Y_2). The policy-maker possesses a rich dataset containing various personal attributes of individuals, such as age, gender, education, and social status, which are collectively represented as $\mathbf{X}$.

The initial concern of the policy-maker is that policies aimed at individuals on sick leave may need to be tailored to account for differences in individual characteristics. For instance, young employees on sick leave might exhibit distinct behavioral patterns compared to their older counterparts. In this context, the analysis necessitates a method to identify the key variables that predominantly influence the diversity in the effects of policies.

The second concern of the policy-maker revolves around the high degree of correlation between the outcomes. In the analysis of multiple outcomes, it's typical to predict each one separately. When there are no missing values in the data (or any missing values occur randomly), analyzing each outcome individually can produce unbiased estimates for policy effects, even if the outcomes are correlated. Neglecting outcome correlations may result in imprecise estimation of the variance-covariance matrix of policy effects. This imprecision can lead to inaccuracies in estimated confidence intervals, subsequently increasing both Type I and Type II error rates.

A simple example illustrates the second concern. Consider the outcomes, the total number of sick leave days and the number of days within a single sickness episode are correlated and jointly normally distributed: $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ with means μ_1, μ_2, and standard deviations σ_1 and σ_2 for outcomes Y_1 and Y_2, respectively. Because the outcomes are correlated, it follows that the estimates of policy effects and their variances are also correlated.

Consider the standardized population policy effect between the group subjected to the policy and the control group:

$$\delta_j = \frac{\mu_j^{(P=1)} - \mu_j^{(P=0)}}{\sqrt{\sigma_{jj}}}, \ j = [1, 2].$$

$\mu_j^{(P=1)}$ and $\mu_j^{(P=0)}$ represent the population means for the group exposed to the policy ($P = 1$) and the control group ($P = 0$), respectively. I use σ_{jj} to denote the pooled population variance for the $j-$th outcome.

The corresponding standardized sample policy effect is given as

$$d_j = \frac{\bar{Y}_j^{(P=1)} - \bar{Y}_j^{(P=0)}}{\sqrt{s_{jj}}},$$

where $\bar{Y}_j^{(P=1)}$ and $\bar{Y}_j^{(P=0)}$ represent the sample means for the group exposed to the policy ($P = 1$) and the control group ($P = 0$), respectively. s_{jj} is the sample analogue of σ_{jj}.

Olkin and Gleser (2009) show that the sampling variance, and covariance between d_1 and d_2, are given as:

$$\sigma_{d_j} \approx \frac{N^{(P=1)} + N^{(P=0)}}{N^{(P=1)}N^{(P=0)}} + \frac{d_j^2}{2(N^{(P=1)} + N^{(P=0)})},$$

and

$$\sigma_{d_1 d_2} \approx \frac{N^{(P=1)} + N^{(P=0)}}{N^{(P=1)}N^{(P=0)}}\rho + \frac{d_1 d_2 \rho^2}{2(N^{(P=1)} + N^{(P=0)})}.$$

$N^{(P=1)}$ and $N^{(P=0)}$ denote the number of subjects in the exposed and control group, correspondingly. ρ is the correlation between the two outcomes. (1) shows that if two outcome variables are positively correlated, the covariance between standardized policy effects (d_1 and d_2) is also positive. Neglecting this connection between outcomes can lead to flawed statistical conclusions (Becker, 2000). To clarify, when we consider outcomes (and, in turn, policy effects) as unrelated, we disregard the shared information provided by each observed policy effect. This leads to underestimation of standard errors. As a result, the confidence intervals become excessively narrow, and the likelihood of making Type I errors increases when estimating and testing the policy effects. Therefore, Becker (2000) concludes that "No reviewer should ever ignore dependence among study outcomes. Even the most simplest ad hoc options are better than pretending such dependence does not exist.”

Another significant concern is the computational expense linked to modeling each outcome individually. As depicted in Figure 11, separate estimations can be notably costly, particularly when dealing with multiple outcomes.

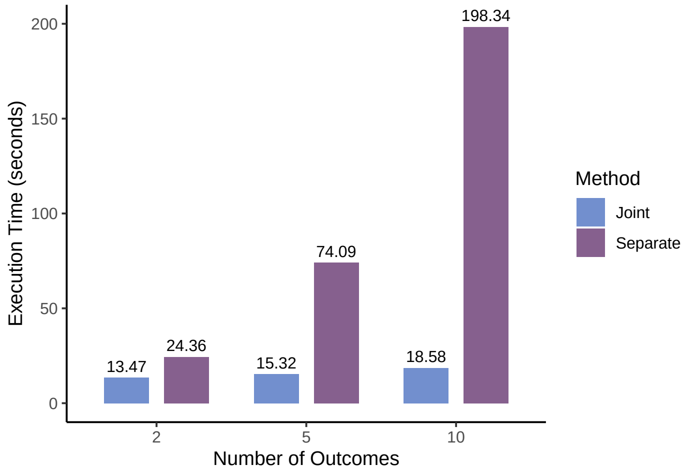


Figure 1: The time spent (in seconds) to predict the outcome jointly and separately. I build the random forest algorithm with 1000 trees in each approach.

Lastly, there’s a concern that certain individual characteristics might not significantly contribute to explaining the variations in policy effects. When this occurs, the estimates of policy effects become more uncertain, displaying greater variance. In classical random forest algorithms, the basic assumption is that each variable has a reasonable chance of affecting the estimation of policy effects, as noted in Athey and Wager (2019). This necessitates a substantial amount of data to accurately identify the most relevant variables. Nareklishvili et al. (2022) propose a dimensionality reduction method to aid the issue.

To address these concerns, I utilize partial least squares to identify the most pertinent characteristics before estimating variations in policy effects. The framework in this article is explicitly tailored to simultaneously address multiple outcomes and policies. A significant benefit of the multivariate approach lies in its simultaneous utilization of data from all outcomes. By leveraging information from multiple outcomes, we enhance the precision of policy effect estimates and maximize the precision of confidence intervals (Jackson et al., 2011). In this paper, I focus on two outcomes and a single binary policy.

4 Empirical Examples

The purpose of this section is to shed light on the empirical applications that serve as motivating examples for the proposed multivariate causal forest algorithm.

Example 1: Multiple Policy Arms and Multiple Outcomes

In the context of randomized controlled trials (RCTs), Athey and Imbens (2016) present a specific form of the linear regression model. This model describes heterogeneous policy effects as follows:

$$\text{outcome}_i = \theta(\mathbf{X}_i) \times \text{policy}_i + \mathbf{X}_i\theta + \text{error}_i. \tag{1}$$

Here, $\theta(\mathbf{X}_i)$ represents the effect of a binary policy on the continuous outcome for each person $i = 1, \ldots, N$. This effect varies based on individual characteristics $\mathbf{X}_i$. The vector-valued parameter θ signifies the impact of $\mathbf{X}_i$ on outcome. Additionally, error accounts for the residuals associated with the outcome and is independent of policy or individual characteristics.

The multivariate causal forest accommodates an extended version of (4) to encompass scenarios involving multiple policies and outcomes. In the case of $p = 1, \ldots, P$ binary policy arms, and a single continuous outcome, the model is defined as follows:

$$\text{outcome}_i = \mathbf{X}_i\theta + \theta_1(\mathbf{X}_i) \times \text{policy}_{1i} + \cdots + \theta_P(\mathbf{X}_i) \times \text{policy}_{Pi} + \text{error}_i.$$

The expression policy_{pi} serves as a binary indicator denoting whether an individual i pertains to policy arm p. Additionally, $\theta_p(\mathbf{X}_i)$ denotes the parameter associated with a $p-$th policy arm. This model not only accommodates multiple policy effects within a specific policy arm but also permits the grouping of policy effects across different arms. This flexibility allows for a more detailed examination of heterogeneous policy effects.

In scenarios involving multiple outcomes, assuming a researcher is dealing with M continuous outcome variables outcome_{mi} with $m = 1, \ldots, M$, the model for evaluating various policy outcomes can be expressed as:

$$\begin{aligned}\text{outcome}_i =& \mathbf{X}_i\theta + \theta_1(\mathbf{X}_i)1(\text{outcome}_i \in \text{outcome}_{1i}) \times \text{policy}_{1i} + \cdots + \\ & \theta_M(\mathbf{X}_i)1(\text{outcome}_i \in \text{outcome}_{Mi}) \times \text{policy}_{Mi} + \text{error}_i.\end{aligned}$$

The term $1(\text{outcome}_i \in \text{outcome}_{mi})$ serves as an indicator for each respective m-th outcome, where m ranges from 1 to M. The variables policy_{mi} and $\theta_m(\mathbf{X}_i)$ represent a binary indicator for a policy and its effect on the $m-$th continuous outcome, respectively.

The joint estimation approach in this paper offers the advantage of aggregating comparable policy effects. This facilitates the customization of policy interventions. Additionally, it enables researchers to investigate the correlation of policy effects across diverse outcome variables, an otherwise challenging task. The joint estimation in addition accommodates variations in policy effects across observable characteristics $\mathbf{X}_i$, ensuring the estimation of correct standard errors.

Example 2: Demand and Supply Analysis

Contemporary likelihood-based methodologies applied to the joint estimation of demand and supply systems encounter challenges in accommodating unobserved heterogeneity within the model framework. Notably, researchers have primarily directed their attention towards modeling common shocks, incorporating purchase histories and demographic variables, while only partially addressing the issue of consumer heterogeneity (Iyer and Villas-Boas, 2003; Draganska and Jain, 2002). Multivariate causal forest can be used to not only control for the individual characteristics but detect heterogeneities of multiple responses simultaneously.

Consider the analysis of the impact of ticket prices (denoted as the policy variable P_i) on air-travel demand (represented by the outcome variable Y_i) (Hartford et al., 2017):

$$\text{air travel demand}_i = f(\text{ticket price}_i, \mathbf{X}_i) + \varepsilon_i.$$

Here, $\mathbf{X}_i$ represents a vector of characteristics, including variables such as holidays. f denotes a nonlinear smooth function. It is essential to recognize that ticket prices are typically not set exogenously; rather, they tend to be influenced by unobservable factors, represented by ε_i. Consequently, adopting a simplistic approach that predicts airline demand solely as a function of price may yield misleading conclusions. A standard solution is to introduce an exogenous driver of prices, in this setting, fuel costs (Z_i, a binary indicator for high fuel costs), and simultaneously estimate a second equation:

$$\text{ticket price}_i = m(\text{fuel costs}_i, \mathbf{X}_i) + v_i,$$

where v_i is the corresponding error of price, and m is another nonlinear smooth function.

Another classic example is provided by Angrist and Evans (1996). The authors estimate the impact of childbearing on women's labor supply. To achieve this, they employ parental preferences for a mixed sibling-sex composition as an exogenous determinant of childbearing:

$$\begin{aligned} \text{hours worked women}_i &= f(\text{childbearing}_i, \mathbf{X}_i) + \varepsilon_i, \\ \text{childbearing}_i &= m(\text{mixed gender preference}_i, \mathbf{X}_i) + v_i. \end{aligned}$$

Within the framework of the multivariate causal forest, I embark on a simultaneous estimation process. Specifically, I analyze the heterogeneity of two distinct parameters simultaneously:

$$\begin{aligned}\theta_1 =& \mathbb{E}(Y_i|Z_i = 1, \mathbf{X}_i) - \mathbb{E}(Y_i|Z_i = 0, \mathbf{X}_i),\\ \theta_2 =& \mathbb{E}(P_i|Z_i = 1, \mathbf{X}_i) - \mathbb{E}(P_i|Z_i = 0, \mathbf{X}_i).\end{aligned}$$

As before, Y_i, P_i, Z_i, and $\mathbf{X}_i$ are defined as the outcome, policy, exogenous instrument, and independent characteristics. This approach enables the analysis of heterogeneity in Wald estimates $\hat{\theta}^{\text{Wald}}$ obtained according to the sample analogue of the ratio (Angrist and Imbens, 1995):

$$\theta^{\text{Wald}} = \frac{\mathbb{E}(Y_i|Z_i = 1, \mathbf{X}_i) - \mathbb{E}(Y_i|Z_i = 0, \mathbf{X}_i)}{\mathbb{E}(P_i|Z_i = 1, \mathbf{X}_i) - \mathbb{E}(P_i|Z_i = 0, \mathbf{X}_i)}.$$

This method shares similarities with the instrumental forest approach introduced by y Athey et al. (2019). However, a key distinction lies in the estimation strategy. Our approach enables the joint estimation of both the numerator and denominator, allowing us to incorporate their covariance into the model.

In this setting, the structural equation model consists of two outcomes. The setup could be generalized to include more than two outcomes (Ullman and Bentler, 2012). The multivariate causal forest framework opens up avenues for exploring various questions, such as estimating supply and demand effects of disability on labor force participation (Stern, 1996); demand analysis with many prices (Chernozhukov et al., 2019).

Example 3: Regression Discontinuity Design

A Regression Discontinuity Design (RDD) is a quasi-experimental methodology designed to estimate the causal effects of interventions. It involves the assignment of interventions based on a predetermined cutoff or threshold (c), where the assignment changes either above or below this threshold. For instance, students may be placed into educational programs according to placement scores. Such a score would represent an exogenously set cutoff. The classical parametric regression discontinuity design may be formulated by the following equation:

$$\text{outcome}_i = \theta_0 + \theta \times \text{policy}_i + C_i\theta + \varepsilon_i. \tag{2}$$

In this context, policy_i takes the value of one when $C_i \geq c$ and zero otherwise. C_i is a continuous variable determining the exposure to the policy, and Y_i is a continuous outcome variable ranging in $(-\infty, \infty)$.

The multivariate causal forest extends the traditional RDD framework to encompass multiple policies, enabling the investigation of variations in policy effects across individual characteristics. I extend (2) to accommodate $p = 1, \ldots, P$ interventions:

$$\text{outcome}_i = \theta_0 + \theta_1(\mathbf{X}_i) \times \text{policy}_{1i} + \cdots + \theta_P(\mathbf{X}_i) \times \text{policy}_{Pi} + C_i\theta + \varepsilon_i,$$

where policy_{pi} and $\theta_1(\mathbf{X}_i)$ represent a binary indicator for a $p-$th policy and its influence on the outcome, correspondingly.

5 Group Average Policy Effects

Let θ_g represent a vector-valued policy effect for a group g, which we seek to investigate formally [1]. We express this effect as the sum of the group average policy effect, denoted by $\mathbb{E}(\theta_g)$, and the randomness $\boldsymbol{\varepsilon}_g$ associated with group $g(\mathbf{X}_i)$ (and

[1] If the parameters and random variables are vector-valued, the operations apply coordinatewise.

uncorrelated with $\mathbb{E}(\theta_g)$):

$$\theta_g = \mathbb{E}(\theta_g) + \varepsilon_g. \tag{3}$$

The group average policy effect represents the disparity in expected outcomes, $\mathbf{E}(\mathbf{Y}_i)$, between two distinct scenarios: one where a specific group is subjected to a particular policy ($P_i = 1$), and another where the same group is not exposed to the policy ($P_i = 0$):

$$\mathbb{E}(\theta_g) = \mathbb{E}\big(\mathbf{Y}_i(1) - \mathbf{Y}_i(0) | g(\mathbf{X}_i)\big). \tag{4}$$

The groups in this paper are data-driven (unknown a priori), continuous, and uncorrelated. For example, if the group consists of two components, then we define the group as $g(\mathbf{X}_i) = [Z_{i1}, Z_{i2}]$, where Z_{ik} for $k = 1, 2$ is a k-th component for an individual i. These components may take any value between $[-\infty, +\infty]$. For example, consider two distinct components:

$$\left\{ \begin{array}{c} [\mathbf{3.50}, \mathbf{4.27}, \mathbf{3.15}, \mathbf{5.25}, 0.33, 0.11, 0.16, -0.71] \\ [0.05, 0.10, -0.25, 0.20, \mathbf{7.22}, \mathbf{8.14}, \mathbf{6.26}, \mathbf{5.28}]. \end{array} \right\}$$

In this scenario, a clear partition emerges, dividing the components into two distinct groups: Group 1 with high values of the first component, and Group 2 corresponding to high values of the second component.

$$\text{Group 1} \sim \left\{ \begin{array}{c} [\mathbf{3.50}, 0.05] \\ [\mathbf{4.27}, 0.10] \\ [\mathbf{3.15}, -0.25] \\ [\mathbf{5.25}, 0.20] \end{array} \right\}, \text{Group 2} \sim \left\{ \begin{array}{c} [0.33, \mathbf{7.22}] \\ [0.11, \mathbf{8.14}] \\ [0.16, \mathbf{6.26}] \\ [-0.71, \mathbf{5.28}]. \end{array} \right\}.$$

To identify specific groups that correspond to each component, I regress these components on individual characteristics. Assume, high values of the first component are strongly indicative of older individuals, while high values of the second component are associated with females. Consequently, the first group comprises older males, while the second group comprises young females. Even though the components within each group are uncorrelated, there exists a slight overlap between them, attributable to their continuous nature.

Figure 2 illustrates the continuous nature of the two groups. In particular, it demonstrates that young females are more inclined to have high values of the second component, while older males are more commonly associated with high values of the first component.

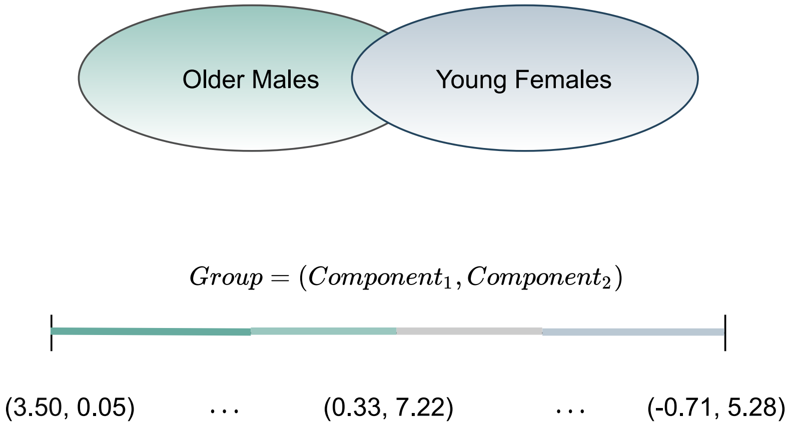


Figure 2: Visualisation of two data-driven groups (older males and young females) that consist of continuous components.

This article revolves around two primary research goals. First, it centers on identifying and forming economically meaningful population groups adaptively. Second, it aims to systematically characterize the most diversified components within the identified groups of the population.

5.1 Identification of Data-Driven Groups of the Population

To ensure the formation of well-defined and meaningful groups, I posit the following assumption:

Assumption 5.1 (group cardinality)**.** *The dimension of the space of latent groups, denoted as dim*$(\mathcal{G})$*, is weakly lower than the dimensionality of the space of the original covariates, denoted as dim*$(\mathcal{X})$*, i.e.,* $\dim(\mathcal{G}) \leq \dim(\mathcal{X})$*.*

By imposing Assumption 5.1, we ensure that the process of subgroup formation remains feasible and interpretable in high-dimensional scenarios. Note that, potential outcomes in (4) remain unobserved, as each group is exclusively exposed to either the policy intervention or the control condition, but never both simultaneously. To estimate group average policy effects, I assume, the policymaker governs the data-generating process through a randomized controlled trial:

Assumption 5.2 (group unconfoundedness)**.** *The data are independently and identically distributed (i.i.d), and conditional on the population groups, the policy is independent of the potential outcomes*

$$\mathbf{Y}_i(1), \mathbf{Y}_i(0) \perp\!\!\!\perp P_i | g(\mathbf{X}_i).$$

Here, $\mathbf{Y}_i(1), \mathbf{Y}_i(0) \in \mathbb{R}^M$ represent potential outcomes. Assumption 5.2 is not inherently stronger than the classical unconfoundedness assumption $\big(\mathbf{Y}_i(1), \mathbf{Y}_i(0) \perp P_i \mid \mathbf{X}_i\big)$ when g does not provide any additional information beyond characteristics $\mathbf{X}_i \in \mathbb{R}^D$. [2]

Due to Assumption 5.2, the group average policy effect is given as

$$\mathbb{E}(\boldsymbol{\theta}_g) = \mathbb{E}\big(\mathbf{Y}_i | P_i = 1, g(\mathbf{X}_i)\big) - \mathbb{E}\big(\mathbf{Y}_i | P_i = 0, g(\mathbf{X}_i)\big). \tag{5}$$

I follow the approach of Nareklishvili et al. (2022) and seek to identify the linear combinations of the characteristics, labeled as target components, that explain the highest variation in $\mathbf{X}_i$ as well as the outcome $\mathbf{Y}_i$. According to Helland (1990, 2014), the vectors of characteristics and outcomes, $\mathbf{X}_i$ and $\mathbf{Y}_i$ respectively, can be

[2] $g : \mathbf{X}_i \mapsto g(\mathbf{X}_i)$ may be an affine transformation that preserves information inherent to $\mathbf{X}_i$. In that case, the classical unconfoundedness assumption implies Assumption 5.2. See Rosenbaum and Rubin (1983) for further details.

decomposed into the latent, unobserved structures. In an iterative manner, the target component that captures the most pronounced variations within the observable characteristics $\mathbf{X}_i$ also unveils the latent structure governing the most significant fluctuations in the outcome $\mathbf{Y}_i$. This approach is known as the partial least squares algorithm.

