UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Advancing Robust and Reliable Estimation of Heterogeneous Treatment Effects: Methodological Innovations and Critical Evaluations

Hugo Gobato Souto

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Hugo Gobato Souto

Advancing Robust and Reliable Estimation of Heterogeneous Treatment Effects: Methodological Innovations and Critical Evaluations

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
September 2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G574a

Gobato Souto, Hugo
Advancing Robust and Reliable Estimation of Heterogeneous Treatment Effects: Methodological Innovations and Critical Evaluations / Hugo Gobato Souto; orientador Francisco Louzada Neto. -- São Carlos, 2025.
182 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Heterogeneous Treatment Effects. 2. Simulation Studies. 3. Bayesian Trees. 4. Difference-in-Differences. 5. Statistical Testing. I. Louzada Neto, Francisco, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Hugo Gobato Souto

Avançando a Estimação Robusta e Confiável de Efeitos de Tratamento Heterogêneos: Inovações Metodológicas e Avaliações Críticas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Setembro de 2025

São Carlos - SP, 08 de Setembro de 2025.

Ref: Carta comprovante da versão final de teses e dissertações

Eu Prof(a). Dr(a). Francisco Louzada Neto, Orientador(a) do(a) pós-graduando(a) Hugo Gobato Souto, do Programa de Pós-Graduação PIPGEs, venho por meio desta, AUTORIZAR os trâmites para a homologação da tese/dissertação do(a) supracitado(a) aluno(a), e ATESTAR que a tese/dissertação intitulada "Advancing Robust and Reliable Estimation of Heterogeneous Treatment Effects: Methodological Innovations and Critical Evaluations", é a **versão final** com as alterações sugeridas pela Banca Examinadora, estando o arquivo tecnicamente correto em sua forma e estrutura e com os devidos agradecimentos aos órgãos de fomento à pesquisa, no caso de recebimento de bolsa e/ou financiamento.

Solicito as devidas providências para o encaminhamento em questão, subscrevo-me.

Atenciosamente,

gov.br Documento assinado digitalmente
FRANCISCO LOUZADA NETO
Data: 10/09/2025 10:50:36-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura orientador(a)

Este trabalho é dedicado ao meu pai, Sérgio Paulo Amaral Souto, e mãe, Yara Galvão Gobato,

que sempre me apoiaram e foram grandes inspirações.

Last but not least, I want to thank God for all the wonders that he gave me.

Ik besef wel dat ik loterijticket na loterijticket gewonnen heb in mijn leven.

Et j'essaie tout simplement de vivre la vie que Dieu me reserve et de faire le bien partout où je passe.

Si Dios está conmigo, ¿quién contra mí?

ACKNOWLEDGEMENTS

Firstly, I would like to thank Carlos M. Carvalho to spark my curiosity about Causal Inference during his talk in the ESOBE 2023 at the University of Glasgow. Secondly, I would like to thank Drew Herren and P. Richard Hahn for creating the StochTree repository, which was certainly highly important for the research performed in this thesis [1]. Thirdly, I would like to thank CAPES and FAPESP for their Master's scholarships [2].

Last but far from least, I would like to thank my group of friends, called GR, and consisting of Alvaro Vale Bolsonaro, André Hoffmann, Bruno Barbosa, David Talhati, Enzo Lucca Gabassa, João Pedro Bahu, João Pedro Milani, João Pedro Salazar, Levy Scalli Neto, Thales Nordi, and Reginaldo Claro Gobato Junior. They have always been there for me to make my days brighter.

[1] In special, a huge thanks for Drew Herren for always helping me with both theoretical and practical questions about the BCF model and other CATE estimation models

[2] CAPES scholarship (88887.949149/2024-00) from March 2024 until July 2024 and FAPESP scholarship (2024/06274-0) from July 2024 until February 2025 (stopped as I started working at Magazine Luiza as Causal Inference Scientist).