5.2 Geometry of Latent Groups

Overall, the model for multivariate partial least squares involves decomposing the input and output space into latent structures:

$$\underbrace{\mathbf{X}}_{N\times D} = \underbrace{\mathbf{Z}}_{N\times G} \times \underbrace{\mathbf{V}^T}_{G\times D} + \underbrace{\mathbf{E}}_{N\times D},$$
$$\underbrace{\mathbf{Y}}_{N\times M} = \underbrace{\mathbf{R}}_{N\times G} \times \underbrace{\mathbf{Q}^T}_{G\times M} + \underbrace{\mathbf{F}}_{N\times M}.$$

$\mathbf{X} \in \mathbb{R}^{N\times D}$ is a matrix of individual characteristics.

$\mathbf{Y} \in \mathbb{R}^{N\times M}$ is a matrix of outcomes.

$\mathbf{Z} \in \mathbb{R}^{N\times G}$ represent the matrix of latent components that make up groups $g(\mathbf{X})$.

$\mathbf{R} \in \mathbb{R}^{N\times G}$ is an output score.

$\mathbf{V}$ and $\mathbf{Q}$ denote $D \times G$ and $M \times G$ loading matrices (weights).

$\mathbf{E}$ and $\mathbf{F}$ represent errors of independent characteristics, and outcomes, respectively.

The fundamental concept behind the identification of the first partial least squares component involves the discovery of unit vectors in two spaces: v in the input space $\mathbb{R}^D$ and w in the output space $\mathbb{R}^M$. The goal is to maximize the product of the projection of an input vector onto the unit vector

$$\underbrace{\mathbf{Z}_1}_{N\times 1} = \underbrace{\mathbf{X}}_{N\times D}\underbrace{v}_{D\times 1}$$

and the projection of an output vector onto its corresponding unit vector

$$\underbrace{\mathbf{R}_1}_{N\times 1} = \underbrace{\mathbf{Y}}_{N\times M}\underbrace{w}_{M\times 1}$$

.

Formally,

$$max_{(v,w)}\mathbb{E}(\mathbf{Z}_1^T \cdot \mathbf{R}_1) = max_{(v,w)}\mathbb{E}(v^T\mathbf{X}^T \cdot \mathbf{Y}w) = max_{(v,w)}v^T\mathbb{E}(\mathbf{X}^T\mathbf{Y})w. \tag{6}$$

The vectors $|v| = 1$ and $|w| = 1$ represent unit vectors, each with a length of one. $C_{XY} = \mathbb{E}(\mathbf{X}^T\mathbf{Y})$ signifies the covariance between the mean-centered random variables $\mathbf{X}$ and $\mathbf{Y}$. The term $\mathbf{Z}_1$ represents the first component that constitutes the data-driven groups in this context. $\mathbf{R}_1$ is known as the output score.

To determine the directions represented by vectors v and w, we can employ the Lagrange multiplier method and solve for them using the following optimization problem:

$$(v^0, w^0) = argmax_{(v,w)}\{v^T C_{XY} w - \frac{1}{2}\lambda_v(v^T v - 1) - \frac{1}{2}\lambda_w(w^T w - 1)\}. \tag{7}$$

In this equation, λ_v and λ_w are Lagrange multipliers. After solving for the unit directions in (7), we end up with

$$C_{XY}C_{YX}v = \lambda_v\lambda_w v,$$
$$C_{YX}C_{XY}w = \lambda_v\lambda_w w.$$

This implies that vector v is an eigenvector of matrix $C_{XY}C_{YX}$, and vector w is an eigenvector of matrix $C_{YX}C_{XY}$. It can be shown that $\lambda_v = \lambda_w = \lambda$. Figure 3 presents the identification of the first component visually.

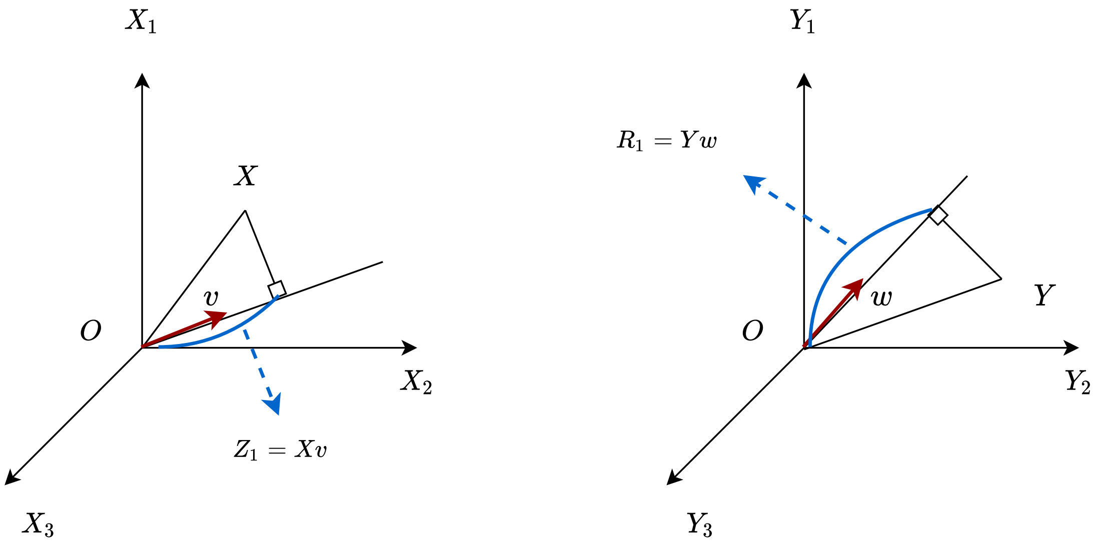


Figure 3: Identification of the first latent group (factor) by the partial least squares algorithm.

The subsequent step involves predicting the input and output as a function of the first component and obtaining residuals:

$$\mathbf{X}^{(1)} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{Z}_1\hat{\beta}_x^T \tag{8}$$

$$\mathbf{Y}^{(1)} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Z}_1\hat{\beta}_y^T. \tag{9}$$

Here, $\hat{\beta}_x$ and $\hat{\beta}_y$ are $D \times 1$ and $M \times 1$ vectors of coefficients obtained through ordinary least squares regression of the input and outcomes on the first component $\mathbf{Z}_1$, respectively. To identify the second component, we iterate the procedure, replacing the original characteristics and outcomes with residualized characteristics $\mathbf{X}^{(1)}$ and outcomes $\mathbf{Y}^{(1)}$. Subsequent components are obtained similarly. In this paper, cross-validation yields the optimal number of components.

5.3 Estimation of Heterogeneous Group Average Policy Effects

Following the categorization of the population into groups, the main objective is to compute the group average policy effects. The application of the partial least squares procedure is specifically designed for multiple outcomes. The formed groups not only encapsulate characteristics pertinent to each outcome but also capture shared information between the two outcomes. Consequently, I propose a causal forest algorithm designed for multiple outcomes.

The approach involves recursively partitioning the space of identified groups, denoted as $\mathcal{G}$, using axis-aligned splits. The primary objective is to estimate the vector of policy effects within each partition. An axis-aligned split is defined as a pair (j, γ), where $j = 1, \ldots, G$ represents a specific group (*splitting coordinate*), and $\gamma \in \mathbb{R}$ is the value of this group (*splitting index*).

Let $\mathcal{P}^{(0)} = \mathcal{G} \in \mathbb{R}^G$ represent the *parent node* of the tree. We strategically select a splitting coordinate $j : 1 \leq j \leq G$ and a splitting index γ to divide $\mathcal{P}^{(0)}$ into two non-overlapping rectangles, denoted as *child nodes*:

$$\mathcal{P}^{(1,1)} = \mathcal{P}^{(0)} \cap \{\widetilde{\gamma}_j \in \mathcal{P}^{(0)} : \widetilde{\gamma}_j \leq \gamma\} \text{ and } \mathcal{P}^{(1,2)} = \mathcal{P}^{(0)} \cap \{\widetilde{\gamma}_j \in \mathcal{P}^{(0)} : \widetilde{\gamma}_j > \gamma\}, \tag{10}$$

where $\widetilde{\gamma}_j$ represents the j-th coordinate of the vector $g(\mathbf{X}_i)$ derived from the training data. Subsequently, this sequential splitting process is iteratively applied to the resulting child nodes, $\mathcal{P}^{(1,1)}$ and $\mathcal{P}^{(1,2)}$, until a specified stopping criterion is met. Crucially, each split is based on data that belong to the corresponding partition, ensuring data-driven decision-making at each step. Figure 4 illustrates the process graphically.

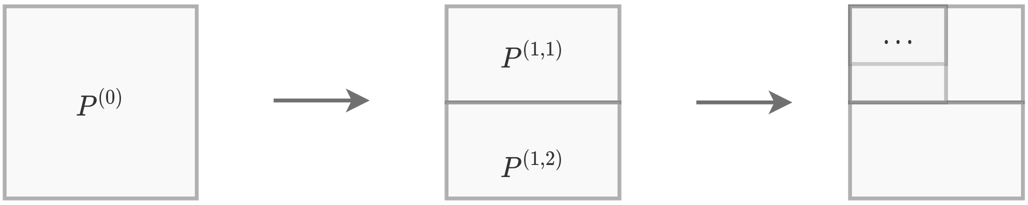


Figure 4: The splitting process of a tree.

The sequence of k splits induces the partition of $\mathcal{G}$ which we denote by Π. This partition consists of a collection of non-overlapping rectangular regions ℓ, referred to as leaves or *terminal nodes* in the tree structure. The amalgamation of these partitions constitutes a latent group space, mathematically expressed as follows:

$$\Pi = \{\ell_1, \ell_2, \ldots, \ell_{|\Pi|}\} \text{ and } \cup_{n=1}^{|\Pi|} \ell_n = \mathcal{G}.$$

The underlying split process ensures that every element within $\mathcal{G}$ is precisely allocated to one of these partitions.

Athey and Imbens (2016) propose a novel approach to estimate heterogeneous policy effects under the assumption of unconfoundedness. Their method involves a two-sample split procedure, with a training sample, denoted as S^{tr}, used to identify and construct the splitting variables and corresponding values. Subsequently, they utilize an estimation sample, denoted as S^{est}, to compute policy effects across various segments of latent groups in the context of interest. In our framework, the unbiased sample analogue of the expectation of $\boldsymbol{\theta}_g$ is given as:

$$\widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \Pi) = \sum_{n=1}^{|\Pi|} \Bigg(\mathbf{1}(\gamma \in \ell_n, P_i = 1) \frac{1}{|i : g(\mathbf{X}_i) \in \ell_n, P_i = 1|} \sum_{i:g(\mathbf{X}_i) \in \ell_n} \mathbf{Y}_i(1) - \\ \mathbf{1}(\gamma \in \ell_n, P_i = 0) \frac{1}{|i : g(\mathbf{X}_i) \in \ell_n, P_i = 0|} \sum_{i:g(\mathbf{X}_i) \in \ell_n} \mathbf{Y}_i(0) \Bigg), \tag{11}$$

where $|\Pi|$ is the total number of the terminal nodes. $\mathbf{1}(\gamma \in \ell_n, P_i = p)$ is a binary variable and equals one when, for a given $p \in \{0, 1\}$, a generic test data point γ belongs to a terminal leaf ℓ_n and zero otherwise. $\mathbf{Y}_i(p)$ denotes an $M-$dimensional outcome variable.

6 Asymptotic Properties

A conventional method for demonstrating the asymptotic properties of trees and forests entails the use of classical trees and forests tailored to predict the outcome variable (Wager, 2014; Wager and Athey, 2018). I follow the same strategy in our investigation and extend the properties to policy effects afterward.

Let $A_i = (\mathbf{Y}_i, g(\mathbf{X}_i))_{i=1}^{N}$ represent the dataset without any policy information. We consider a prediction of a classical tree for a given individual with the latent group value of γ. The prediction is obtained by averaging the vector-valued outcome within each terminal leaf (node):

$$T(\gamma, \xi, A_1, \ldots, A_N) = \sum_{n=1}^{|\Pi|} \mathbf{1}(\gamma \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i:g(\mathbf{X}_i)\in\ell_n} \mathbf{Y}_i.$$

Here, ξ represents an external source of randomization, facilitating randomized split selection procedures during the construction of a tree. The function $\mathbf{1}(\gamma \in \ell_n)$ acts as an indicator, equaling one if the point γ belongs to the terminal leaf ℓ_n, and zero otherwise. The term N_{ℓ_n} denotes the number of observations in the terminal node ℓ_n. In essence, a tree $T(\gamma, \xi, A_1, \ldots, A_N)$ provides a prediction at the point γ based on the dataset $\{A_i\}_{i=1}^{N}$ and a randomization parameter ξ. For a more comprehensive and intuitive understanding of classification and regression trees, detailed insights can be found in the works of Lewis (2000) and Kingsford and Salzberg (2008).

Trees are known for their ease of interpretation and implementation, as well as their robustness against outliers and missing data. However, they suffer from high variance, instability, and a tendency to overfit the training data, making it challenging to determine the optimal tree structure. To address these limitations, the random forest algorithm was introduced by Breiman (2001). In the context of random forests, let $s < N$ be a subset of size s sampled from a population indexed by $i = \{1, \ldots, N\}$. The value of s is typically set as $s = N^{\beta}$, with β being chosen sufficiently close to 1 (Wager and Athey, 2018). Following Breiman (2001) and Wager and Athey (2018), the random forest estimator is defined as the average of the individual tree estimators, aggregated over all possible size-s subsamples of the training data, while also taking into account the auxiliary noise ξ. Specifically, the prediction of a random forest estimator for a specific individual (with a group value of γ) is defined as

$$\mathcal{F}(\gamma, A_1, \ldots, A_N) = \frac{1}{\binom{N}{s}} \sum_{1 \leq i_1 \leq \cdots \leq i_s \leq N} \mathbb{E}_{\xi} T(\gamma, \xi, A_{i_1}, \ldots, A_{i_s}), \tag{12}$$

where $i_1, \ldots, i_s$ are the size-s subsamples of the population $\{i = 1, \ldots, N\}$. $\binom{N}{s}$ denotes the binomial coefficient representing the number of ways to choose s subsamples from a population of size N. The prediction is the average of the expected predictions over all possible combinations of these subsamples.

In practice, we estimate a random forest using Monte Carlo averaging, which can be expressed as follows:

$$\mathcal{F}(\gamma, A_1, \ldots, A_N) \approx \frac{1}{B} \sum_{b=1}^{B} T(\gamma, \xi^*, A_1^*, \ldots A_N^*). \tag{13}$$

Here, the sets $\{A_1^*, \ldots A_N^*\}$ are drawn without replacement from the original dataset $\{A_1, \ldots A_N\}$. The parameter B denotes the number of sub-samples considered for the averaging process. The output of $\mathcal{F}(\gamma, A_1, \ldots, A_N)$ is a $1 \times M$ vector for a given individual, and all the arithmetic operations mentioned in this context are defined coordinate-wise in $\mathbb{R}^M$.

6.1 Assumptions

Random forests for group average policy effects rest on similar assumptions introduced by Wager and Athey (2018). The first assumption to impose is the ”honesty” of a tree.

Assumption 6.1 (Honesty)**.** *Given the identified latent groups $g(\mathbf{X}_i)$, we assert that the outcome variable Y_{im} and the splitting parameters (i.e. the splitting coordinates and splitting indices denoted by (j, γ)) are statistically independent of each other. Specifically, for each individual i where the outcome Y_{im} contributes to the final prediction:*

$$F(Y_{im}|g(\mathbf{X}_i), (j, \gamma)) = F(Y_{im}|g(\mathbf{X}_i)).$$

F represents the probability density function associated with the respective m-th outcome variable, with m ranging from 1 to M.

This requirement can be addressed through various approaches. In this article, inspired by the work of Athey and Imbens (2016), I adopt a method involving the division of the dataset into two distinct partitions: a training set (S^{tr}) and an estimation set (S^{est}). Observations in S^{tr} and the groups $g(\mathbf{X}_i) \in S^{est}$ determine the optimal splitting coordinates and indices for constructing the decision trees. Subsequently, the predicted outcomes are generated based on the estimation sample S^{est}.

Another fundamental assumption in this article concerns the data-generating process.

Assumption 6.2 (Data Generating Process 1)**.** *Let $\mathbf{Y}_i = f(b_0 + P_i\theta(g(\mathbf{X}_i)) + \mathbf{X}_i\mathbf{b}) + \varepsilon_i$ where $\mathbf{b}$ is a $D \times 1$ vector of coefficients, b_0 is a constant and f is an infinitely differentiable non-linear mapping. Assume, $\mathbf{X}_i$ have a joint Elliptical distribution with the mean $\mu_{\mathbf{X}}$ and a variance $\Sigma_{\mathbf{XX}}$. Let $\mathbf{X}_i$ be independent of ε_i. Moreover, let S_{xx} and s_{xy} converge in probability to $\Sigma_{\mathbf{XX}}$ (the population variance of $\mathbf{X}_i$) and $\sigma_{\mathbf{Xy}}$ (the population covariance of $\mathbf{X}_i$ and $\mathbf{Y}_i$) when $N \to \infty$. Moreover, let there exist a pair of eigenvectors and eigenvalues (v_j, λ_j) for which $\sigma_{\mathbf{Xy}} = \sum_{j=1}^{M} \lambda_j v_j$ (with λ_j non-zero for each $j = 1, \ldots, M$). Assume also $\mathbb{E}(|f(U_i)|) < \infty$ and $\mathbb{E}(U_i|f(U_i)|) < \infty$ with $U_i = b_0 + \mathbf{X}_i\mathbf{b}$ and $q = M$.*

According to Assumption 6.2, the relationship between the response variable and the independent characteristics adheres to a predetermined functional form. Moreover, the subject characteristics follow an elliptical distribution, exhibiting an ellipse-like shape in a multi-dimensional coordinate system. Although this assumption may not always hold in real-world scenarios, empirical evidence has demonstrated that the results obtained under this assumption are not significantly divergent from those obtained when the features have alternative types of distributions (see Brillinger, 2012). Under Assumption 6.2, Nareklishvili et al. (2022) show that the partial least squares estimator is consistent up to a proportionality constant. The proof relies on the analytic solution of partial least squares weights proposed by Helland (1990) and Stein's lemma discussed by Brillinger (2012).

Assumption 6.3 (Data Generating Process 2)**.** *The identified latent groups $g(\mathbf{X}_i)$ are supported on the unit cube $g(\mathbf{X}_i) \in [0,1]^G$, with a density that is bounded away from 0 and ∞. First and second moments, $\mathbb{E}(Y_{im}|g(\mathbf{X}_i) = \gamma)$, $\mathbb{E}\big((Y_{im})^2|g(\mathbf{X}_i) = \gamma\big)$, are Lipschitz-continuous for each m-th outcome variable, respectively ($m = 1, \ldots, M$). Furthermore, $Var(Y_{im}|g(\mathbf{X}_i) = \gamma)$ is bounded away from 0 (i.e., $inf_{\gamma \in \mathcal{G}} Var(Y_{im}|g(\mathbf{X}_i) = \gamma) > 0$).*

Lipschitz-continuity and bounded variances represent widely used assumptions in the literature (Wager and Athey, 2018; Biau, 2012). The outcomes of the paper are not explicitly contingent on the distributional assumptions of $g(\mathbf{X}_i)$, but they affect the constants employed throughout the study (Lemma 2 and Theorem 3 in Section 3.2 in Wager and Athey, 2018).

Assumption 6.4 (Random Split Trees)**.** *At each recursive step, the probability that the next split occurs at j-th group is bounded below by π/d for $\pi \in (0,1]$, for all $j = 1, \ldots, G$.*

According to the works by Meinshausen and Ridgeway (2006) and Wager and Athey (2018), we introduce Assumption 6.4, which guarantees that during each recursive step of the tree-building process, each identified group is chosen with a probability of at least π/d for some $0 < \pi \leq 1$ at every splitting step.

Assumption 6.5 (The Splitting Algorithm is (α, k)-regular)**.** *There exists $\alpha > 0$, where each split leaves at least a fraction α of the available training examples on each side of the split, and moreover, the splitting ceases at a node when it contains less than k observations for some $k \in \mathbb{N}$.*

Assumption 6.5 guarantees that each partitioned half-space contains a sufficient number of observations (individuals). It has been demonstrated by Wager and Walther (2015) that under this assumption, the half-spaces formed by the algorithm are also large in Euclidean volume. Moreover, the assumption imposes an upper bound on the number of observations in terminal nodes, leading to fully grown trees of depth k, where each terminal node contains between k and $2k - 1$ observations. Consequently, this property places an upper limit on the variance of the tree estimator at any given group value γ.

Assumption 6.6 (Overlap). *We assume that for some* $0 < \epsilon < 1$ *and all* $\gamma \in [0, 1]^G$*:*

$$\epsilon < \mathbb{P}(P_i = 1 | g(\mathbf{X}_i) = \gamma) < 1 - \epsilon.$$

Assumption 6.6 guarantees that for large enough N, there will be enough individuals with and without the policy intervention.

6.2 Consistency and Asymptotic Normality

A random forest estimator can be viewed as a U-statistic, a concept introduced by Hoeffding (1961) and further developed in the theory of statistics (Korolyuk and Borovskich, 2013). The Höeffding decomposition, also known as the Hajek projection, is described by Hájek (1968) and der Vaart (1998) in the univariate case. To explore the large sample theory of random forests within a multivariate framework, I extend the fundamental properties of the Höeffding decomposition to encompass multiple outcomes. An established method for exploring the large sample theory of random forests involves determining the lower bound of its Höeffding decomposition. In line with this conventional approach, I pursue the same strategy in this section.

Consider a vector-valued function denoted as $T \in \mathbb{R}^M$. This function is characterized by being measurable and permutation symmetric, where the latter property implies that $T(\pi\gamma) = T(\gamma)$ holds true for all $\pi \in \mathit{\Pi}$ (a tree in this context). The

Hajek projection of this function is defined as follows:

$$\dot{T} = \mathbb{E}(T) + \sum_{i=1}^{N} \left[\mathbb{E}(T|g(\mathbf{X}_i)) - \mathbb{E}(T) \right] = \sum_{i=1}^{N} \mathbb{E}(T|g(\mathbf{X}_i)) - (N-1)\mathbb{E}(T). \tag{14}$$

Intuitively, (14) represents a projection of the vector-valued function T onto the linear subspace encompassing all random variables of the form $\sum_{i=1}^{N} f_i(g(\mathbf{X}_i))$, where $f_i : \mathbb{R}^G \mapsto \mathbb{R}$ are arbitrary measurable functions satisfying $\mathbb{E}(f_i^2(g(\mathbf{X}_i)) < \infty$ for $i = 1, \ldots, N$. This projection ensures that the conditional expectation of $\dot{T}$ in (14) coincides with the conditional expectation of T, denoted as:

$$\begin{aligned} \mathbb{E}(\dot{T}|f_i(g(\mathbf{X}_i))) &= \mathbb{E}(T|f_i(g(\mathbf{X}_i))), \text{ and} \\ \mathbb{E}(\dot{T}) &= \mathbb{E}(T). \end{aligned} \tag{15}$$

Now consider a vector-valued random forest estimator denoted as $\mathcal{F}(\gamma, A_1, \ldots, A_N) \in \mathbb{R}^M$, with a corresponding vector of means μ. Let $\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N)$ represent the Hajek projection of this multivariate random forest estimator, and let Σ denote the covariance matrix of the Hajek projection. It is important to note that the trees in $\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N)$ are symmetric, and the observations are independently and identically distributed (i.i.d) as before. Under these conditions, Lemma 6.1 holds.

Lemma 6.1. *The Hajek projection, denoted as $\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N)$, is given by the expression:*

$$\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N) - \mu = \frac{s}{N} \sum_{i=1}^{N} \left(T_1(A_i) - \mu \right),$$

where $\dot{T} = \sum_{i=1}^{s} T_1(A_i)$ with $T_1(a) = \mathbb{E}_{\xi, A_2, \ldots, A_N} T(\gamma, \xi, a, A_2, \ldots, A_N)$ represents the Hajek projection of a tree $T(\gamma, A_1, \ldots, A_N) = \mathbb{E}_{\xi} T(\gamma, \xi, A_1, \ldots, A_N) \in \mathbb{R}^M$. The parameter $s = N^{\beta}$ as before, and M represents the dimension of the outcome variables in each *terminal node.*

The covariance matrix Σ of the Hajek projection is given by:

$$\Sigma = \frac{s}{N}\mathbb{V}(\dot{T}) \in \mathbb{R}^{M \times M},$$

where $\mathbb{V}$ denotes the covariance matrix of the projected elements of the tree.

Proof. See Appendix A.0.1. □

In Appendix A.0.1, Figure 16 illustrates the projection of random forests onto the subspace defined by variables of the form $\sum_{i=1}^{N} f_i(g(\mathbf{X}_i))$. The projection meets the required conditions for the Lindeberg central limit theorem (Billingsley, 2013; DiCiccio and Romano, 2022), therefore, the Hajek projection of the multivariate random forest estimator is asymptotically normally distributed:

$$\Sigma^{-1/2}\big(\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N) - \mu\big) \xrightarrow{d} \mathcal{N}(0, I_{M \times M}),$$

where 0 is a $\mathbb{R}^M$ vector of zeros and $I_{M \times M}$ is an identity matrix.

To establish the asymptotic normality of the multivariate random forest estimator, we introduce an insightful relationship between the estimator and its projection by adding and subtracting $\Sigma^{-1/2}\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N)$ into the expression for $\Sigma^{-1/2}\big(\mathcal{F}(\gamma, A_1, \ldots, A_N) - \mu\big)$. This manipulation leads to the following decomposition:

$$\Sigma^{-1/2}\big(\mathcal{F}(\gamma, A_1, \ldots, A_N) - \mu\big) = \Sigma^{-1/2}\big(\mathcal{F}(\gamma, A_1, \ldots, A_N) - \dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N)\big) + \Sigma^{-1/2}\big(\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N) - \mu\big).$$

Formally, the objective of this article is to show that:

$$\Sigma^{-1/2}\big(\mathcal{F}(\gamma, A_1, \dots, A_N) - \dot{\mathcal{F}}(\gamma, A_1, \dots, A_N)\big) \xrightarrow{p} 0.$$

Then by Slutsky's theorem, the multivariate random forest estimator is asymptotically normally distributed.

In line with Wager and Athey (2018), I derive the lower bound of the variance of a vector-valued forest $\dot{\mathcal{F}}(\gamma, A_1, \dots, A_N)$ and demonstrate its' convergence to zero. The primary focus lies in proving the convergence in squared mean of the expression $\Sigma^{-1/2}\big(\mathcal{F}(\gamma, A_1, \dots, A_N) - \dot{\mathcal{F}}(\gamma, A_1, \dots, A_N)\big)$. For the sake of brevity and clarity, we shall use the more concise notations $\mathcal{F}$ and $\dot{\mathcal{F}}$ to represent $\mathcal{F}(\gamma, A_1, \dots, A_N)$ and $\dot{\mathcal{F}}(\gamma, A_1, \dots, A_N)$, respectively. Lemma 6.2 obtains the upper bound of the squared deviation between the forest and its' Hajek projection.

Lemma 6.2. *The mean squared difference of $\mathcal{F}$ and $\dot{\mathcal{F}}$ has the upper bound:*

$$\mathbb{E}\big(\mathcal{F} - \dot{\mathcal{F}}\big)^T \Sigma^{-1}\big(\mathcal{F} - \dot{\mathcal{F}}\big) \leq \frac{s}{N} tr\Big(\big(\mathbb{V}(\dot{T})\big)^{-1}\mathbb{V}(T)\Big),$$

where tr is a trace operator, and $\mathbb{V}(T)$ and $\mathbb{V}(\dot{T})$ denote the variance of a multivariate tree and its' Hajek projection, respectively.

Proof. See Appendix A.0.2. □

Under Assumptions 6.1-6.6, Theorem 6.1 shows that $\frac{s}{N} tr\Big(\big(\mathbb{V}(\dot{T})\big)^{-1}\mathbb{V}(T)\Big)$ approaches zero in the limit.

Theorem 6.1. *The entries of $\mathbb{V}(T)$ are bounded and its diagonal elements are bounded away from zero. Moreover, the lower bound of the off-diagonal terms of $\mathbb{V}(\dot{T})$ are on the order of $o\big(\frac{1}{\log^p(s)}\big)$. The upper bound in Lemma 6.2 converges to zero in the limit:*

$$\frac{s}{N} tr\Big(\big(\mathbb{V}(\dot{T})\big)^{-1}\mathbb{V}(T)\Big) \to 0.$$

Proof. See Appendix A.0.3. □

By Slutsky's theorem, Theorem 6.1 implies that the multivariate random forest estimator is asymptotically normally distributed. Appendices A.0.4, A.0.5, and A.0.6 generalize the proofs to accommodate correlated parameters across groups (leaves) of the population, quantiles of the outcome, and group average policy effects, respectively.

7 Estimation and Inference

Estimation and inference of group average policy effects rely on the presence of conditional moment functions within each subset of the identified group space. We define these population conditional moment functions as follows:

$$\mathbb{E}\left[\left(\rho(A_i, \theta_g^{\ell})\right) | g(\mathbf{X}_i) = \gamma\right] = 0. \tag{16}$$

Here, the vector-valued parameter θ_g^{ℓ} belongs to the ℓ-th partition. Parameters can exhibit heterogeneity across a smaller subset of the latent group space denoted as $\mathcal{Z}$ (which coincides with $\mathcal{G}$ in this article). The first assumption we make is the existence of a solution within each ℓ-th subset.

Assumption 7.1. *For all $\gamma \in \mathcal{G}$, the conditional expectation $\mathbb{E}\left[\left(\rho(A_i, \theta_g^{\ell})\right) | g(\mathbf{X}_i)\right]$ is bounded, and its supremum norm* [3] *converges to zero as the sample size increases, i.e.,*

$$\sup_{\gamma \in \mathcal{G}} ||\mathbb{E}\left[\left(\rho(A_i, \theta_g^{\ell}\right) | g(\mathbf{X}_i) = \gamma\right]|| = o(1).$$

.

This assumption ensures that the error in each partition remains bounded and does not grow infinitely large with increasing data.

[3] Also labeled as Forbeius norm with $||A||_F = \sqrt{tr(AA^T)}$. Here, A^T is the transpose of A.

Assumption 7.2. *For each subset ℓ, there exists a matrix $\Omega(X_i) \in \mathbb{R}^{p\times p}$ such that the eigenvalues of $\Omega(X_i)$ are uniformly bounded by a constant λ and*

$$\mathbb{E}\left[\Omega(X_i)\frac{\partial\rho(A_i,\theta_g^{\ell})}{\partial\theta_g^{\ell}}\right] \tag{17}$$

is strictly positive definite.

Assumption 7.2 ensures that the covariance structure of the parameters is invertible with non-zero eigenvalues. Let Assumptions 7.1 and 7.2 hold. Denote Σ as the covariance matrix of $\widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \Pi)$. We aim to minimize the discrepancy between the group-level policy effect, represented by θ_g, and its expected value (see also Athey and Imbens, 2016).

$$\mathbb{E}_{S^{tr},S^{est}}\left[\left(\theta_g - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \Pi)\right)^T \Sigma^{-1}\left(\theta_g - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \Pi)\right)\right]. \tag{18}$$

Algebraic transformations (as shown in Appendix A.0.7) lead to the unbiased empirical analogue that we maximize at each split of the population group space:

$$\hat{\theta}(\gamma, S^{est}, \Pi) = \arg\max_{\widetilde{\theta}} \frac{1}{N^{tr}} \sum_{\ell} N_{\ell}^{tr}\left(\widetilde{\theta}(g(\mathbf{X}_i), \Pi)^T \widehat{\Sigma}^{-1} \widetilde{\theta}(g(\mathbf{X}_i), \Pi) | g(\mathbf{X}_i) = \gamma\right), \tag{19}$$

where the covariance matrix can be estimated as $\hat{\Sigma} = \hat{\Sigma}\big(\tilde{\theta}(g(\mathbf{X}_i), S^{tr}, \Pi) | N^{est}\big)$. In this article, training and estimation samples have an equal number of observations, $N^{tr} = N^{est}$.

Inference of policy effects relies on the bootstrap method. Let $b = 1, \ldots, B$ be the b-th bootstrapped sample. We use a tree Π_b and the corresponding estimation sample S_b^{est} to obtain $\hat{\theta}(\gamma, S_b^{est}, \Pi_b)$ for a generic individual with a group value of γ. Next, the average of the individual tree estimates is $\hat{\theta}\big(\gamma, \{S_b^{est}\}_{b=1}^{B}\}, \{\Pi_b\}_{b=1}^{B}\big) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}(\gamma, S_b^{est}, \Pi_b) = \bar{\theta}$, where $\bar{\theta}$ is an $m-$dimensional vector of parameters. Then the estimation of the following pairwise variances and covariances can lead to valid confidence ellipses:

$$\begin{aligned}
&Var\left[\hat{\theta}_m(x, S_b^{est}, \Pi_b)\right] = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_m(x, S_b^{est}, \Pi_b) - \bar{\theta}_m\right)^2 \qquad (20)\\
&Cov\left[\hat{\theta}_m(x, S_b^{est}, \Pi_b), \hat{\theta}_{m'}(x, S_b^{est}, \Pi_b)\right] = \frac{1}{B}\sum_{b=1}^{B}\Delta,\\
&\text{where } \Delta = \left(\hat{\theta}_m(x, S_b^{est}, \Pi_b) - \bar{\theta}_m\right)\left(\hat{\theta}_{m'}(x, S_b^{est}, \Pi_b) - \bar{\theta}_{m'}\right).
\end{aligned}$$

The confidence ellipse is given as:

$$(\hat{\theta}(x, S_b^{est}, \Pi_b) - \bar{\theta})^T \hat{S}^{-1}(\hat{\theta}(x, S_b^{est}, \Pi_b) - \bar{\theta}) = r^2. \qquad (21)$$

Here, $\hat{S}^{-1}$ represents the bootstrapped variance-covariance matrix, wherein pairwise covariances are estimated using (20). r^2 follows the chi-squared distribution with M degrees of freedom. Algorithm 1 summarises the steps required to implement the method [4].

Algorithm 1: Multivariate Causal Forest for Group Average Policy Effects

```
Require: number of trees, tree depth, number of leaves $|\Pi|$, the number of
  observations for each bootstrapped data set ($s$), number of observations in each
  leaf, data $\left(\{\mathbf{X}_i\}_{i=1}^N, \{\mathbf{Y}_i\}_{i=1}^N, \{P_i\}_{i=1}^N\right)$.
Ensure: Predicted Group Average Policy Effects.
  1. Identify the optimal number of latent groups based on the partial least
  squares algorithm and $k-$fold cross-validation (see Section 5).
  2. Divide data into train ($S^{tr}$), estimation ($S^{est}$) and test samples ($S^{te}$).
  3. Identify the optimal partitions (leaves) based on $S^{tr}$ and the moment function
  in (19); estimate the policy effects in $S^{est}$.
  4. Predict the group average policy effects by using $S^{te}$.
```

[4] Throughout the article, I use the grf package Tibshirani et al. (2023) to implement the algorithm. I also provide a Python code at the GitHub repository `https://github.com/MariaRevili/multivariate-causal-forest`.

8 Simulated Experiment: Group Average Policy Effects

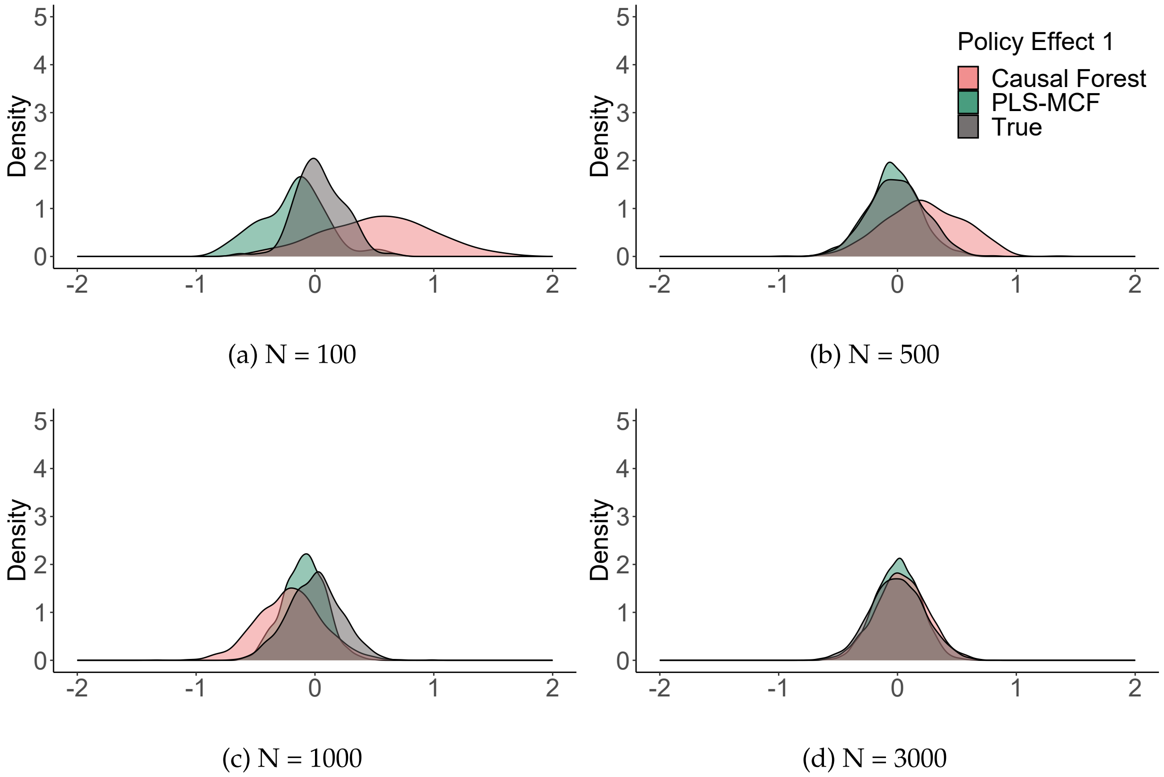
The simulation design emulates a randomized controlled trial proposed by Nareklishvili et al. (2022) within a multivariate framework. The decision-maker is equipped with two distinct outcomes, namely Y_{i1} and Y_{i2}, along with the policy P_i, and four unique variables denoted as X_{i1}, X_{i2}, X_{i3}, and X_{i4}. Concretely, the outcomes are determined by a combination of the policy intervention and the individual characteristics pertaining to each participant $i = 1, \ldots, N$.

$$
\begin{aligned}
& X_{i1}, X_{i2}, X_{i3}, X_{i4}, \sim \mathcal{N}(N, \mu, \Sigma),\ \mu = (-1, 1, 2, 0),\ \Sigma = \mathbf{1}_{4\times 4}, \qquad (22)\\
& P_i \sim \mathcal{B}(N, 1, 0.5),\ \varepsilon_i \sim \mathcal{N}(N, 0, 1)\\
& Y_{i1} = 100 \cdot X_{i1} + 100 \cdot X_{i2} + P_i \cdot (X_{i3} \cdot X_{i4}) + \varepsilon_{i1},\\
& Y_{i2} = 100 \cdot X_{i1} + 100 \cdot X_{i2} + P_i \cdot (X_{i1} \cdot X_{i2}) + \varepsilon_{i2}.
\end{aligned}
$$

The outcomes exhibit a strong and significant correlation of 95.602%. The policy variable (P_i) is generated using a binomial distribution denoted as $\mathcal{B}$. Conversely, the individual characteristics (X_{i1}, X_{i2}, X_{i3}, and X_{i4}) are sampled from a normal distribution, denoted as $\mathcal{N}$. Each outcome also consists of a normally distributed noise (ε_{i1} and ε_{i2}). To investigate the performance of the algorithm, I conduct a Monte Carlo simulation of a randomized controlled trial (RCT) a hundred times. Each simulation is performed with different sample sizes represented by N = 100, 500, 1000, and 3000. The results from these simulations are then averaged to obtain more robust and reliable findings. The variance-covariance matrix (Σ) is an identity matrix, indicating the independence of the individual characteristics.

As highlighted by Nareklishvili et al. (2022), the current design presents notable challenges for the proposed algorithm. The identification of policy-relevant groups becomes intricate due to the considerable influence of independent variables on the outcomes (X_{i1} and X_{i2}). These variables, however, do not contribute significantly to explaining the heterogeneity of policy effects. Moreover, the outcomes exhibit a high degree of correlation and encompass redundant characteristics that do not play a substantial role in either outcome prediction or the estimation of policy effect heterogeneity.

Figure 5 illustrates the density of policy effects ($X_{i3} \cdot X_{i4}$) based on Y_{i1} alongside the corresponding predictions obtained by using the multivariate causal forest and causal forest algorithms.