"…you can give me nothing now yet I love you so that you stand in my way of loving anyone else —
but I want you to stand there. You, dead, are so much better than anyone else alive."
(Richard Feynman)

"The first principle is that you must not fool yourself –
and you are the easiest person to fool."
(Richard Feynman)

RESUMO

SOUTO, H. G. **Avançando a Estimação Robusta e Confiável de Efeitos de Tratamento Heterogêneos: Inovações Metodológicas e Avaliações Críticas**. 2025. 180 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Contexto: A estimação própria de Efeitos de Tratamento Heterogêneos (HTEs) é crucial em muitas disciplinas científicas, mas enfrenta desafios metodológicos significativos, incluindo complexidade do modelo, confounding e carga computacional. **Objetivos**: Esta tese teve como objetivo avançar o campo da estimação de HTEs, desenvolvendo e avaliando criticamente metodologias que aprimoram as técnicas e práticas de inferência causal. **Contribuições Metodológicas**: O trabalho apresenta três contribuições principais: (1) Uma validação empírica da importância de estudos de ablação para modelos causais não-paramétricos complexos, examinando especificamente a Bayesian Causal Forest (BCF) e o papel de seu componente de propensity score; (2) O desenvolvimento do Test-Informed Simulation Count Algorithm (TISCA), uma abordagem para determinar o número necessário de replicações em estudos de simulação para avaliação de modelos usando princípios estatísticos; e (3) A introdução da Floresta Causal Bayesiana de Diferenças em Diferenças (DiD-BCF), um novo estimador não-paramétrico para inferência causal robusta em configurações de DiD, abordando particularmente de forma eficaz a heterogeneidade do efeito do tratamento por meio de uma reparametrização baseada na Parallel Trends Assumption (PTA). **Principais Descobertas**: Estudos de ablação revelaram que o componente de propensity score na BCF não é essencial para o desempenho e sua omissão pode reduzir o tempo de computação em aproximadamente 21%. O TISCA demonstrou fornecer contagens de simulação estatisticamente justificadas. O DiD-BCF demonstrou desempenho consideravelmente superior em relação aos benchmarks estabelecidos e revelou efeitos de tratamento condicionais sutis em uma aplicação empírica à política de salário mínimo dos EUA. **Conclusão Geral e Implicações**: Esta tese defende coletivamente um paradigma de maior rigor, eficiência e compreensão nuances na estimação de HTEs. Ela fornece aos pesquisadores insights criticamente avaliados e novas ferramentas — defesa de estudos de ablação, um algoritmo de design de simulação estatisticamente fundamentado e um estimador DiD não-paramétrico avançado — para gerar evidências causais mais robustas, confiáveis e acionáveis, fortalecendo assim a base para a tomada de decisões baseada em evidências em várias disciplinas.

Palavras-chave: Efeitos de Tratamento Heterogêneos, Inferência Causal, Métodos Não Paramétricos, Estudos de Simulação, Árvores Bayesianas.

ABSTRACT

SOUTO, H. G. **Advancing Robust and Reliable Estimation of Heterogeneous Treatment Effects: Methodological Innovations and Critical Evaluations**. 2025. 180 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Context: The robust and reliable estimation of Heterogeneous Treatment Effects (HTEs) is crucial across many scientific disciplines, yet it faces significant methodological challenges including model complexity, confounding, computational burden, and rigorous evaluation practices. **Objectives**: This thesis aimed to advance the field of HTE estimation by developing and critically evaluating methodologies that enhance the rigor, efficiency, and practical utility of causal inference techniques. **Methodological Contributions**: The work presents three primary contributions: (1) An empirical validation of the importance of ablation studies for complex nonparametric causal models, specifically examining the Bayesian Causal Forest (BCF) and the role of its propensity score component; (2) The development of the Test-Informed Simulation Count Algorithm (TISCA), a principled approach for determining the necessary number of replications in simulation studies for model evaluation using statistical principles; and (3) The introduction of the Difference-in-Differences Bayesian Causal Forest (DiD-BCF), a novel non-parametric estimator for robust causal inference in DiD settings, particularly effectively addressing treatment effect heterogeneity through a Parallel Trends Assumption (PTA)-based reparameterization. **Principal Findings**: Ablation studies revealed that the propensity score component in BCF is not essential for performance and its omission can reduce computation time by approximately 21%. TISCA was shown to provide statistically justified simulation counts, promoting efficiency and enhancing the credibility of comparative model evaluations. DiD-BCF demonstrated considerably superior performance over established benchmarks and uncovering nuanced conditional treatment effects in an empirical application to U.S. minimum wage policy. **Overall Conclusion and Implications**: This thesis collectively champions a paradigm of increased rigor, efficiency, and nuanced understanding in HTE estimation. It provides researchers with critically evaluated insights and novel tools—ablation study advocacy, a statistically grounded simulation design algorithm, and an advanced non-parametric DiD estimator—to generate more robust, reliable, and actionable causal evidence, thereby strengthening the foundation for evidence-based decision-making across various disciplines.

Keywords: Heterogeneous Treatment Effects, Causal Inference, Nonparametric Methods, Simulation Studies, Bayesian Trees.