(a) N = 100

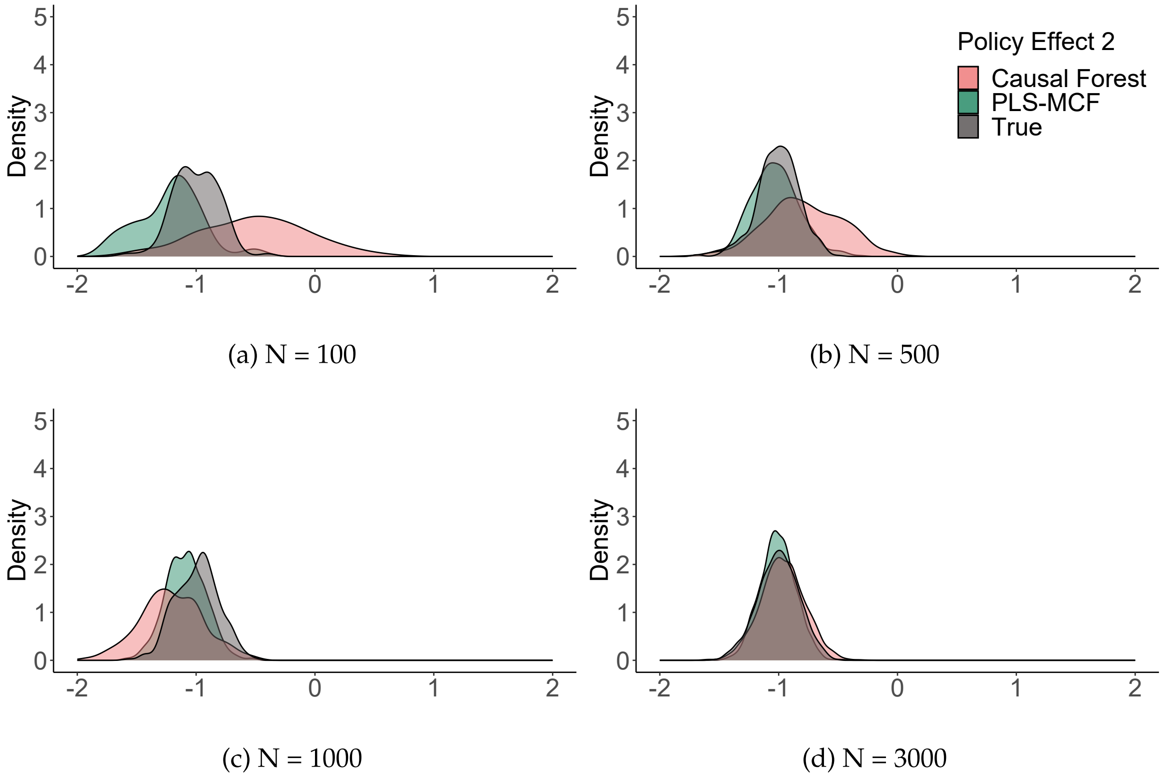
(b) N = 500

(c) N = 1000

(d) N = 3000

Figure 5: The density of simulated and estimated policy effects ($X_{i3} \cdot X_{i4}$) based on Y_{i1} for various number of samples denoted by $N = 100, 500, 1000, 3000$. The densities are estimated by the causal forest and multivariate causal forest after reducing the dimensionality using the partial least squares method (labeled as PLS-MCF). All results are averaged across a hundred different Monte Carlo simulation experiments. The number of components is set to four in each simulated experiment.

The success of the multivariate causal forest is evident in two distinct aspects. First, the algorithm demonstrates remarkable resilience in recovering the true density of policy effects, even when the dataset consists of a limited number of observations. Second, the multivariate random forest exhibits a notable reduction in the variance of group average policy effects compared to the causal forest method, especially when confronted with small sample sizes. This characteristic signifies the stability of the multivariate random forest approach in estimating policy effects across various latent groups. Figure 6 shows the simulated and estimated density of policy effects based on the second outcome Y_{i2}. The implications remain the same.



(a) N = 100

(b) N = 500

(c) N = 1000

(d) N = 3000

Figure 6: The density of simulated and estimated policy effects $(X_{i1} \cdot X_{i2})$ based on Y_{i2} for various number of samples denoted by $N = 100, 500, 1000, 3000$. The results are averaged across a hundred different replications of the experiment. The densities are estimated by the causal forest and multivariate causal forest (labeled as PLS-MCF). The number of components is set to four in each simulated experiment.

As shown by Nareklishvili et al. (2022), the partial least squares method captures latent groups from different segments of the features and outcomes simultaneously. This allows us to recover characteristics relevant to the outcomes as well as policy effect heterogeneity. Furthermore, the multivariate causal forests method is well-tailored for correlated coefficients and contributes to a reduction in the variance of the policy effects. Despite being specifically designed to handle correlated coefficients, Appendix A.1 provides empirical evidence that, in this setting, the effectiveness of the method can be attributed to its capability to efficiently identify and extract latent groups present within the population.

8.1 Multivariate Causal Forest for GATE and Sorted GATE

This section highlights a nuanced distinction between the multivariate random forest for estimating group average policy effects and the Sorted Group Average Treatment Effects (Sorted GATE) method proposed by Chernozhukov et al. (2018).

The authors define distinct (sorted) groups based on the quantiles of the estimated policy effects. Within the context of interest, Chernozhukov et al. (2018) estimate heterogeneous policy effects through a classical causal forest algorithm, labeled as proxy conditional average treatment effect (Proxy CATE). Subsequently, the proxy CATE is partitioned into four distinct quantiles. This division allows for the creation of four groups with unique characteristics: the first group represents observations with policy effects falling between the 0th and 25th percentiles, the second and third groups encapsulate the range between the 25th and 75th percentiles, and the fourth group encompasses individuals with policy effects spanning from the 75th to the 100th percentile. Lastly, the authors run ordinary least squares regression within each group and estimate heterogeneous policy effects.

Sorted GATE is a powerful estimator in diverse scenarios when the Proxy CATE is equipped with reliable, trustworthy information concerning the unobservable policy effects. In the context of the simulation design outlined in this section, we observe that the estimated Proxy CATE (which is analogous to the estimated heterogeneous policy effects using the causal forest approach) displays significant deviations from the simulated density of policy effects under two distinct conditions. First, when the sample size is constrained or limited, the variance and accuracy of the estimated Proxy CATE tend to be affected, leading to less reliable estimates of policy effects. Second, when the input data consist of large noise, the Proxy CATE estimated by the causal forest approach exhibits substantial deviations from the simulated policy effects.

Considering the aforementioned challenges, the sorted groups based on the Proxy CATE may not reflect the correct dimensions that reflect heterogeneous policy effects. Furthermore, given that Sorted GATE utilizes linear regression as its foundation, the resulting estimations of policy effects are susceptible to significant variability, particularly in the presence of noisy data.

Table 1: Estimated Sorted GATEs described by Chernozhukov et al. (2018). Groups 1, 2, 3, 4, 5, and 6 correspond to observations with the proxy CATE falling between 0- 10th, 10-25th, 25-50th, 50-75th, 75-90th, and 90-100th percentiles, respectively. Proxy CATE is estimated by the conventional causal forest algorithm. We adopt the same simulation design summarised by the setup in (23).

	Dependent variable:
Baseline	1.299^{***}
	(0.013)
Sorted GATE - Group 1	26.011^{***}
	(9.066)
Sorted GATE - Group 2	3.420
	(7.375)
Sorted GATE - Group 3	-4.654
	(5.594)
Sorted GATE - Group 4	-5.890
	(5.673)
Sorted GATE - Group 5	1.644
	(7.265)
Sorted GATE - Group 6	-6.934
	(8.991)
Constant	0.941
	(1.411)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1 presents the GATE coefficients alongside their respective standard errors, categorized by groups. The findings from Table 1 indicate that the coefficients obtained through Sorted GATE demonstrate considerable variability attributed to the presence of noise in the data. Furthermore, the observed coefficients do not display significant heterogeneity within the studied groups.

9 Simulation Design: Confidence Ellipses

The multivariate causal forest provides a framework for devising a joint hypothesis test by utilizing high-dimensional confidence intervals. In this section, I investigate the confidence ellipses within a two-dimensional coordinate system. Consider, a simulated experiment of the following form:

$$\begin{aligned}
&X_{i1}, X_{i2}, X_{i3} \sim \mathcal{N}(N, \mu, \Sigma),\ \mu = (0,0,0),\ \Sigma = \mathbf{1}_{3\times 3}, \qquad (23)\\
&P_i \sim \mathcal{B}(N, 1, 0.5),\ \varepsilon_i \sim \mathcal{N}(N, 0, 1)\\
&Y_{i1} = 0.5 + P_i \cdot (X_{i2} + 0.5X_{i1}) + \varepsilon_{i1},\\
&Y_{i2} = 0.5 + P_i \cdot (X_{i3} + 0.5X_{i1}) + \varepsilon_{i2}.
\end{aligned}$$

All variables are defined as before. However, the independent characteristics come from the preferential attachment algorithm (Jeong et al., 2003). Each node of the network represents one feature. The resulting network follows a power-law degree distribution, and thus, is scale-free. That means, only a few variables (characteristics) in the network have a relatively large number of ”neighbors”. The distance between two characteristics is the shortest path between them in the network. We calculate a $D \times D$ ($D = 3$) pairwise distance matrix L, for $\mathbf{X}_i \in \mathbb{R}^D$. Next, this distance matrix is transformed into a covariance matrix $\Sigma_{(m,m')} = 0.5^{L_{(m,m')}}$, where (m, m') represents the element in each row m and column m' of a matrix L ($m, m' = 1, \ldots, D$). Overall, we obtain the following variance-covariance matrix for the three-dimensional characteristic space:

$$\Sigma = \begin{pmatrix} 1 & 0.500 & 0.250 \\ 0.50 & 1 & 0.125 \\ 0.25 & 0.125 & 1 \end{pmatrix}$$

Figure 7 illustrates such a network visually. In this simulation design, I generate a network with a node size of one thousand, and the number of observations N equals one thousand.

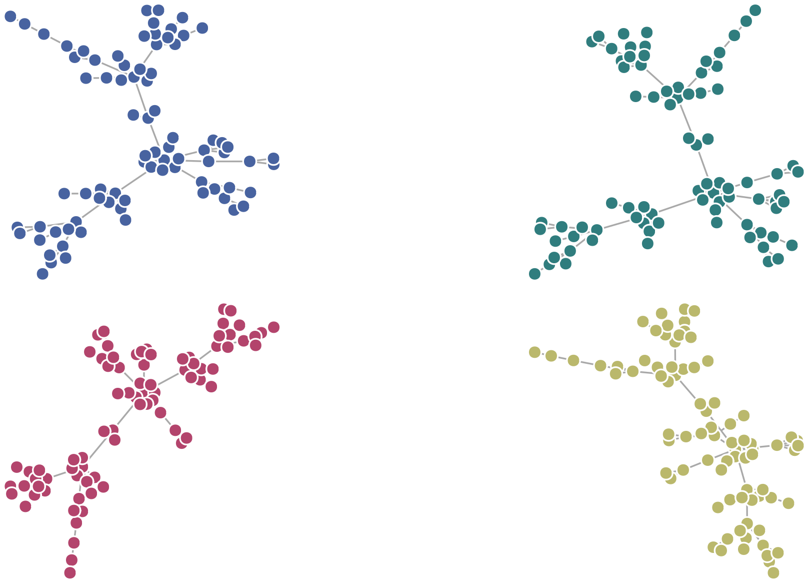


Figure 7: Visualization of four different network structures inherent to the space of independent characteristics. The number of nodes is set to fifty in each case.

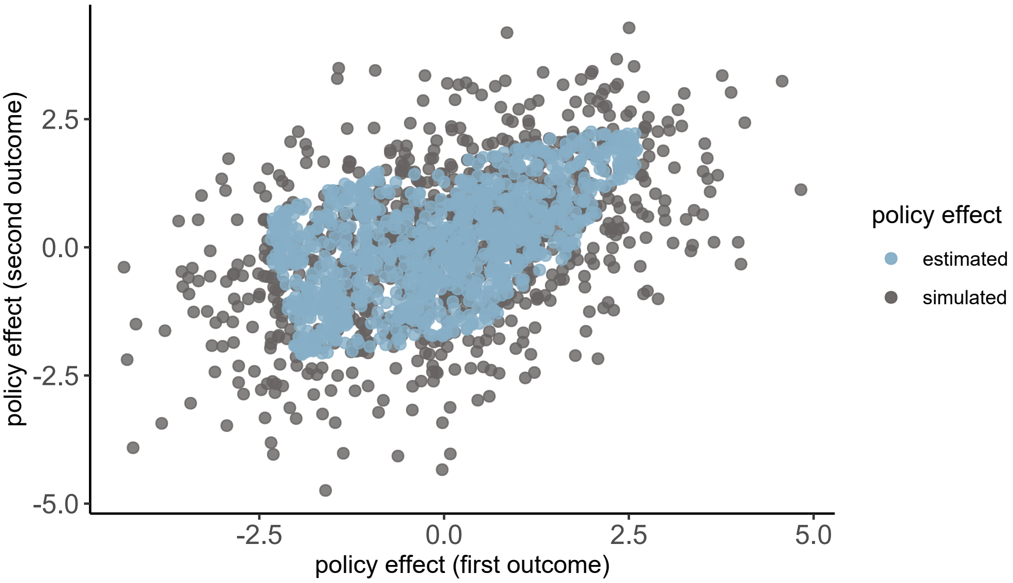
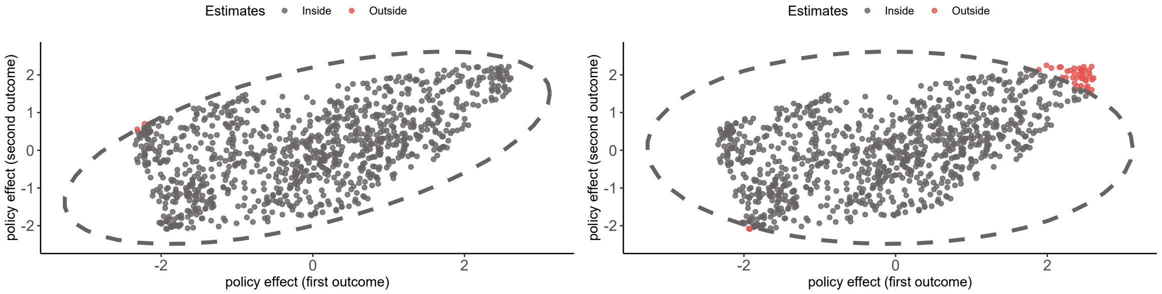


Figure 8: Simulated and estimated policy effects in a two-dimensional coordinated system. The number of simulated observations is equal to one thousand, and the number of trees in a multivariate causal forest is set to one thousand. The mean squared error for the policy effects stemming from the first and second outcomes are 0.257 and 0.305, respectively.

To estimate policy effects, I construct a multivariate causal forest with one thousand trees. Then I aggregate the results by averaging policy effect estimates across each tree. Moreover, I estimate the variance-covariance matrix by using the bootstrap approach. Figure 8 shows the simulated and estimated policy effects. The results closely resemble the simulated effects, and the mean squared error is negligible in each case. The computed mean covariance between the first and second policy effects is 0.772, closely approximating the true value of 1.024.

Figure 9 illustrates confidence ellipses for policy effects with and without accounting for covariance. Panel (a) incorporates covariance, while Panel (b) does not. Notably, in Panel (b), around 5% of individuals fall outside the ellipse, potentially leading to misleading inference on policy effect estimates.

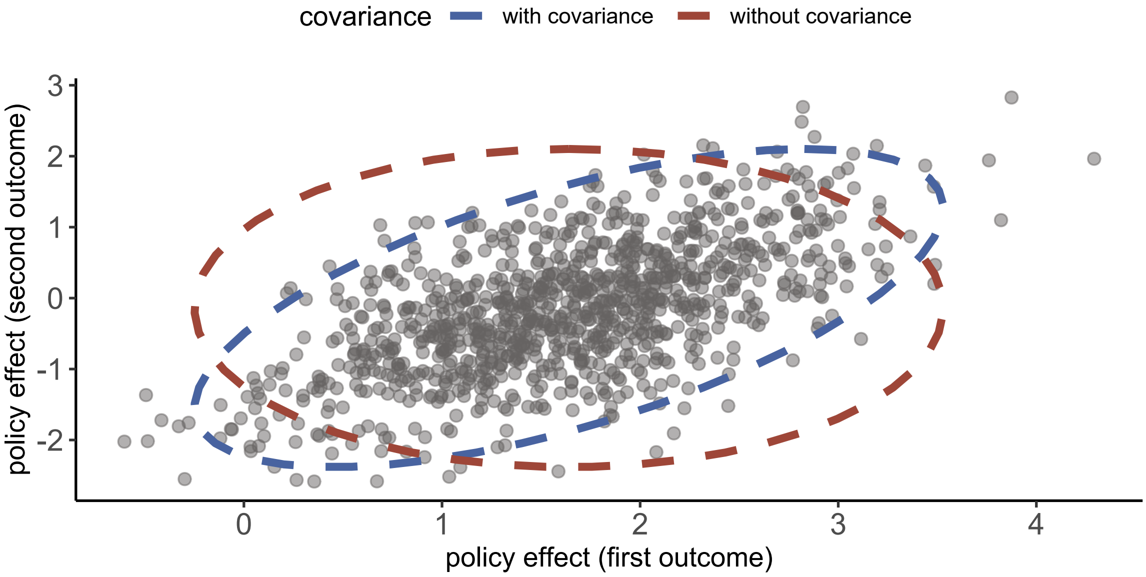


(a) Confidence ellipse with covariance

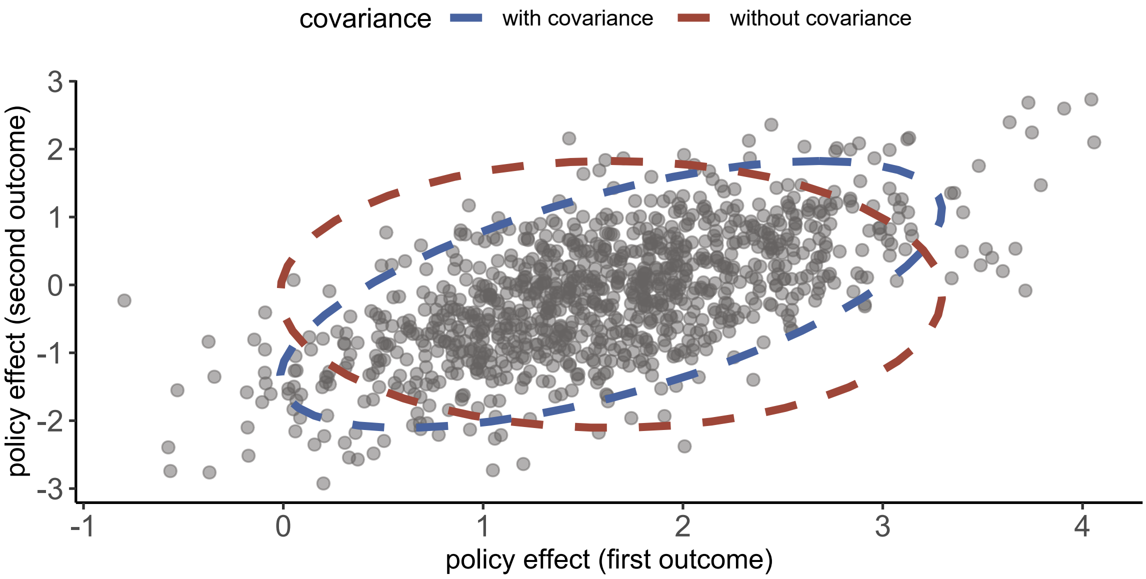
(b) Confidence ellipse without covariance

Figure 9: 95% confidence ellipses with a radius that equals the square root of chi-squared distribution with two degrees of freedom. The confidence ellipse without covariance induces 49 out of 1000 observations to fall outside the interval.

Figure 10 shows 95% and 90% confidence ellipses for a particular individual. The improved coverage and precision of confidence ellipses are evident when considering the covariance of policy effects. Appendix A.1, Figure 19, illustrates confidence ellipses for groups with varying signs of covariance. Appendix A.3 in addition illustrates the rates of convergence of the multivariate causal forest after the dimensionality reduction and GRF. According to Appendix A.3, the distinction is negligible.



(a) 95% confidence ellipse



(b) 90% confidence ellipse

Figure 10: Panel (a) and (b) correspond to 95% and 90% confidence ellipses, respectively, for one particular individual. Data are simulated from a joint normal distribution of policy effects with the means of 1.64, -0.140, variances of 0.596, 0.839, and covariance of 0.445, respectively.

10 Dialogue Meetings and Sickness Absence: A Field Experiment

The Norwegian insurance system grants all employees the entitlement to sick leave benefits for a period of up to one year, following a qualifying span of four weeks of employment. During this period, the initial few days of sick leave (either 3 or 8 days) can be self-certified, but any further sick leave must be certified by a physician. Norway has consistently ranked among the European leaders in terms of sick-leave rates, with the rate hovering around 6 percent. As a result, efforts to combat absenteeism and promote a swift return to work have become central to the political agenda for an extended period. In this attempt, the use of dialogue meetings (DMs) has emerged as one of the prominent measures adopted to address the issue.

The Norwegian Labour and Welfare Administration (NAV) facilitates and coordinates dialogue meetings by extending invitations to the sick-listed employee, the employer, and the physician involved. Additionally, a caseworker from NAV actively participates and presides over the meeting, and documents the duration of sick leave. These dialogue meetings are conducted within a specified timeframe of 26 weeks from the onset of the sickness spell. During these gatherings, the involved parties engage in a comprehensive assessment of the situation at hand, jointly devising a well-coordinated plan of action aimed at reintegrating the sick-listed worker back into the workplace. Although the meeting does not culminate in a legally binding agreement, a NAV caseworker summarises the meeting and drafts a comprehensive timeframe of the sickness absence of a worker.

Alpino et al. (2022) conduct an extensive, pre-registered, and randomized field experiment in collaboration with the Norwegian Labour and Welfare Administration (NAV). The primary aim is to investigate the effects of summoning sick-listed individuals to a dialogue meeting, as compared to those who are not summoned but still have the option to request one. The experiment was carried out between 2016 and 2018, employing a thoughtfully designed randomization scheme. Specifically, the internet page implements a random draw of individuals, allocating the absentee to one of eight different treatment arms.

The present paper builds upon data provided by Alpino et al. (2022), which extensively utilize three primary sources: i) The randomization data set, which encompasses information on the assignment of individuals on sick leave to different treatment arms. ii) Caseworker surveys, which provide valuable insights into the caseworkers' perspectives and experiences during the experiment. iii) Outcomes and independent characteristics delivered by NAV, which offers comprehensive information on various relevant features of the absentees and outcome measures.

We investigate the heterogeneous effects of dialogue meetings on two distinct outcome measures: total days of sickness absence (*Total days*) and days of sick leave within the current spell (*Days within spell*). Total days of sickness absence represent the cumulative count of days an individual was absent due to sickness from the initial draw date until the date of data extraction. On the other hand, days within spell refer to the specific number of sick leave days experienced within the ongoing sick leave spell. These outcomes are highly correlated, with a rate of 78.4%. Independent characteristics consist of gender (a binary indicator for *Female*), *Age*, *Marital status*, *Nationality*, the *Grade* (i.e. percentage) of sickness absence at the time of the draw, *Days before* representing the total number of days on sickness absence since 2015 up until the date of the draw. *Symptoms* which reflect the share of absentees classified by the physician as having symptoms rather than diagnoses according to the International Classification of Primary Care (ICPC-2), *nr employees* being the number of employees at the absentee's workplace (at the time of the draw). A detailed description of data and summary statistics can be found in the work of Alpino et al. (2022).

10.1 Interpreting Identified Latent Groups

Cross-validation, as detailed in Appendix A.2, reveals the presence of four distinct target components. These latent components are continuous and encompass different groups of the population. To interpret the groups, we employ the ordinary least squares regression where each component is regressed on all the characteristics of sick-listed workers.

Table 2 investigates the relationship between the components and various characteristics of sick-listed workers. The results of Table 2 highlight several key observations regarding the identified groups. First, it is evident that the characteristics fully explain the generated groups ($R^2 = 1$). This finding is not surprising as the identified groups of the population represent linear combinations of the covariates. Second, the coefficients associated with each characteristic play a crucial role in determining the composition of each segment within the identified groups. For instance, the positive coefficient attributed to Married in Component 2 implies that the high values of Component 2, among others, are represented by married sick-listed workers. Lastly, the magnitude of these coefficients provides insights into the extent of influence each individual characteristic has over a given component.

Table 2: Ordinary Least Squares regression of four distinct components and the characteristics they comprise of. Highlighted values represent the coefficients with a value greater than 0.500 for each. Comp is equivalent to a given component.

	Dependent variable:			
	Comp 1	Comp 2	Comp 3	Comp 4
	(1)	(2)	(3)	(4)
Female	0.098***	−**0.605*****	−0.495***	**1.196*****
	(0.000)	(0.000)	(0.000)	(0.000)
Age	−**0.601*****	−0.170***	0.057***	**0.514*****
	(0.000)	(0.000)	(0.000)	(0.000)
Never-married	0.401***	−**1.109*****	0.168***	0.138***
	(0.000)	(0.000)	(0.000)	(0.000)
Grade	**1.033*****	**0.748*****	−**1.021*****	−0.277***
	(0.000)	(0.000)	(0.000)	(0.000)
nr employees	0.156***	0.138***	−0.176***	−0.067***
	(0.000)	(0.000)	(0.000)	(0.000)
Married	−0.060***	**1.272*****	0.071***	0.104***
	(0.000)	(0.000)	(0.000)	(0.000)
Norwegian	0.150***	0.0003***	**0.788*****	−0.418***
	(0.000)	(0.000)	(0.000)	(0.000)
Widow	−0.136***	**1.587*****	0.038***	−**3.304*****
	(0.000)	(0.000)	(0.000)	(0.000)
Days before	−0.004***	−0.003***	−0.007***	−0.003***
	(0.000)	(0.000)	(0.000)	(0.000)
Symptoms	**0.649*****	0.081***	−0.278***	**0.849*****
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	−0.557***	0.321***	1.220***	−0.293***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	10,230	10,230	10,230	10,230
R^2	1.000	1.000	1.000	1.000
Adjusted R^2	1.000	1.000	1.000	1.000
Residual Std. Error (df = 10219)	0.000	0.000	0.000	0.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Larger coefficients imply a more substantial impact, emphasizing the degree of exposure each characteristic has within a particular group. In this article, we focus on primary characteristics with the absolute value of the coefficient larger than 0.500 (highlighted in Table 2). Overall, according to Table 2, the low values of each corresponding component are primarily equivalent to:

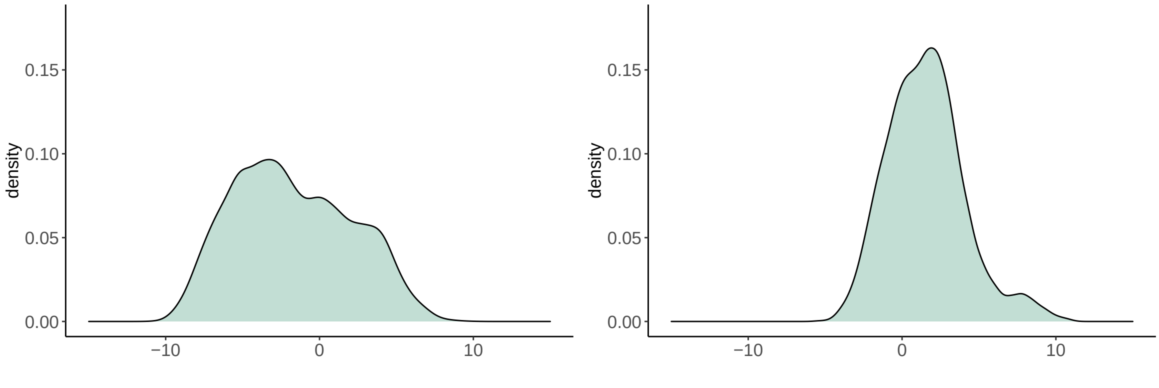
- Component 1 $\sim$ older absentees with a low grade of sickness absence at the time of the draw and no symptoms,

- Component 2 $\sim$ females who have never been married or widowed, and have a low grade of sickness absence at the time of the draw,
- Component 3 $\sim$ Norwegians with a low grade of sickness absence at the time of the draw,
- Component 4 $\sim$ young widowed males with no symptoms.

Likewise, elevated values within the components can be perceived as opposite to the characteristics observed in the low-value segments. Multiple collections of four components comprise groups. A notable advantage of the derived components lies in their continuity. They reflect a diverse spectrum of individuals represented with varying degrees within each group segment.

10.2 The effect of dialogue meetings on sick leave

Figure 11 depicts the probability density functions representing the influence of dialogue meetings (DMs) on two correlated outcomes: ”Total days” and ”Days within spell”. The independent variables for the multivariate causal forest are the identified components of the population.



(a) The effect of DM on Total days

(b) The effect of DM on Days within spell

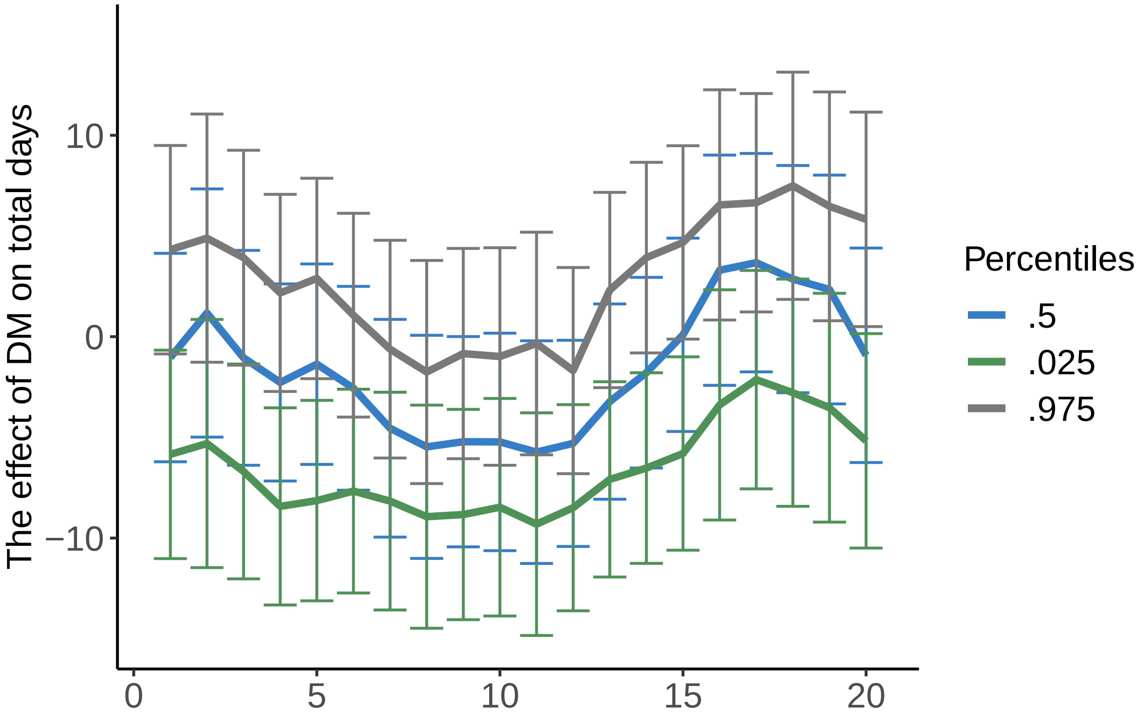
Figure 11: The densities of the effect of dialogue meetings (DM) on total days of sick leave and days of sick leave within the spell, respectively.

The impact of DMs on total days displays noteworthy heterogeneity. This heterogeneity suggests that certain individuals experience a substantial reduction of 10 days in sick leave due to DMs, while others encounter an increase of 10 days in sick leave as a consequence of these meetings. The findings highlight the variability of the relationship between DMs and sick leave, warranting further investigation and consideration in decision-making processes.

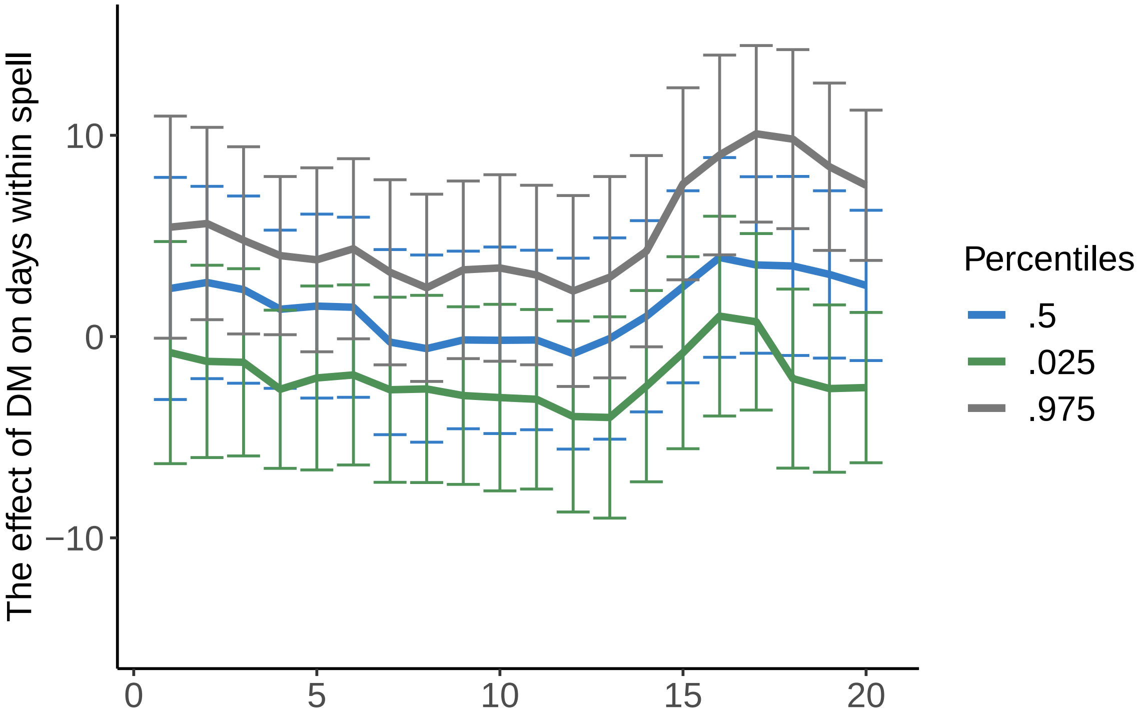
Figure 12 presents the primary findings of this article, depicting distinct quantiles of the impact of dialogue meetings (DMs) on Total days (a) and Days within spell (b), respectively. The figure reveals a two-dimensional pattern of heterogeneity. Specifically, the effect of DMs on each outcome displays substantial variation within and across the vigintiles of the second component [5].

Among never-married females with a low percentage (grade) of sickness absence (representing the bottom vigintiles), there is a significant reduction in the total number of days on sick leave. On average, never-married females can contract up to 10 days of sick leave. The effect on Days within spell is positive, yet statistically non-significant. Conversely, for married sick-listed workers, dialogue meetings lead to a significantly prolonged sick leave, with an extension of up to 10 days.

[5] I do not observe statistically significant heterogeneities among the other identified components; consequently, I do not present them in the article.



(a) Component 2 vigintiles (based on total days)

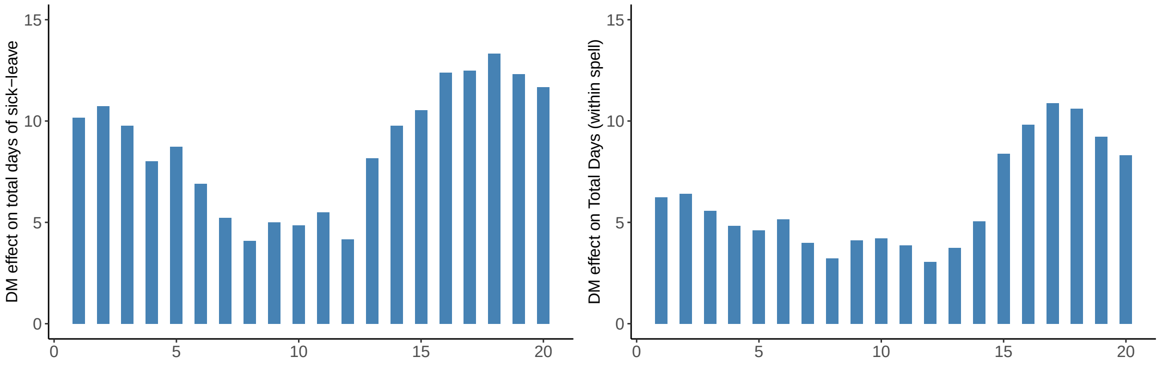


(b) Component 2 vigintiles (based on days within spell)

Figure 12: The effect of dialogue meetings (DM) on total days of sick leave and days of sick leave within the spell, respectively.

Moreover, the policy effect of the married exhibits a higher variance compared to never-married females. Figure 13 visually depicts this difference, showing a significant increase in the spread between the top 0.975th and 0.025th percentiles, particularly among the segment of married sick-listed workers. In contrast, Appendices A.3.1 and A.3.2 show that the conventional causal forest method does not detect significant heterogeneities across the given components or the original covariates.

These empirical findings underscore the necessity of adopting a highly personalized and targeted policy-making, specifically tailored to address the needs of married workers who exhibit a higher grade of severity and significantly extended sick leave as a result of dialogue meetings.



(a) Component 2 vigintiles (based on total days) (b) Component 2 vigintiles (based on days within spell)

Figure 13: The difference between top and bottom quantiles

10.3 Explaining Policy Effect Heterogeneity

I further investigate the economic significance of the findings by displaying the proportion of never-married females across the vigintiles of the second component. As Figure 14 shows, the share of never-married females is more than 15% in the initial five vigintiles and progressively diminishes to zero in the last five.

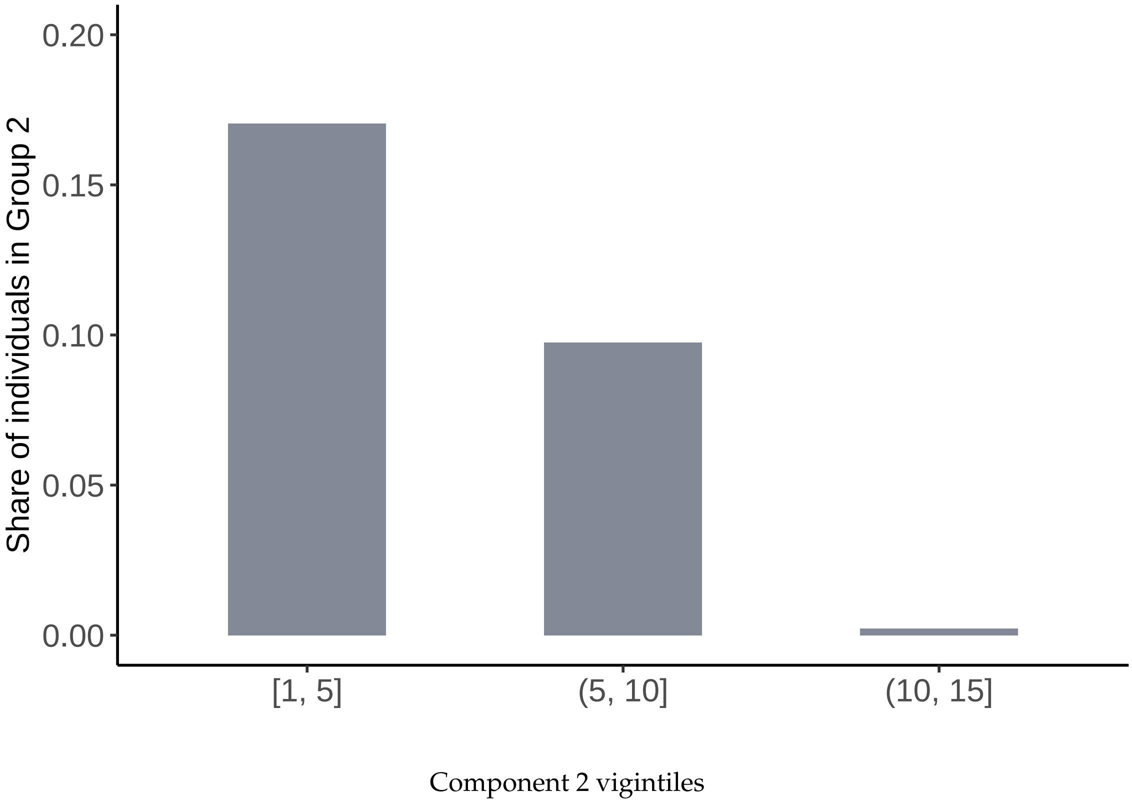
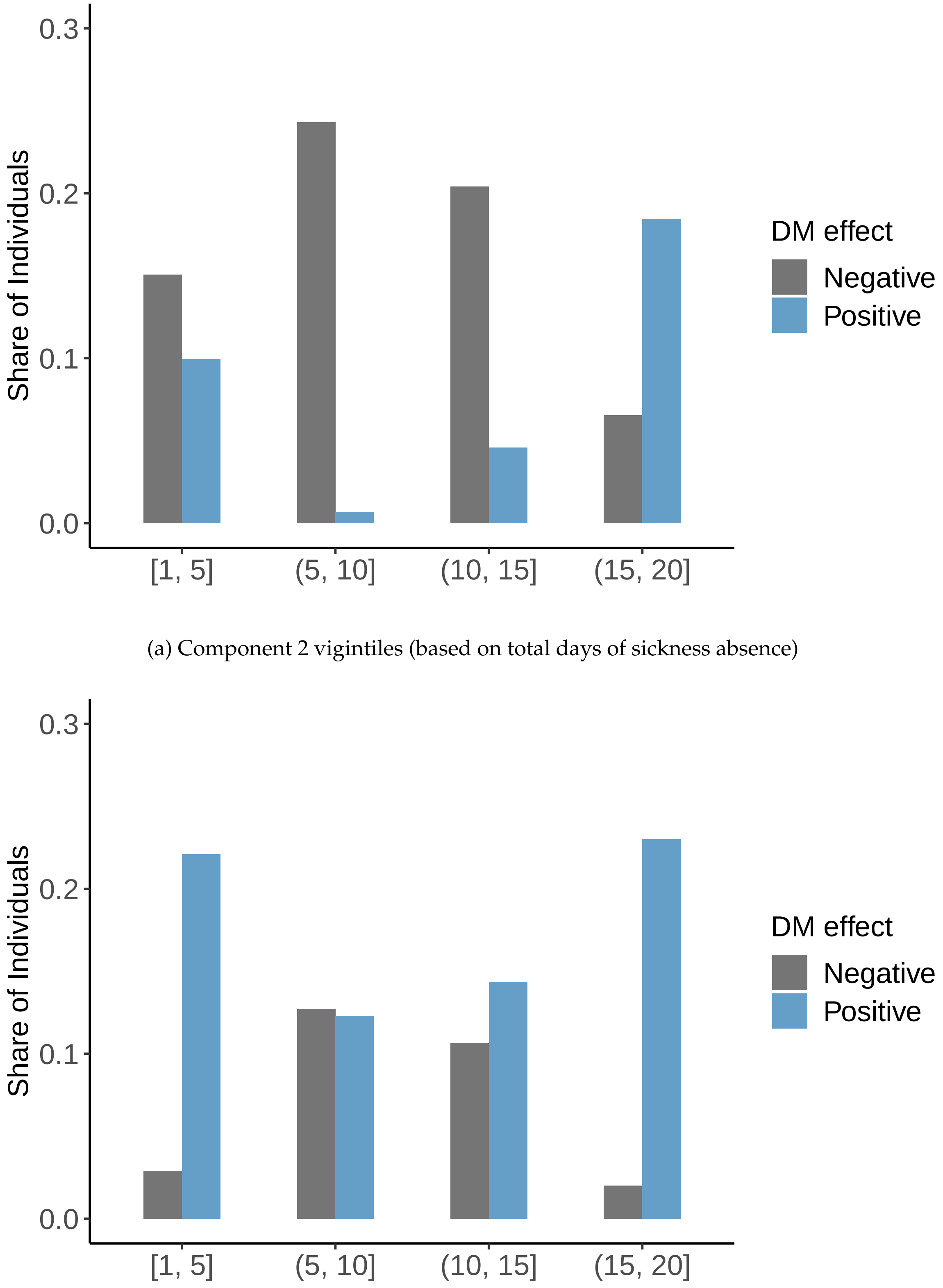


Figure 14: Share of never-married females within vigintiles of the second component.