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ATE	Average Treatment Effect
ATT	Average Treatment effect on the Treated
BART	Bayesian Additive Regression Trees
BCF	Bayesian Causal Forest
BH	Benjamini-Hochberg
CATE	Conditional Average Treatment Effect
CATT	Conditional Average Treatment effect on the Treated
CFFE	Causal Forest with Fixed Effects
CHTE	Covariate-Heterogeneous Treatment Effects
CIL	Confidence Interval Length
CLT	Central Limit Theorem
DGP	Data-Generating Process
DiD	Difference-in-Differences
DiD-BCF	Difference-in-Differences Bayesian Causal Forest
DR	Doubly Robust
GATE	Group Average Treatment Effect
GATT	Group Average Treatment effects on the Treated
HTE	Heterogeneous Treatment Effect
IPW	Inverse Probability Weighting
ITC	Investment Tax Credit
LLM	Large Language Model
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCMC	Markov chain Monte Carlo
MDE	Minimum Detectable Effect
MH	Metropolis-Hastings
MVBART	Multivariate Version of BART
OR	Outcome Regression
PEHE	Precision in Estimation of Heterogeneous Effect
PTA	Parallel Trends Assumption

RIC	Regularization-induced Confounding
RMSE	Root Mean Squared Error
SDID	Synthetic Difference-in-Differences
SUTVA	Stable Unit Treatment Value Assumption
TISCA	Test-Informed Simulation Count Algorithm
TWFE	Two-Way Fixed Effects
wsBCF	warm-start Bayesian Causal Forest
XBART	Accelerated Bayesian Additive Regression Tree

LIST OF SYMBOLS

Y — Outcome Variable

$\mathbb{X}$ — Vector of Covariates

D — Treatment Indicator Variable

$\pi(\mathbf{X})$ — Propensity Score

$\mathcal{N}(0,1)$ — Standard Gaussian Distribution

$\mathbf{1}(\cdot)$ — Indicator Functions

$E\left[\cdot\right]$ — Expected Value

ε — Error Term

$\forall$ — for all

$exp\{\dot{\}}$ — Exponetial Function

$|\dot{|}$ — Absolute Value

$\boldsymbol{y}$ — Observed Vector of Values of the Outcome Variable

$\boldsymbol{x}$ — Observed Matrix of Values of the Covariates

$\boldsymbol{\varepsilon}$ — Observed Vector of Values of the Error Term

CONTENTS

CHAPTER

1

INTRODUCTION

The estimation of treatment effects, particularly understanding how these effects vary across different individuals or subgroups—known as Heterogeneous Treatment Effects (HTEs)—stands as a cornerstone of quantitative research in a multitude of fields including statistics (CURTH; SCHAAR, 2021; HAHN; MURRAY; CARVALHO, 2020; CALLAWAY; SANT'ANNA, 2021; YAO *et al.*, 2021), econometrics (VARIAN, 2016; HOOVER, 1990; CAUSAL..., 2024), public policy (IMBENS, 2024; GANGL, 2010; GRIMMER, 2014), and the health sciences (ROTHMAN; GREENLAND, 2005; OHLSSON; KENDLER, 2020; VANDENBROUCKE; BROADBENT; PEARCE, 2016). A nuanced comprehension of HTEs is paramount as it enables the formulation of personalized interventions, the design of more effective and targeted policies, and a more profound insight into underlying causal mechanisms (CINTRON *et al.*, 2022; REHILL; BIDDLE, 2023; HITSCH; MISRA; ZHANG, 2024).

However, the path to achieving robust and reliable estimates of HTEs is laden with challenges (HAHN; MURRAY; CARVALHO, 2020; CURTH; SCHAAR, 2021; CURTH *et al.*, 2021). Researchers must navigate the complexities of selecting or developing flexible models capable of capturing intricate, non-linear relationships, while simultaneously mitigating risks associated with model misspecification (CHERNOZHUKOV *et al.*, 2024). In the context of observational studies, confounding bias presents a persistent hurdle (CHERNOZHUKOV *et al.*, 2024). Furthermore, the sophisticated estimation techniques often required can impose a significant computational burden (CHERNOZHUKOV *et al.*, 2024). Beyond the models themselves, the very practices used to evaluate and compare new HTE estimators warrant careful scrutiny to ensure the validity, reproducibility, and efficiency of research findings in this domain (CURTH *et al.*, 2021; SOUTO; NETO, 2024b).