I further conduct a comparative analysis on the distribution of individuals experiencing positive and negative policy effects, as depicted in Figure 15. The graph shows a notable prevalence of negative effects of DMs on total days of sick leave among the initial fifteen vigintiles of the second component. Conversely, this trend reverses in the uppermost five vigintiles of the second component, which are entirely comprised of married sick-listed workers.

Findings indicate that the disparities in the effects of DMs on the days of sick leave are predominantly attributed to marital status.



(a) Component 2 vigintiles (based on total days of sickness absence)

(b) Component 2 vigintiles (based on days within the sickness spell)

Figure 15: The effect of dialogue meetings (DM) on total days of sick leave and days of sick leave within the spell, respectively.

11 Conclusion

Policymakers frequently encounter scenarios where they have access to diverse policies or outcomes and wish to assess variations in policy effects among distinct population groups. In response, this paper introduces a multivariate causal forest method specifically designed to estimate such variations, both within and across individual groups of the population.

The underlying groups in this article are initially unknown and can be both continuous and unordered. The process of identifying the groups is adaptive, utilizing the partial least squares algorithm. The key aspect of this approach is that the identified target components that comprise a group are represented as linear combinations of individual characteristics. As a result, I show that the formed groups possess meaningful economic interpretations. As a next step in policy effect evaluation, the multivariate causal forest allows us to detect specific segments of the groups that are most susceptible to a policy or intervention. In this article, I illustrate that the estimates of the multivariate causal forest are asymptotically normally distributed.

To illustrate the applicability of the algorithm, I revisit a field experiment, the influence of dialogue meetings on the duration of sickness absence (described by Alpino et al., 2022). Among the identified groups, I find that marital status explains a significant variation in policy effects. On average, never-married females can contract up to 10 days of sick leave. Conversely, for married sick-listed workers, dialogue meetings lead to a significantly prolonged sick leave, with an extension of up to 10 days. Moreover, the policy effect of the married exhibits a higher variance compared to never-married females. These empirical findings underscore the necessity of adopting a highly personalized and targeted policy-making, specifically tailored to address the needs of married workers who exhibit a higher grade of severity and significantly extended sick leave as a result of dialogue meetings.

The multivariate causal forest methodology is highly suitable for randomized experiments that involve multiple outcomes or policy options. One useful extension would be to generalize the theoretical properties of the multivariate causal forest to non-i.i.d. data. Minh et al. (2023) introduce Hoeffding decomposition for U-statistic in the presence of network effects. I plan to adopt their approach and accommodate network structures in the setup. The second useful theoretical extension is to investigate the theory of partial least squares when the individual characteristics are not elliptically distributed. Lastly, I plan to extend the empirical results of a paper by proposing an economic model that theoretically explains the differences between never-married females and married absentees.

References

Keshav Agrawal, Susan Athey, Ayush Kanodia, and Emil Palikot. Personalized recommendations in edtech: Evidence from a randomized controlled trial. *arXiv preprint arXiv:2208.13940*, 2022.

Matteo Alpino, Karen Evelyn Hauge, Andreas Kotsadam, and Simen Markussen. Effects of dialogue meetings on sickness absence—evidence from a large field experiment. *Journal of Health Economics*, 83:102615, 2022.

Joshua Angrist and William N Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size, 1996.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

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

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

Susan Athey, Guido W Imbens, et al. Machine learning for estimating heterogeneous causal effects. Technical report, 2015.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2), 2019.

Betsy J Becker. Multivariate meta-analysis. *Handbook of applied multivariate statistics and mathematical modeling*, pages 499–525, 2000.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095, 2012.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Leo Breiman. Consistency for a simple model of random forests. *University of California at Berkeley. Technical Report*, 670, 2004.

David R Brillinger. A generalized linear model with "gaussian" regressor variables. *Selected Works of David Brillinger*, pages 589–606, 2012.

Martin Browning, Jesus Carro, et al. Heterogeneity and microeconometrics modelling. *Econometric Society Monographs*, 43:47, 2007.

Domagoj Ćevid, Loris Michel, Jeffrey Näf, Nicolai Meinshausen, and Peter Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *arXiv preprint arXiv:2005.14458*, 2020.

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2): 491–525, 2006.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

Victor Chernozhukov, Jerry A Hausman, and Whitney K Newey. Demand analysis with many prices. Technical report, National Bureau of Economic Research, 2019.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning*, pages 665–673. PMLR, 2014.

AW Asymptotic der Vaart, Van. Cambridge university press: New york. *NY, USA*, 1998.

Cyrus DiCiccio and Joseph Romano. Clt for u-statistics with growing dimension. *Statistica Sinica*, 32(1), 2022.

Michaela Draganska and Dipak Jain. Structural models of competitive market behavior: An estimation approach using disaggregate data. Technical report, Society for Computational Economics, 2002.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

Jaroslav Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346, 1968.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

Inge Helland. Partial least squares regression. *Wiley StatsRef: Statistics Reference Online*, 2014.

Inge S Helland. On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607, 1988.

Inge S Helland. Partial least squares regression and statistical models. *Scandinavian journal of statistics*, pages 97–114, 1990.

Wassily Hoeffding. The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics, 1961.

Ganesh Iyer and J Miguel Villas-Boas. A bargaining theory of distribution channels. *Journal of marketing research*, 40(1):80–100, 2003.

Dan Jackson, Richard Riley, and Ian R White. Multivariate meta-analysis: potential and promise. *Statistics in medicine*, 30(20):2481–2498, 2011.

Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *Europhysics letters*, 61(4):567, 2003.

Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

Stanislav Kolenikov and Kenneth A Bollen. Testing negative error variances: Is a heywood case a symptom of misspecification? *Sociological Methods & Research*, 41 (1):124–167, 2012.

Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*, volume 273. Springer Science & Business Media, 2013.

Michael Lechner. *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*. Springer, 2001.

Roger J Lewis. An introduction to classification and regression tree (cart) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, volume 14. Citeseer, 2000.

Kevin Li. Asymptotic normality for multivariate random forest estimators. *arXiv preprint arXiv:2012.03486*, 2020.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Siyun Liu and Tao Yu. Kernel density estimation in mixture models with known mixture proportions. *Statistics in Medicine*, 40(28):6360–6372, 2021.

Evan Mayo-Wilson, Nicole Fusco, Tianjing Li, Hwanhee Hong, Joseph K Canner, Kay Dickersin, et al. Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *Journal of clinical epidemiology*, 86: 39–50, 2017.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Tâm Le Minh, Sophie Donnet, François Massol, and Stéphane Robin. Hoeffding-type decomposition for *u*-statistics on bipartite networks. *arXiv preprint arXiv:2308.14518*, 2023.

Maria Nareklishvili, Nicholas Polson, and Vadim Sokolov. Feature selection for personalized policy analysis. *arXiv preprint arXiv:2301.00251*, 2022.

Denis Nekipelov, Paul Novosad, and Stephen P Ryan. Moment forests. 2018.

I Olkin and L Gleser. Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, 2:357–376, 2009.

Giovanni Peccati. Hoeffding-anova decompositions for symmetric statistics of exchangeable observations. 2004.

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

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. 2015.

Steven Stern. Semiparametric estimates of the supply and demand effects of disability on labor force participation. *Journal of Econometrics*, 71(1-2):49–70, 1996.

Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. Package 'grf', 2023.

Jodie B Ullman and Peter M Bentler. Structural equation modeling. *Handbook of Psychology, Second Edition*, 2, 2012.

Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

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

Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

Guihua Wang, Jun Li, and Wallace J Hopp. An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science*, 68(5):3399–3418, 2022.

You-Gan Wang and Xu Lin. Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics*, 61(2):413–421, 2005.

A.0.1 Lemma 6.1

Proof. Define the Hajek projection of the multivariate random forest estimator:

$$\dot{\mathcal{F}}(\gamma, A_1, \ldots, A_N) - \mu = \sum_{i=1}^{N} \mathbb{E}\big(\mathcal{F}(\gamma, A_1, \ldots, A_N) - \mu | A_i\big) = \tag{24}$$
$$\frac{1}{\binom{N}{s}} \sum_{i=1}^{N} \mathbb{E}\Bigg(\sum_{1 \leq i_1 \leq \cdots \leq i_s \leq N} \mathbb{E}_{\xi} T(\gamma, \xi, A_{i_1}, \ldots, A_{i_s}) - \mu | A_i \Bigg),$$

where $\binom{N}{s}$ is the number of $i_1 \leq \cdots \leq i_s$ size-s subsets from $1, \ldots, N$ observations.

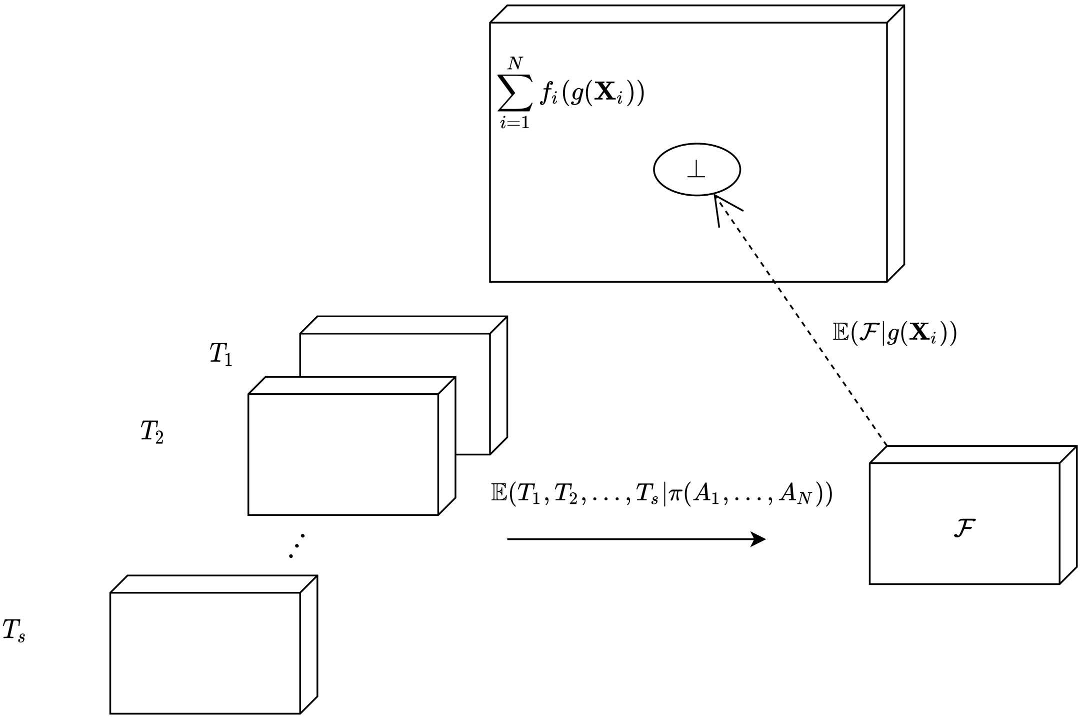

$\sum_{i=1}^{N} f_i(g(\mathbf{X}_i))$
$\perp$
$\mathbb{E}(\mathcal{F}|g(\mathbf{X}_i))$
T_1
T_2
T_s
$\mathbb{E}(T_1, T_2, \ldots, T_s|\pi(A_1, \ldots, A_N))$
$\mathcal{F}$

Figure 16: Hajek Projection of a (vector-valued) U-statistic, formed by the expectation of symmetric functions (i.e., trees), and aggregated over subsamples $i_1, \ldots, i_s$. $\pi(A_1, \ldots, A_N)$ denotes permutations of the data and $\perp$ is the orthogonality. Projection is the expectation of $\mathcal{F}$ conditional on covariates $g(\mathbf{X}_i)$. It can be shown that $\mathbb{E}\left[(T - \dot{T}) \sum_{i=1}^{N} f(g(\mathbf{X}_i))\right] = 0$.

When the observation i is not in samples $1 \leq i_1 \leq \cdots \leq i_s$, then the conditional expectation of the tree (aggregated over the randomization) is the same as the unconditional one. Therefore:

$$\mathbb{E}\big(\mathbb{E}_{\xi} T(\xi, A_{i_1}, \ldots, A_{i_s}) | A_i\big) = \mathbb{E}_{\xi, A_{i_1}, \ldots, A_{i_s}} T(\gamma, \xi, A_{i_1}, \ldots A_{i_s}) = \mu.$$

Overall, there are $\binom{N-1}{s-1}$ samples that contain observation i. Moreover, the sequence of observations is *i.i.d*. and the trees are permutation symmetric. Therefore, for each sample,

$$\mathbb{E}\big(\mathbb{E}_{\xi}T(\gamma,\xi,A_{i_1},\dots,A_{i_s})-\mu|A_i\big)=T_1(A_i)-\mu, \tag{25}$$

where $T_1(a)=\mathbb{E}_{\xi,A_2,\dots,A_N}T(\gamma,\xi,a,A_2,\dots,A_N)$.

Incorporating (25) in (24) yields:

$$\dot{\mathcal{F}}(\gamma,A_1,\dots,A_N)-\mu=\frac{\binom{N-1}{s-1}}{\binom{N}{s}}\sum_{i=1}^{N}\big(T_1(A_i)-\mu\big)=\frac{s}{N}\sum_{i=1}^{N}\big(T_1(A_i)-\mu\big). \tag{26}$$

Since the observations $A_1,\dots,A_N$ are i.i.d, the same property holds for $T_1(A_i)$. By taking the expectation of both sides in (26), we can easily verify that $\mathbb{E}\big(\dot{\mathcal{F}}(\gamma)\big)=\mu$ where $\dot{\mathcal{F}}(\gamma)=\dot{\mathcal{F}}(\gamma,A_1,\dots,A_N)$. Define Σ to be the covariance matrix of $\dot{\mathcal{F}}(\gamma,A_1,\dots,A_N)$. Then:

$$\Sigma=\mathbb{V}\left[\frac{s}{N}\sum_{i=1}^{N}\big(T_1(A_i)-\mu\big)\right]=\frac{s^2}{N}\mathbb{V}\big(T_1(A_i)\big)=\frac{s}{N}\mathbb{V}\big(\sum_{i=1}^{s}T_1(A_i)\big)=\frac{s}{N}\mathbb{V}(\dot{T})\in\mathbb{R}^{M\times M}, \tag{27}$$

where $\dot{T}=\sum_{i=1}^{s}T_1(A_i)$ is the Hajek projection of a tree $T(\gamma,A_1,\dots,A_N)=\mathbb{E}_{\xi}T(\gamma,\xi,A_1,\dots,A_N)\in\mathbb{R}^M$. Note that, a tree T is symmetric in its arguments, and observations $i=1,\dots,N$ are *i.i.d*. Therefore, the Hajek projection of a tree estimator reduces to $\sum_{i=1}^{s}T_1(A_i)$ (as in (26)). We disregard the second (constant) term, as it does not enter in the variance $\mathbb{V}$. Note that since the statistic $T_1(A_i)$ is a vector, the operation $\mathbb{V}$ applies coordinate-wise.

□

A.0.2 Lemma 6.2

Proof. Define the mean squared deviation of the multivariate forest estimator and its projection:

$$\mathbb{E}(\mathcal{F}-\dot{\mathcal{F}})^T\Sigma^{-1}(\mathcal{F}-\dot{\mathcal{F}}) = \mathbb{E}\left[\mathrm{tr}\Sigma^{-1}(\mathcal{F}-\dot{\mathcal{F}})(\mathcal{F}-\dot{\mathcal{F}})^T\right] = \\ \mathrm{tr}\Sigma^{-1}\mathbb{E}(\mathcal{F}-\dot{\mathcal{F}})(\mathcal{F}-\dot{\mathcal{F}})^T = \mathrm{tr}\Sigma^{-1/2}\mathbb{V}\big(\mathcal{F}-\dot{\mathcal{F}}\big)\Sigma^{-1/2}. \tag{28}$$

Assume there exist functions T_i, such that the following equality holds:

$$\mathbb{E}\big(T_i(g(\mathbf{X}_i)\in B)|g(\mathbf{X}_i)\notin B)\big) = 0. \tag{29}$$

Equation (29) is the necessary condition for the weak independence of the exchangeable sequences of $g(\mathbf{X}_i)$. Assume, $T_i(g(\mathbf{X}_i)\in B)$ are symmetric, square-integrable, vector-valued functions. Then each T_i and $T_{i'}$ are pairwise independent. Since $i = 1,\ldots,N$ is an exchangeable ($i.i.d$) sequence, Theorem 6 of Peccati (2004) applies. In addition, Proposition 1 of Li (2020) applies to our case as well. We define Höeffding decomposition of a multivariate U-statistic:

$$\mathcal{F}-\dot{\mathcal{F}} = \frac{1}{\binom{N}{s}}\left[\sum_{i<j}\binom{N-2}{s-2}\big(T_2(A_i,A_j)-\mu\big) + \sum_{i<j<m}\binom{N-3}{s-3}\big(T_3(A_i,A_j,A_m)-\mu\big)+\ldots\right. \tag{30}$$

where $T_2, T_3\ldots$ are second, third, and higher order projections of a tree T that meet the following conditions:

$$\mathbb{E}\big(T_i-\mu\big)^T\Sigma^{-1}\big(T_{i'}-\mu\big) = 0 \text{ for each } i\neq i', \text{ and} \tag{31}$$

$$\mathbb{E}\big(T_i-\mu\big)^T\Sigma^{-1}\big(T_i-\mu\big) \leq \mathbb{E}\big(T-\mu\big)^T\Sigma^{-1}\big(T-\mu\big), \tag{32}$$

where T_i and $T_{i'}$ are the i- th and i'-th projections of the tree, with $i = 1,\ldots,N$.

We fix the variance (Σ) of the multivariate random forest estimator. Moreover, we notice that $\binom{N}{s} \geq \binom{N-1}{s-1} \geq \binom{N-2}{s-2} \geq \binom{N-3}{s-3} \geq \dots$. Therefore:

$$\mathcal{F} - \dot{\mathcal{F}} \leq \frac{s}{N}\left[\sum_{i<j}\left(T_2(A_i, A_j) - \mu\right) + \sum_{i<j<m}\left(T_3(A_i, A_j, A_m) - \mu\right) + \dots\right], \tag{33}$$

where $\frac{s}{N} = \frac{\binom{N-1}{s-1}}{\binom{N}{s}}$. Based on Equation (32), the variance of $\mathcal{F} - \dot{\mathcal{F}}$ has an upper bound:

$$\mathbb{V}(\mathcal{F} - \dot{\mathcal{F}}) \leq \left(\frac{s}{N}\right)^2 \mathbb{V}(T). \tag{34}$$

In (27) we derived $\Sigma = \frac{s}{N}\mathbb{V}(\dot{T})$. Plugging the value of Σ and (34) in (28) leads to the upper bound of the squared deviation:

$$\mathbb{E}(\mathcal{F} - \dot{\mathcal{F}})^T \Sigma^{-1} (\mathcal{F} - \dot{\mathcal{F}}) \leq tr\left((\frac{s}{N}\mathbb{V}(\dot{T}))^{-1/2} \left(\frac{s}{N}\right)^2 \mathbb{V}(T) (\frac{s}{N}\mathbb{V}(\dot{T}))^{-1/2} \right) = \frac{s}{N} tr\left((\mathbb{V}(\dot{T}))^{-1} \mathbb{V}(T) \right) \tag{35}$$

In the final equality, we use the cyclic property of the trace operator: $tr(XYZ) = tr(YZX) = tr(ZXY)$.

□

A.0.3 Theorem 6.1

Proof. Bounded elements of $\mathbb{V}(T)$ directly follow from the proposed assumptions. According to Assumption 6.5, the number of observations in each terminal node is bounded above. This implies that the variance of the tree is bounded above by constant times $\mathbb{V}(Y_{im}|g(\mathbf{X}_i) = \gamma)$. Moreover, Assumption 6.3 guarantees that $\mathbb{V}(Y_{im}|g(\mathbf{X}_i) = \gamma)$ is bounded away from zero.

In the context at hand, we rely on the findings presented by Wager and Athey (2018) concerning the order of variance terms. Specifically, they show that:

$$\mathbb{V}(\dot{T})_{ii} = \frac{C}{\log^G(s)}, \text{ for some constant } C. \tag{36}$$

$\mathbb{V}(\dot{T})_{ii}$ denotes the diagonal terms of the variance of the projection of a tree estimator. We show that the off-diagonal terms $\mathbb{V}(\dot{T})_{ij} = o(\frac{1}{\log^G(s)})$ for all $i \neq j$.

Start with the definition of a Hajek projection of a tree:

$$\dot{T} - \mu = \sum_{i=1}^{s} \mathbb{E}(T|A_i) \tag{37}$$

Since the observations are *i.i.d.*, then:

$$\mathbb{V}\big(\dot{T}\big) = s\mathbb{V}\big(\mathbb{E}(T|A_1)\big). \tag{38}$$

Then it is clear to see that:

$$\mathbb{V}\big(\mathbb{E}(T|A_1)\big) = \mathbb{V}\big(\mathbb{E}(T|A_1) - \mathbb{E}(T|g(\mathbf{X}_1))\big) + \mathbb{V}\big(\mathbb{E}(T|g(\mathbf{X}_1)\big). \tag{39}$$

Consider m-th outcome variable, where $m = 1, \dots, M$. Since the tree is honest, the diagonal terms in (39) simplify as follows (see the Proof of Theorem 5 in Wager and Athey, 2018):

$$\begin{aligned} \mathbb{V}\big(\mathbb{E}(T|A_1) - \mathbb{E}(T|g(\mathbf{X}_1))\big)_{mm} = \mathbb{V}\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))(Y_{1m} - \mathbb{E}(Y_{1m}|g(\mathbf{X}_1))\big)_{mm} \approx \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big]\mathbb{E}\big[\big(Y_{1m} - \mathbb{E}(Y_{1m}|g(\mathbf{X}_1))\big)^2\big] = \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] Var(Y_m|g(\mathbf{X}_1) = \gamma), \end{aligned} \tag{40}$$

and

$$\mathbb{V}\big(\mathbb{E}(T|g(\mathbf{X}_1)\big)_{mm} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] Var\big(\mathbb{E}(Y_m|g(\mathbf{X}_1) = \gamma)\big). \tag{41}$$

where S_{ℓ_n} is the indicator function and equals one if $g(\mathbf{X}_1) \in \ell_n(\gamma, \Pi)$, and zero otherwise.

The off-diagonal terms equal to:

$$\mathbb{V}\big(\mathbb{E}(T|A_1) - \mathbb{E}(T|X_1)\big)_{mm'} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big]\mathbb{E}\big[(Y_{1m} - \mathbb{E}(g(\mathbf{X}_1)))(Y_{1m'} - \mathbb{E}(Y_{1m'}|g(\mathbf{X}_1)))\big]. \tag{42}$$

According to Assumption 6.3, the variance of each outcome variable is bounded away from zero. Cauchy-Schwarz inequality implies that $|Cov(Y_{im}, Y_{im'}|g(\mathbf{X}_1))|$ is also bounded away from zero [6].

$$|Cov(Y_{1m}, Y_{1m'}|g(\mathbf{X}_1) = \gamma)| \leq \sqrt{Var(Y_{1m}|g(\mathbf{X}_1) = \gamma)Var(Y_{1m'}|g(\mathbf{X}_1) = \gamma)}.$$

Wager and Athey (2018) show that

$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] \geq \frac{(G-1)!}{2^{G+1}\log^G(s)} \cdot \frac{1}{ks}, \tag{43}$$

where k is the minimum number of observations in a given terminal node. Combining (38) and (43) yields the order of diagonal and off-diagonal terms:

$$\mathbb{V}(\dot{T})_{mm} = o\left(\frac{1}{log^G(s)}\right), \text{ and } \mathbb{V}(\dot{T})_{mm'} = o\left(\frac{1}{log^G(s)}\right). \tag{44}$$

Now we prove that $\frac{s}{N} tr\left((\mathbb{V}(\dot{T}))^{-1}\mathbb{V}(T)\right) \to 0$ in a more general framework. Consider, we have two square matrices C and B with diagonal (c_{ii}, b_{ii}) and non-diagonal terms (c_{ij}, b_{ij}), respectively. Moreover, they exhibit the following proper-

[6] An alternative argument is to notice that the term in the integrand consists of multiples of the first and second moments of the outcome variables Y_{1m} and $Y_{1m'}$. Since these moments are continuous, they are bounded. Thus, their expectation is also bounded.

ties:

$$1.\ b_{ii} \geq \eta \text{ for some } \eta \in \mathbb{R}^+ \text{ and for all } i = 1, \dots M, \tag{45}$$

$$2.\ c_{ii} \geq \frac{b_{ii}}{\log(N)}, \tag{46}$$

$$3.\ c_{ij} = o\left(\frac{1}{\log(N)}\right). \tag{47}$$

Then we show that $\frac{s}{N} tr(C^{-1}B) \to 0$. Recall that the Leibniz formula for the determinant is given as follows:

$$det(C) = \sum_{\pi} \left(\text{sgn}(\pi) \prod_{i=1}^{M} c_{i,\pi_i} \right), \tag{48}$$

where π is a permutation function that reorders the set $\{1, \dots, M\}$. Diagonal and off-diagonal terms are on the same order, their product is also on the same order. Therefore, $det(C)$ is asymptotically equivalent to either $\prod_{i=1}^{M} c_{ii}$ or $\prod_{i=1}^{M} c_{ij}$ where $i \neq j$. For simplicity, we keep the notation that $det(C) \sim^a \prod_{i=1}^{M} c_{ii}$, where $'' \sim^a ''$ denotes asymptotic equivalence. Based on Cramer's rule, we can write i-th diagonal term of the inverse of C:

$$(C^{-1})_{ii} = \frac{det(C_{-i})}{det(C)}.$$

C_{-i} is the matrix where we remove the i-th row and the i-th column. By the same argument, $det(C_{-i}) \sim^a \prod_{j=1}^{M-1} c_{jj}$. Then we end up with:

$$(C^{-1})_{ii} \sim^a \frac{\prod_{j=1}^{M-1} c_{jj}}{\prod_{i=1}^{M} c_{ii}} = \frac{1}{c_{ii}}.$$

The i-th diagonal entry of the matrix

$$(C^{-1}B)_{ii} = (c^{-1})_{ii} b_{ii} + \sum_{j \neq i} (c^{-1})_{ij} b_{ji} \sim^a \frac{b_{ii}}{c_{ii}} \leq \log(N).$$

The last equality follows from Property 2 in (46). Therefore, the trace of $(C^{-1}B)$ is also on the order of $\log(N)$. We take the limit of $\frac{s}{N} tr(C^{-1}B)$, where $s = N^{\beta}$ and

$\beta < 1$. L'Hôpital's rule yields:

$$\lim_{N\to\infty} \frac{s}{N}\log(N) = \lim_{N\to\infty} \frac{1}{(1-\beta)N^{1-\beta}} \to 0. \tag{49}$$

The proof is complete by letting $C = \big(\mathbb{V}(\dot{T})\big)^{-1}$ and $B = \mathbb{V}(T)$.

□

A.0.4 Outcomes Correlated Across Groups

Denote two different leaves as ℓ and ℓ'. The aim is to show that

$$\frac{s}{N} tr\big(Var(\dot{T})^{-1} Var(T)\big) \to 0.$$

Consider the Hajek projection of a tree

$$\dot{T} - \mu = \sum_{i=1}^{s} \mathbb{E}(T|A_i), \text{ so that } Var(\dot{T}) = sVar(\mathbb{E}(T|A_1).$$

The last equality follows as the data $A_i = (\mathbf{Y}_i, g(\mathbf{X}_i))$ are *i.i.d.* Moreover, the variance of the conditional expected tree can be expanded as:

$$Var\mathbb{E}(T|A_1) = Var\big[\mathbb{E}(T|A_1) - \mathbb{E}(T|g(\mathbf{X}_i))\big] + Var\big[\mathbb{E}(T|g(\mathbf{X}_i))\big].$$

The algorithm is honest, therefore, the difference $\mathbb{E}(T|A_1) - \mathbb{E}(T|g(\mathbf{X}_i))$ simplifies to

$$\mathbb{E}(T_\ell|A_1) - \mathbb{E}(T_\ell|g(\mathbf{X}_1)) = \mathbb{E}(S_\ell|g(\mathbf{X}_i))(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, g(\mathbf{X}_i))).$$

T_ℓ is a tree estimate at γ_ℓ and S_ℓ is an indicator of whether $g(\mathbf{X}_1)$ and γ_ℓ belong to the same terminal node. The outcome Y_{m1} denotes the m-th outcome for the first observation. Note that the covariance matrix in each tree now consists of not only the covariance between the outcomes within a leaf but across terminal nodes as well. Therefore, we focus on the covariance of an outcome from a leaf ℓ with the same outcome from another leaf ℓ' and with another outcome $Y_{m'}$ from another leaf ℓ':

$$\mathbb{E}\big[\mathbb{E}(S_\ell|g(\mathbf{X}_i))\mathbb{E}(S_{\ell'}|g(\mathbf{X}_i))(Y_{m1}-\mathbb{E}(Y_{m1}|S_\ell=1,g(\mathbf{X}_i)))(Y_{m1}-\mathbb{E}(Y_{m1}|S_{\ell'}=1,g(\mathbf{X}_i)))\big],$$
$$\mathbb{E}\big[\mathbb{E}(S_\ell|g(\mathbf{X}_i))\mathbb{E}(S_{\ell'}|g(\mathbf{X}_i))(Y_{m1}-\mathbb{E}(Y_{m1}|S_\ell=1,g(\mathbf{X}_i)))(Y_{m'1}-\mathbb{E}(Y_{m'1}|S_{\ell'}=1,g(\mathbf{X}_i)))\big].$$

The terms $\mathbb{E}(Y_{m1}-\mathbb{E}(Y_{m1}|S_\ell = 1,g(\mathbf{X}_i)))(Y_{m1}-\mathbb{E}(Y_{m1}|S_{\ell'} = 1,g(\mathbf{X}_i)))$ and $\mathbb{E}(Y_{m1}-\mathbb{E}(Y_{m1}|S_\ell = 1,g(\mathbf{X}_i)))(Y_{m'1}-\mathbb{E}(Y_{m'1}|S_{\ell'} = 1,g(\mathbf{X}_i)))$ are the polynomials of degree at most two. Since the first and second moments of the outcome are bounded, these terms are also bounded. Next, Cauchy-Schwarz inequality implies

$$\sqrt{\mathbb{E}\big[\mathbb{E}(S_\ell|g(\mathbf{X}_i))\big]^2\mathbb{E}\big[\mathbb{E}(S_{\ell'}|g(\mathbf{X}_i))\big]^2} \leq \sqrt{\mathbb{E}\big[\mathbb{E}(S_\ell|g(\mathbf{X}_i))\mathbb{E}(S_{\ell'}|g(\mathbf{X}_i)\big)}.$$

Therefore, the lower bound of the covariance is on the order $o\left(\frac{1}{log^G(s)}\right)$. Equivalently, we can show that the off-diagonal terms of $Var\big[\mathbb{E}(T|g(\mathbf{X}_i))\big]$ are on the same order. Then Theorem 6.1 applies, and the asymptotic normality of the random forest estimator holds.

A.0.5 Quantiles of the Outcomes

This section generalises theory of the multivariate causal forest to accomodate quantiles rather than means of the outcome. Meinshausen and Ridgeway (2006) defines the density of the $m-$th outcome variable as follows:

$$\mathbb{E}(y|g(\mathbf{X}_i)=\gamma) = \mathbb{P}(Y_{im}\leq y|g(\mathbf{X}_i)=\gamma) = \mathbb{E}(1_{\{Y_{im}\leq y\}}|g(\mathbf{X}_i)=\gamma).$$

$1_{\{Y_{im}\leq y\}}$ is the binary variable indicating the outcome is weakly less than some value y. Just as $\mathbb{E}(Y_{im}|g(\mathbf{X}_i)=\gamma)$ is approximated by the weighted mean over the observations of Y_{im}, define the approximation to $\mathbb{E}(1_{\{Y_{im} \le y\}} | g(\mathbf{X}_i) = \gamma)$ by a tree as

$$\hat{F}(y|1_{\{Y_{im} \le y\}} = \gamma) = T(\gamma, \xi, A_i, \ldots, A_N) = \sum_{n=1}^{|\Pi|} 1(\gamma \in \ell_n) 1_{\{Y_{im} \le y\}}.$$

Under the assumption that the Hoeffding decomposition exists for quantile regression forests with multiple outcomes, Lemma 6.1 and Lemma 6.2 and their corresponding proofs in Appendix A.0.1 and A.0.2 apply directly. The goal is to show that

$$\lim_{N \to \infty} \frac{s}{N} tr\Big(\big(\mathbb{V}(\dot{T})\big)^{-1} \mathbb{V}(T) \Big) = 0. \tag{50}$$

In Theorem 6.1, we introduce a Hajek projection of a tree, and show the convergence of the deviation of the multivariate forest and its projection to zero. The proof is analogous, except the variance of the tree is defined as

$$\begin{aligned} \mathbb{V}\big(\mathbb{E}(T|A_1) - \mathbb{E}(T|g(\mathbf{X}_1))\big)_{mm} = \mathbb{V}\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))(1_{\{Y_{1m} \le y\}} - \mathbb{E}(1_{\{Y_{1m} \le y\}}|g(\mathbf{X}_1))\big)_{mm} \approx \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] \mathbb{E}\big[\big(1_{\{Y_{1m} \le y\}} - \mathbb{E}(1_{\{Y_{1m} \le y\}}|g(\mathbf{X}_1))\big)^2\big] = \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] Var(1_{\{Y_m \le y\}}|g(\mathbf{X}_1)) \le \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big], \end{aligned} \tag{51}$$

and

$$\mathbb{V}\big(\mathbb{E}(T|g(\mathbf{X}_1)\big)_{mm} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] Var\big(\mathbb{E}(1_{\{Y_{1m} \le y\}}|g(\mathbf{X}_1))\big) \le \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big], \tag{52}$$

where as defined before, S_{ℓ_n} is an indicator function and equals one if $g(\mathbf{X}_1) \in \ell_n(\gamma, \Pi)$, and zero otherwise.

The off-diagonal terms equal to:

$$\mathbb{V}\big(\mathbb{E}(T|A_1) - \mathbb{E}(T|g(\mathbf{X}_1))\big)_{mm'} = \tag{53}$$
$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big]\mathbb{E}\big[(1_{\{Y_{1m}\leq y\}} - \mathbb{E}(1_{\{Y_{1m}\leq y\}}|g(\mathbf{X}_1)))(1_{\{Y_{1m'}\leq y\}} - \mathbb{E}(1_{\{Y_{1m'}\leq y\}}|g(\mathbf{X}_1)))\big]$$
$$\leq \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big].$$

The last inequality in each case follows by the fact that the expectation and variance of $1_{\{Y_{1m}\leq y\}}$ is bounded by one from above, as it is a binary variable. The rest of the proof is analogous to Theorem 6.1.

A.0.6 Theoretical Results for Group Average Policy Effects

The proof is equivalent for the multivariate causal forest with group average policy effects. Wager and Athey (2018) show that for the policy effect,

$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|g(\mathbf{X}_1))\big)^2\big] \geq \frac{(p-1)!}{2^{p+1}\log^p(s)} \cdot \frac{\epsilon}{ks}, \tag{54}$$

where, ϵ is a constant from Assumption 6.6. This does not change the results of the proofs, as the order of $\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|(\mathbf{X}_1))\big)^2\big]$ is still $o\left(\frac{1}{\log^p(s)}\right)$.

A.0.7 Method of Moments

Rescaling the moment function and additional algebraic transformations lead to a simplified unbiased estimator:

$$\mathbb{E}_{S^{tr},S^{est}}\Big[\big(\theta_g - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)^T \Sigma^{-1} \big(\theta_g - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big) - \theta_g^T \Sigma^{-1} \theta_g\Big] = \tag{55}$$

$$\mathbb{E}_{S^{tr},S^{est}}\Big[\big(\underbrace{\theta_g - \theta(g(\mathbf{X}_i), \mathit{\Pi})}_{A} + \underbrace{\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \tilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})}_{B}\big)^T \Sigma^{-1} \big(\underbrace{\theta_g - \theta(g(\mathbf{X}_i), \mathit{\Pi})}_{A} +$$

$$\underbrace{\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \tilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})}_{B}\big) - \theta_g^T \Sigma^{-1} \theta_g\Big] = \tag{56}$$

$$\mathbb{E}_{S^{tr}}\Big(\theta_g^T \Sigma^{-1} \theta_g - 2\theta_g^T \Sigma^{-1} \theta(g(\mathbf{X}_i), \mathit{\Pi}) +$$

$$\theta(g(\mathbf{X}_i), \mathit{\Pi})^T \Sigma^{-1} \theta(g(\mathbf{X}_i), \mathit{\Pi}) - \theta_g^T \Sigma^{-1} \theta_g\Big) + \tag{57}$$

$$\mathbb{E}_{g(\mathbf{X}_i),S^{est}}\Big(\big(\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)^T \Sigma^{-1} \big(\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)\Big) = \tag{58}$$

$$- \mathbb{E}_{g(\mathbf{X}_i)}\big(\theta(g(\mathbf{X}_i), \mathit{\Pi})^T \Sigma^{-1} \theta(g(\mathbf{X}_i), \mathit{\Pi})\big) + \mathbb{E}(tr(I))_{2\times 2}. \tag{59}$$

The second equality follows after taking into account the independence of the train and estimation data, $cov(A, B) = 0$. The final equality is based on the fact that $\theta(g(\mathbf{X}_i), \mathit{\Pi}) = \mathbb{E}(\theta_g | g(\mathbf{X}_i) \in \ell(\gamma, \mathit{\Pi}))$, $\mathbb{E}\big(\widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big) = \theta(g(\mathbf{X}_i), \mathit{\Pi})$, and:

$$\mathbb{E}_{g(\mathbf{X}_i),S^{est}}\Big(\big(\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)^T \Sigma^{-1} \big(\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)\Big) =$$

$$tr\Big(\Sigma^{-1} \mathbb{E}\big(\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)^T \big(\theta(g(\mathbf{X}_i), \mathit{\Pi}) - \widetilde{\theta}(g(\mathbf{X}_i), S^{est}, \mathit{\Pi})\big)\Big) =$$

$$tr(\Sigma^{-1}\Sigma) = tr(I)_{2\times 2},$$

where $tr(I)_{2\times 2}$ is the trace of a 2×2 identity matrix. Since $\mathbb{E}(tr(I))_{2\times 2}$ does not depend on the parameter of interest, we can disregard it. Hence, the optimal parameter maximizes the unbiased estimator of the negative mean squared error:

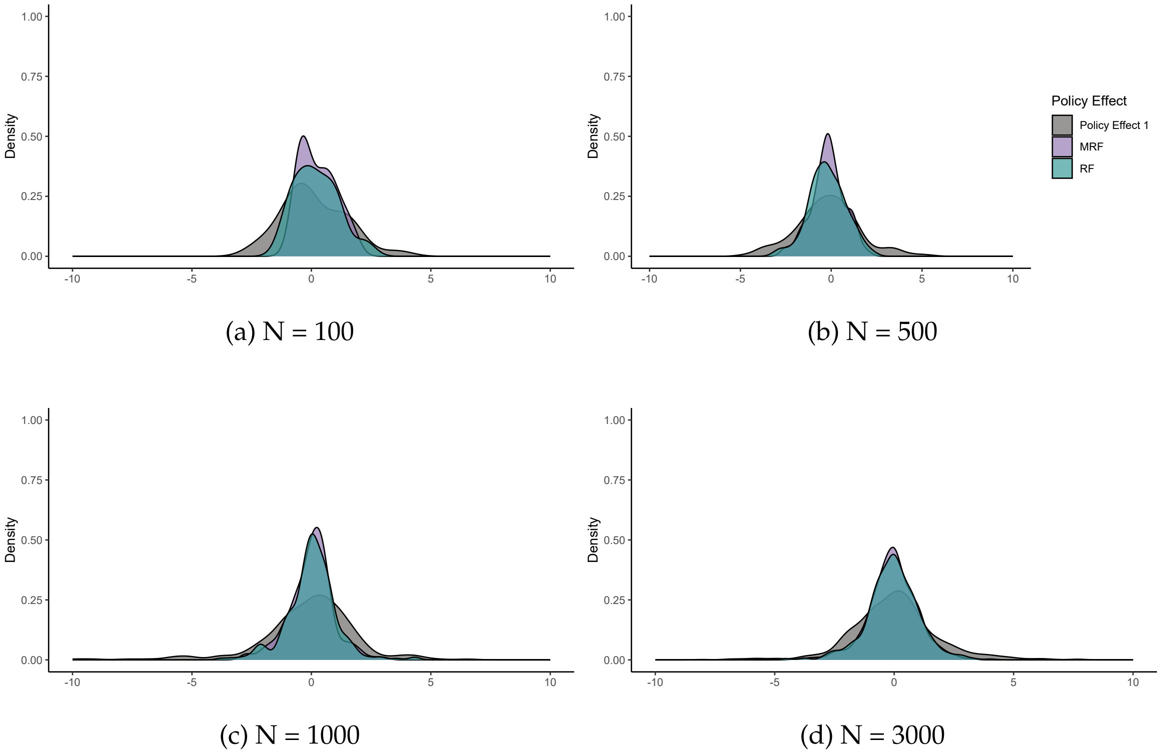
$$\hat{\theta}(\gamma, S^{est}, \mathit{\Pi}) = \arg\max_{\widetilde{\theta}} \frac{1}{N^{tr}} \sum_{\ell} N_\ell^{tr} \big(\widetilde{\theta}(g(\mathbf{X}_i), \mathit{\Pi})^T \widehat{\Sigma}^{-1} \widetilde{\theta}(g(\mathbf{X}_i), \mathit{\Pi}) | g(\mathbf{X}_i) = \gamma\big), \tag{60}$$

where the covariance matrix can be estimated as $\hat{\Sigma} = \hat{\Sigma}\big(\tilde{\theta}(g(\mathbf{X}_i), S^{tr}, \Pi)|N^{est}\big)$. In this article, training and estimation samples have an equal number of observations, $N^{tr} = N^{est}$.