This Master's thesis endeavors to partially address these multifaceted challenges and meaningfully contribute to the existing Causal Inference academic literature. The

research encapsulated herein offers a collection of methodological innovations and critical evaluations aimed at augmenting the statistical toolkit available for estimating HTEs. The overarching goal is to foster more rigorous practices in model development, enhance the soundness of evaluation methodologies, and promote the effective application of these techniques in empirical research. The contributions of this monograph are presented across three core research chapters. Incidentally, these papers assume a certain familiarity with nonparametric Bayesian causal models. For readers without this familiarity, Appendix A presents a detailed introduction to these models.

The thesis commences with Chapter 2, *The Importance of Ablation Studies for Complex Nonparametric Causal Models*. This chapter emphasizes the critical need for thorough evaluation of individual components within complex nonparametric causal models. Focusing on the widely-used Bayesian Causal Forest (BCF) model, the research employs a series of ablation studies to investigate the practical utility of including the estimated propensity score, $\hat{\boldsymbol{\pi}}(\mathbf{X})$, a component intended to mitigate regularization-induced confounding. The findings reveal that omitting this component often does not detract from the model's performance in estimating treatment effects or quantifying uncertainty, and notably reduces computational time by approximately 21%. This chapter advocates for the routine adoption of ablation studies in the development and assessment of treatment effect estimators to ensure models are both parsimonious and efficient, preventing unnecessary complexity. The implications of this research suggest that practitioners can potentially simplify established complex models without sacrificing inferential quality, leading to more computationally efficient and interpretable analyses. This work contributes to the literature by formally championing the use of ablation studies, a standard practice in other machine learning domains, within the specific context of causal effect estimation, thereby fostering a culture of critical model component evaluation and refinement.

Building upon the principle of rigorous methodological assessment, Chapter 3, *Beyond Arbitrary Replications: A Principled Approach to Simulation Design in Causal Inference*, confronts a common pitfall in the comparative evaluation of HTE estimators: the often arbitrary determination of the number of simulation replications and the lack of formal statistical comparisons. This chapter introduces the Test-Informed Simulation Count Algorithm (TISCA), a novel framework that integrates Welch's t-tests with statistical power analysis. TISCA iteratively conducts simulations until a predefined statistical power is achieved for detecting a user-specified minimum detectable effect size, thereby yielding a statistically justified number of replications and facilitating rigorous model comparisons. A bibliometric study presented within the chapter confirms the considerable heterogeneity in current simulation practices, and a case study revisiting McJames *et al.* (2024) demonstrates TISCA's potential to optimize computational resources while ensuring the statistical validity of simulation outcomes. The work in this chapter seeks to promote more robust, efficient, and sustainable simulation practices within causal infer-

ence research and beyond.

The primary implication of TISCA is a shift towards more credible, resource-conscious, and statistically sound model evaluation practices. Its adoption can significantly enhance the reliability and comparability of findings across simulation studies. This chapter contributes to the literature not only by identifying and empirically substantiating a widespread methodological concern but also by providing a concrete, actionable algorithmic solution that promotes statistical rigor and supports sustainable computational research in the evaluation of causal estimators.

The third primary research contribution is detailed in Chapter 4, *Forests for Differences: Robust Causal Inference Beyond Parametric DiD*. This chapter introduces a methodological innovation for HTE estimation within the Difference-in-Differences (DiD) framework, a quasi-experimental design extensively used for treatment evaluation. The proposed Difference-in-Differences Bayesian Causal Forest (DiD-BCF) model is a non-parametric approach designed to address contemporary challenges in DiD settings, such as staggered treatment adoption and the presence of heterogeneous treatment effects. A key feature of DiD-BCF is its theoretically grounded reparameterization based on the Parallel Trends Assumption (PTA), aimed at enhancing estimation accuracy and stability, particularly in complex panel data scenarios. Extensive simulation studies demonstrate the DiD-BCF model's superior performance over established benchmarks, especially when confronted with non-linearity, selection biases, and effect heterogeneity. Furthermore, an application to U.S. minimum wage policy illustrates the model's capacity to uncover significant conditional treatment effect heterogeneity related to county population, insights often obscured by traditional, more restrictive DiD methods. This chapter offers DiD-BCF as a versatile and robust tool for achieving more nuanced causal inference in modern DiD applications.

The implications of developing DiD-BCF are far-reaching, offering practitioners a powerful tool capable of navigating the complexities of modern DiD analyses, including staggered interventions and covariate-dependent effects, within a unified framework. This allows for the extraction of more granular, policy-relevant insights from observational panel data. This research contributes a novel non-parametric Bayesian estimator to the DiD literature, extending the capabilities of causal forests to this important quasi-experimental setting. Furthermore, the introduction of the PTA-based reparameterization presents a distinct methodological advancement for enhancing the stability and accuracy of non-parametric estimators in DiD contexts.

Following these research chapters, Chapter 5 offers a consolidated conclusion, synthesizing the principal findings and contributions of the monograph and outlining potential directions for future research. The thesis is further supplemented by appendices. Appendix B provides additional results related to the ablation studies discussed in Chap-

ter 2.

In sum, this monograph aims to make substantive contributions to the field of heterogeneous treatment effect estimation. By presenting critical evaluations of existing complex models, proposing novel algorithms for the rigorous design of simulation studies, and developing innovative non-parametric estimators for challenging causal inference settings, this work strives to equip researchers with improved methodologies for generating more robust, reliable, and insightful causal evidence.

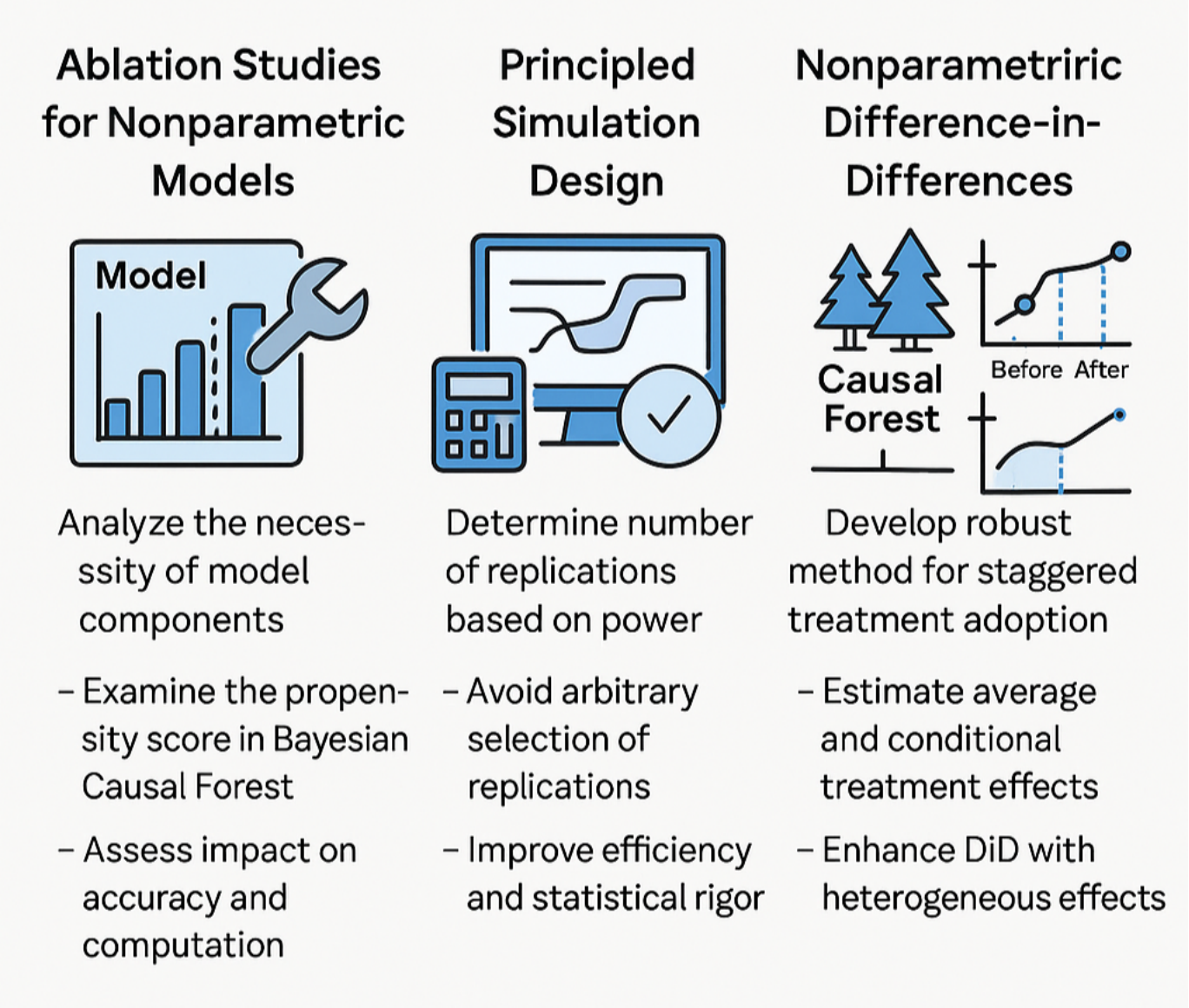

Figure 1 – Visual Abstract

CHAPTER

2

THE IMPORTANCE OF ABLATION STUDIES FOR COMPLEX NONPARAMETRIC CAUSAL MODELS

Abstract

Ablation studies are essential for understanding the contribution of individual components within complex models, yet their application in nonparametric treatment effect estimation remains limited. This paper emphasizes the importance of ablation studies by examining the Bayesian Causal Forest (BCF) model, particularly the inclusion of the estimated propensity score $\hat{\boldsymbol{\pi}}(\mathbf{X})$ intended to mitigate regularization-induced confounding (RIC). Through a partial ablation study utilizing a total of nine synthetic, we demonstrate that excluding $\hat{\boldsymbol{\pi}}(\mathbf{X})$ does not diminish the model's performance in estimating average and conditional average treatment effects or in uncertainty quantification. Moreover, omitting $\hat{\boldsymbol{\pi}}(\mathbf{X})$ reduces computational time by approximately 21%. These findings could suggest that the BCF model's inherent flexibility suffices in adjusting for confounding without explicitly incorporating the propensity score. The study advocates for the routine use of ablation studies in treatment effect estimation to ensure model components are essential and to prevent unnecessary complexity.

Key Words: Conditional Average Treatment Effect, Average Treatment Effect, Bayesian Additive Regression Trees, Bayesian Causal Forest Model, Continuous Treatment Effect.

2.1 Introduction

In recent years, the field of machine learning has witnessed a substantial proliferation of complex models, particularly within the realm of neural networks (SOUTO; MORADI, 2024a; OLIVARES *et al.*, 2023; BIANCHINI; SCARSELLI, 2014; ADHIKARI *et al.*, 2019; HASANPOUR *et al.*, 2016). A critical aspect of model development and validation that has gained prominence is the use of *ablation studies*. Ablation studies involve systematically removing or altering components of a model to assess their individual contributions to overall performance (CAO *et al.*, 2020; ZEILER; FERGUS, 2014; GOMEZ *et al.*, 2017; WU *et al.*, 2022b). This methodology enables researchers to dissect models, understand the significance of each component, and optimize architectures for enhanced efficiency and effectiveness.