A.1 Additional Simulations

In this section, we replicate the design of Section 8. However, the independent characteristics come from the preferential attachment algorithm (Jeong et al., 2003). Each node of the network represents one feature. The resulting network follows a power-law degree distribution, and thus, is scale-free. That means, only a few variables (characteristics) in the network have a relatively large number of ”neighbors”. The distance between two characteristics is the shortest path between them in the network. We calculate a $p \times p$ ($p = 50$) pairwise distance matrix L. Next, this distance matrix is transformed into a covariance matrix $\Sigma_{z,(i,j)} = 0.5^{L_{(i,j)}}$, where (i, j) represents the element in each row i and column j of a matrix L $(i, j = 1, \dots, p)$.

According to Figures 17 and 22, in practice, multivariate random forests and random forests do not significantly differ from each other in this setting.



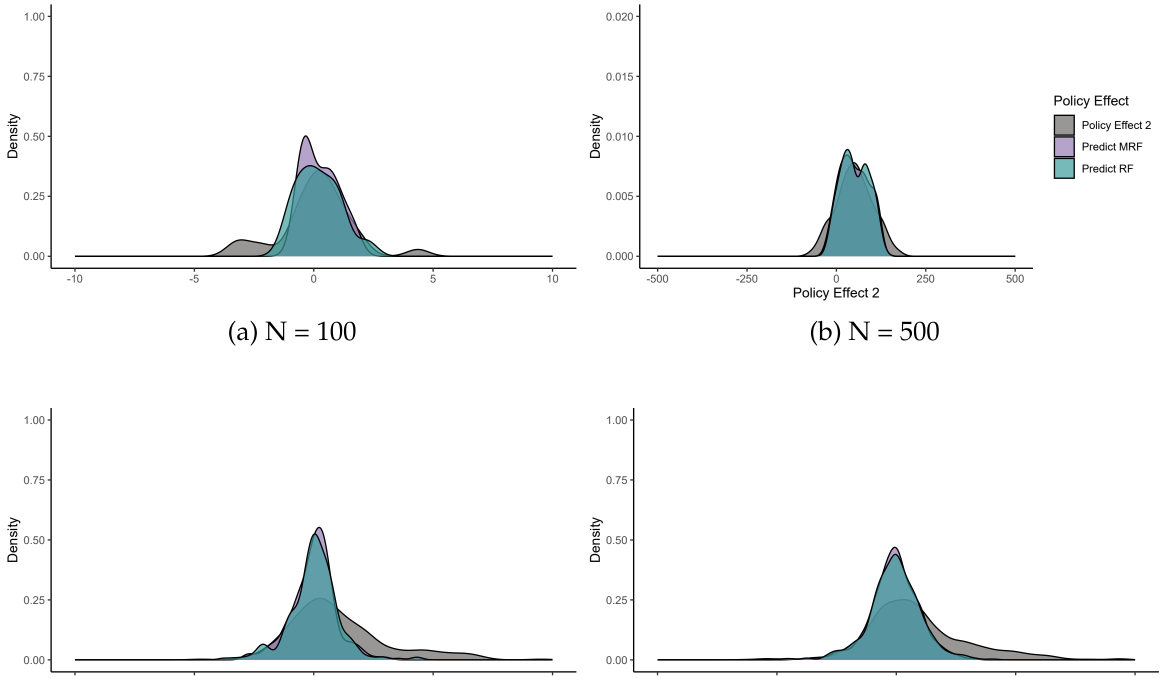
(a) N = 100

(b) N = 500

(c) N = 1000

(d) N = 3000

Figure 17: The density of simulated and estimated policy effects ($X_{i3} \cdot X_{i4}$) based on Y_{i1} for various number of samples denoted by $N = 100, 500, 1000, 3000$. The orange color corresponds to the estimated density of policy effects through multivariate random forests (labeled as MRF). On the other hand, the color green is employed to visually illustrate the estimated density of policy effects derived from classical random forests (labeled as RF). We train and predict the methods on two different sets of data.



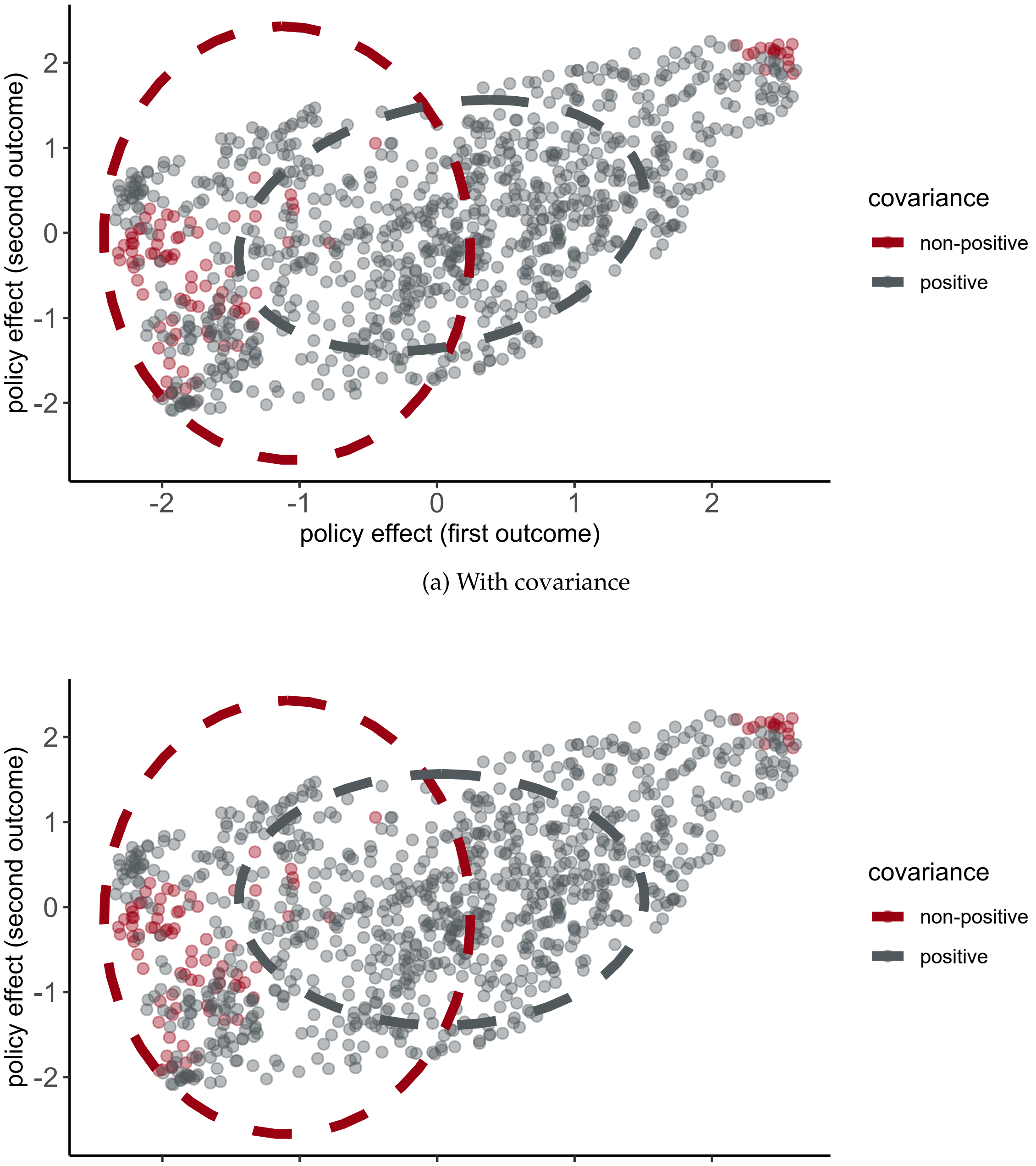
(a) N = 100

(b) N = 500

(c) N = 1000

(d) N = 3000

Figure 18: The density of simulated and estimated policy effects ($X_{i1} \cdot X_{i2}$) based on Y_{i2} for various number of samples denoted by $N = 100, 500, 1000, 3000$. The orange color corresponds to the estimated density of policy effects through multivariate random forests (labeled as MRF). On the other hand, the color green is employed to visually illustrate the estimated density of policy effects derived from classical random forests (labeled as RF). We train and predict the methods on two different sets of data.



(a) With covariance

(b) Without Covariance

Figure 19: Panel (a) and (b) correspond to confidence ellipses with and without incorporating covariance, respectively. Data are grouped according to the estimated sign of covariance.

A.2 Optimal Number of Components

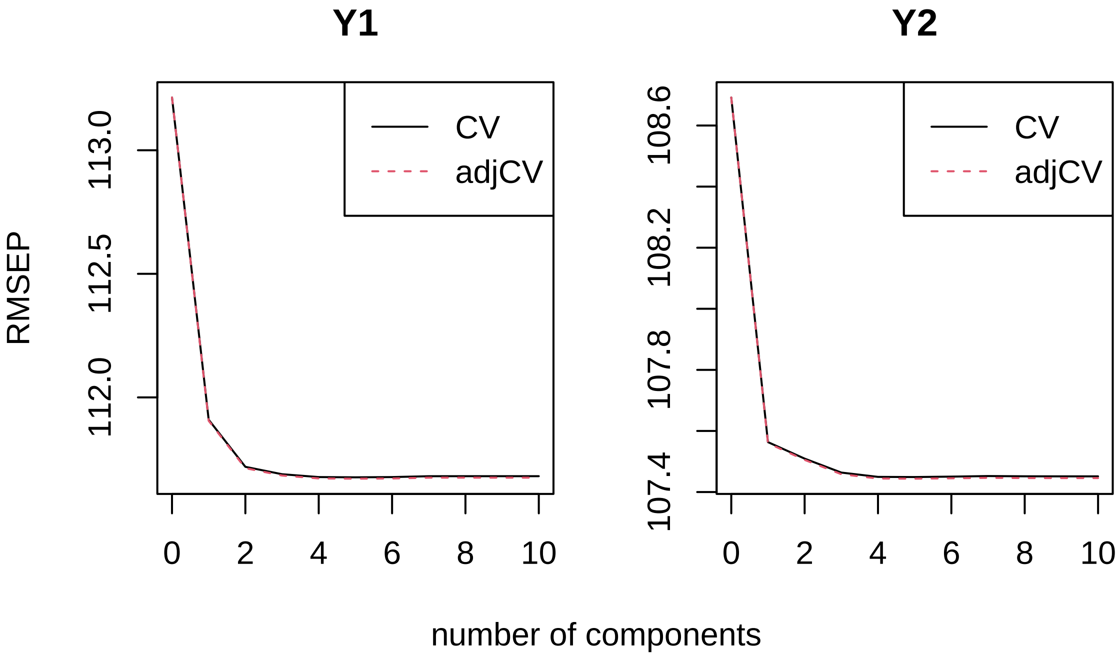
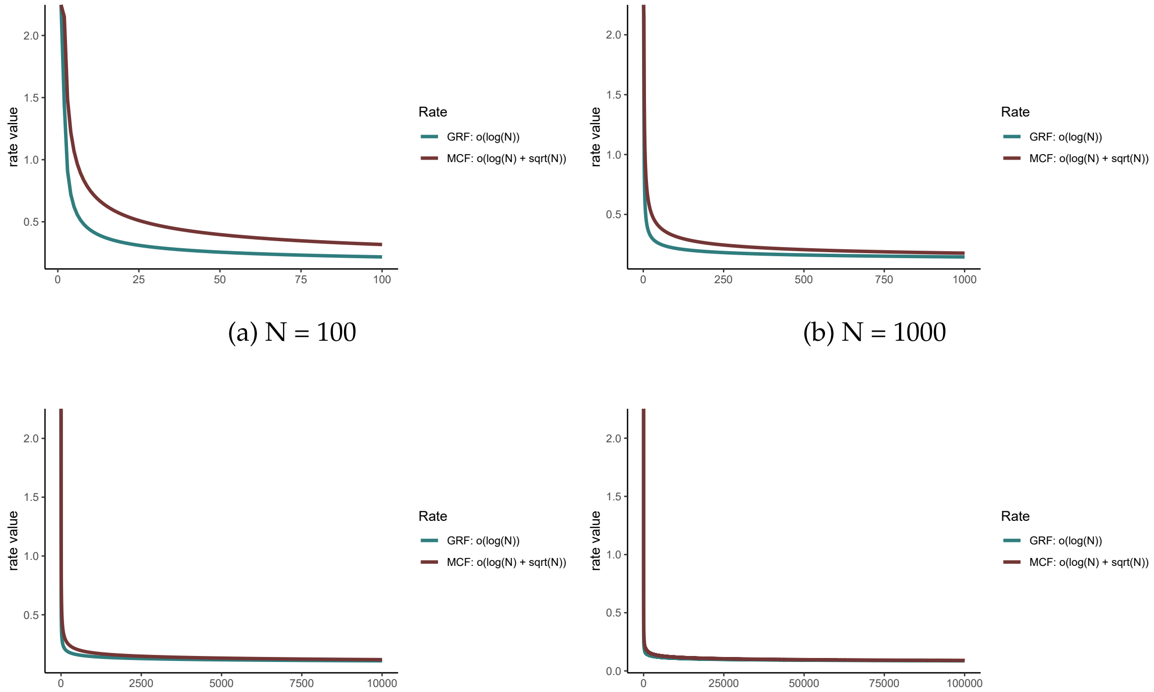


Figure 20: Cross-validation results. The optimal number of groups is the minimum number of groups with the least root mean squared error. See Nareklishvili et al. (2022) for further details.

A.3 The Rate of Convergence



(a) N = 100

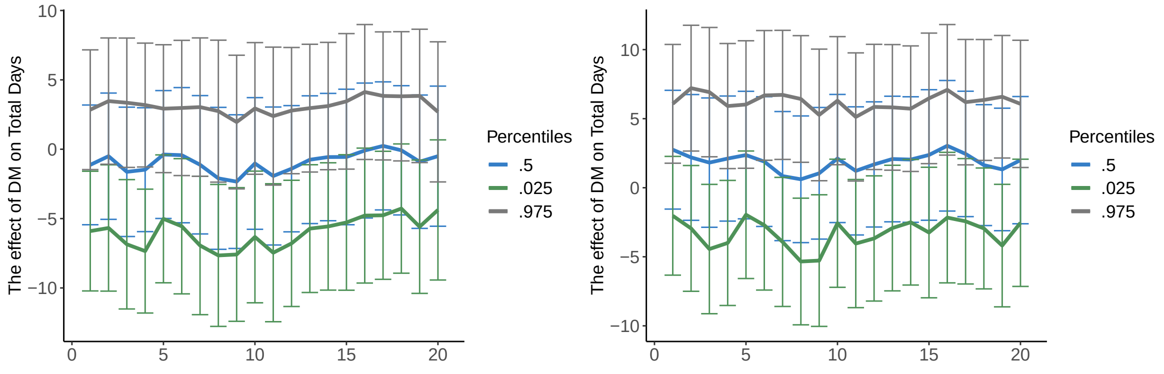
(b) N = 1000

(c) N = 10,000

(d) N = 100,000

Figure 21: Rates of convergence for generalized random forest (GRF; Athey et al., 2019) and multivariate causal forest (MCF) in this paper.

A.3.1 Causal Forest

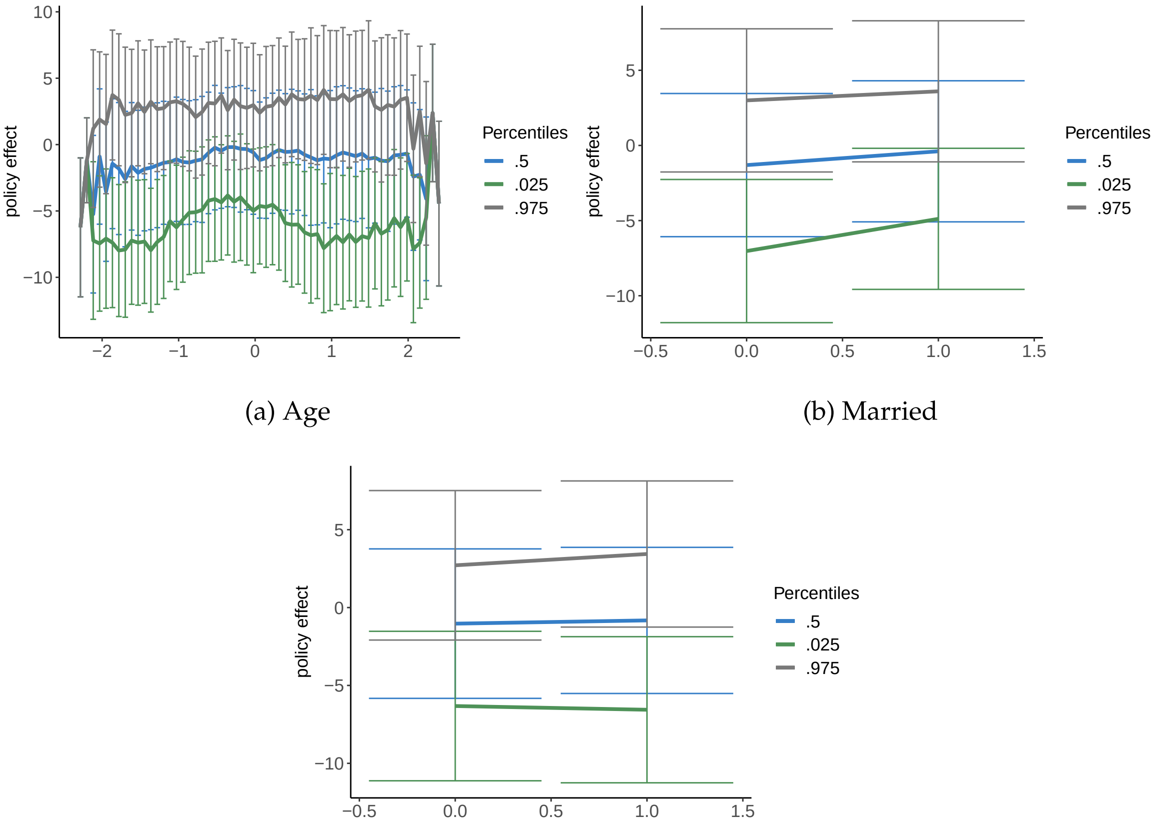


(a) Group 2 vigintiles (total days of sick leave)

(b) Group 2 vigintiles (days of sick leave within spell)

Figure 22: Heterogeneity of the effect of DMs on the duration of sick leave. The densities are estimated by the GRF algorithm with 1000 trees (Athey and Wager, 2019).

A.3.2 Causal Forest: Heterogeneity across Individual Characteristics



(a) Age

(b) Married

(c) Female

Figure 23: Heterogeneity of the effect of DMs on total days of sick leave. The densities are estimated by the GRF algorithm with 1000 trees (Athey and Wager, 2019).