The importance of ablation studies is well-recognized in the machine learning literature (SHEIKHOLESLAMI *et al.*, 2021; DU, 2020; MEYES *et al.*, 2019), especially in the development of neural networks for tasks such as image recognition and natural language processing. For example, He *et al.* (2016) utilized ablation studies to demonstrate the impact of residual connections on deep convolutional neural network performance, leading to the development of the ResNet architecture. Similarly, Szegedy *et al.* (2015) employed ablation experiments to refine the Inception network by analyzing the contributions of various convolutional filter sizes to the network's ability to capture different levels of image features.

Despite their widespread application in general machine learning research, the practice of conducting ablation studies is noticeably lacking in the literature on treatment effect estimation, particularly with complex nonparametric models used for estimating the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE). Many studies in this domain introduce sophisticated models without thoroughly investigating the necessity and influence of their individual components (HASSANPOUR; GREINER, 2020; SHALIT; JOHANSSON; SONTAG, 2017; WAGER; ATHEY, 2018; THAL *et al.*, 2024; HILL, 2011; HAHN; MURRAY; CARVALHO, 2020; WU *et al.*, 2022a). This omission can be problematic, as it may lead to the adoption of unnecessarily complex models, potential overfitting, and inefficiencies that could be mitigated through careful analysis.

In this paper, we highlight the importance of ablation studies in the context of treatment effect estimation by examining the Bayesian Causal Forest (BCF) model proposed by Hahn, Murray and Carvalho (2020). The BCF model is a prominent example, already used in various applied studies (YEAGER *et al.*, 2019; YEAGER *et al.*, 2022; BAIL *et al.*, 2019), that incorporates the estimated propensity score into the baseline function to mitigate Regularization-induced Confounding (RIC). The inclusion of the propensity score is presented as a critical component for accurately estimating treatment

effects in observational studies (HAHN; MURRAY; CARVALHO, 2020).

We challenge this assertion by conducting an ablation study focused on the use of the estimated propensity score within the BCF model. Utilizing synthetic data generated based on the designs proposed by Ballinari and Bearth (2024), we assess the necessity of the propensity score in varying data-generating processes (DGPs) and sample sizes. Our findings indicate that the inclusion of the estimated propensity score does not enhance the performance of the BCF model in any of the tested scenarios. Furthermore, incorporating the propensity score results in a significant decrease in computational efficiency, with the model running approximately 21% slower. These results underscore the critical role that ablation studies can play in the development and evaluation of models for treatment effect estimation. By systematically examining the contributions of individual model components, researchers can avoid unnecessary complexity, improve computational efficiency, and enhance the interpretability of their models.

The remainder of this paper is structured as follows: In Subchapter 2.2, we explore the role of the estimated propensity score in the BCF model and discuss the implications of the RIC problem. Subchapter 2.3 provides a brief overview of the synthetic data used for our analysis and outlines the evaluation methods employed. In Subchapter 2.4 and 2.5, we present the results of our main ablation study and conditional ablation studies respectively, and offer a detailed analysis of the findings. Finally, Subchapter 2.6 concludes the paper and highlights the importance of incorporating ablation studies in future research on treatment effect estimation.

2.2 BCF Model Architecture

The BCF model addresses a critical issue known as RIC, which arises in the context of high-dimensional covariate spaces and complex outcome models (HAHN *et al.*, 2018). Incidentally, RIC is accentuated when the outcome variable is largely determined by the covariates rather than the treatment (HAHN *et al.*, 2018) and when there is considerable target selection (i.e., the propensity score is monotone in the prognostic function) (HAHN; MURRAY; CARVALHO, 2020).

RIC refers to the bias that can occur when the regularization inherent in flexible modeling approaches inadvertently induces a correlation between the estimated treatment effect and the propensity score, potentially leading to incorrect causal inferences (HAHN *et al.*, 2018). The essence of RIC lies in the fact that, in observational studies with strong confounding and weak signal-to-noise ratios, the model's attempt to regularize towards simpler functions can cause the estimated treatment effect to absorb residual variation that is actually due to confounding, especially in the presence of considerable target selection. As a result, the posterior distribution of the treatment effect may be substantially

influenced by the prior over the outcome model, rather than being driven by the data (HAHN; MURRAY; CARVALHO, 2020).

To mitigate the effects of RIC, the BCF model incorporates the estimated propensity score $\hat{\pi}(\mathbf{x_i})$ directly into the outcome model. By doing so, it adjusts for the treatment assignment mechanism in a manner that is integrated into the Bayesian framework of the model. The modified outcome model is specified as:

$$\mathbb{E}[Y_i \mid \mathbf{X_i} = \mathbf{x_i}, D_i = D_i] = \mu(\mathbf{x_i}, \hat{\pi}(\mathbf{x_i})) + \tau(\mathbf{x_i})D_i, \tag{2.1}$$

where Y_i is the outcome variable, $\mathbf{X_i}$ represents the covariates, D_i is the treatment indicator, $\mu(\mathbf{x_i}, \hat{\pi}(\mathbf{x_i}))$ is the prognostic function capturing the baseline outcome adjusted for the propensity score, and $\tau(\mathbf{x_i})$ is the heterogeneous treatment effect function. The inclusion of $\hat{\pi}(\mathbf{X_i})$, which is usually estimated with the standard BART model for binary outcomes, in the prognostic term $\mu(x_i, \hat{\pi}(\mathbf{X_i}))$ serves to account for the relationship between the covariates and the treatment assignment, effectively controlling for confounding in the estimation of the baseline outcome. This adjustment is particularly important in scenarios where treatment assignment is related to the potential outcomes, such as in cases of *targeted selection*, where individuals are assigned to treatment based on predicted outcomes under control (HAHN; MURRAY; CARVALHO, 2020).

For the estimation of $\mu(\mathbf{X_i}, \hat{\pi}(\mathbf{X_i}))$, and $\tau(\mathbf{X_i})$, the BCF model employs separate BART priors with different regularization parameters to reflect their distinct roles in the model. The prognostic function $\mu(\mathbf{X_i}, \hat{\pi}(\mathbf{X_i}))$ is modeled using a sum of regression trees with a relatively diffuse prior to capture potentially complex baseline relationships. Specifically, the prior settings for μ use 200 trees, a shrinkage parameter $\eta = 0.95$, and a tree depth parameter $\beta = 2$, following the defaults recommended by Chipman, George and McCulloch (2010). Additionally, a half-Cauchy prior is placed on the scale parameter of the leaf nodes, with a prior median set to twice the marginal standard deviation of Y (HAHN; MURRAY; CARVALHO, 2020). In contrast, the treatment effect function $\tau(\mathbf{X_i})$ is modeled with a stronger regularization to reflect the assumption that treatment effect heterogeneity may be more modest in magnitude and complexity. The prior for $\tau(\mathbf{X_i})$ utilizes 50 trees, a shrinkage parameter $\eta = 0.25$, and a depth parameter $\beta = 3$, imposing a stricter control over the flexibility of the function. A half-normal prior is placed on the scale parameter of the leaf nodes, anchoring the prior median to the marginal standard deviation of Y (HAHN; MURRAY; CARVALHO, 2020).

By accounting for the treatment assignment mechanism within the outcome model, BCF aims to produce more reliable estimates of the ATE and CATE, particularly in observational studies where confounding is a significant concern. Nonetheless, despite these theoretical advantages, the necessity of including $\hat{\pi}(\mathbf{X_i})$ in the BCF model has not been thoroughly examined through ablation studies. Such studies are essential to ascertain whether the inclusion of the propensity score genuinely contributes to improved treatment

effect estimation or if it introduces unnecessary complexity and computational burden. In our investigation, we aim to assess the impact of excluding $\hat{\pi}(\mathbf{X_i})$ from the BCF model by conducting a partial ablation study, challenging the assumption that the inclusion of the estimated propensity score is crucial for superb performance of the BCF model, even in cases of target selection and the outcome variable being largely determined by the covariates.

By systematically analyzing the model's performance, we provide insights into the practical significance of incorporating $\hat{\pi}(\mathbf{X})$ and contribute to the discourse on model simplification and efficiency in treatment effect estimation, making the improved BCF model 21% faster than the proposed model by Hahn, Murray and Carvalho (2020).

2.3 Methodology

2.3.1 Considered Synthetic Data

To evaluate the impact of including the estimated propensity score $\hat{\pi}(\mathbf{X})$ in the BCF model, we conducted a partial ablation study using synthetic data, which includes three models, namely the original BCF model (BCF ($\hat{\pi}(\mathbf{X})$)), the BCF model with the true propensity score BCF ($\pi(\mathbf{X})$), and the BCF model without the use of the estimated propensity score (BCF (no $\hat{\pi}(\mathbf{X})$)) (i.e., using $\hat{\pi}(\mathbf{X}) = 0.5 \; \forall \; X \in \mathbb{X}$). The data were generated inspired in the designs proposed by Ballinari and Bearth (2024), yet modified to vary the extent to which the outcome variable is determined by the covariates relatively to the treatment and the propensity score is a monotone function of the prognostic function as the greater both are the greater the bias of the estimated treatment effect is due to RIC.

The different data-generating processes (DGPs) are given as:

$$
\begin{aligned}
\mathbf{X_i} &\sim \text{Uniform}(0,1)^5, \\
D_i \mid \mathbf{X_i} &\sim \text{Bernoulli}(\pi(\mathbf{X_i})), \\
\varepsilon_i &\sim \mathcal{N}(0,1), \\
Y_i &= b(\mathbf{X_i}) + (D_i - 0.5) \cdot \left(\frac{X_{i,1} + X_{i,2}}{2\alpha} \right) + \varepsilon_i,
\end{aligned}
$$

where $\mathbf{X_i}$ is a 5-dimensional vector of covariates for individual i, D_i is the binary treatment indicator, Y_i is the observed outcome , and α is the hyperparameter that controls the extent to which the outcome variable is determined by the covariates relatively to the treatment. Additionally, the functions $b(\mathbf{X_i})$ and $\pi(\mathbf{X_i})$ define the baseline main effect (a.k.a the prognostic function) and the propensity score function, respectively. Three different values of α are considered in this paper, namely 1, 2, and 4, leading to the average amount of times that the baseline main effect being greater than the treatment effect being approximately 3, 7 and 13 times. Remember that the negative impact of RIC

on treatment effect estimation is greater as the conditional expectation of Y is largely determined by X rather than D (HAHN; MURRAY; CARVALHO, 2020).

The considered baseline and propensity scores are:

$$
\begin{aligned}
b(\mathbf{X_i}) &= \sin(\pi X_{i,1} X_{i,2}) + 2(X_{i,3} - 0.5)^2 + X_{i,4} + 0.5 X_{i,5}, \\
\pi_{\text{extreme}}(\mathbf{X_i}) &= 0.05 + 0.9 \cdot \text{BetaCDF}_{2,4}(\text{sigmoid}(b(\mathbf{X_i}))) \\
\pi_{\text{moderate}}(\mathbf{X_i}) &= 0.05 + 0.75 \cdot \text{BetaCDF}_{2,4}(\text{sigmoid}(b(\mathbf{X_i}))) + 0.15 \cdot \text{BetaCDF}_{2,4}(\min(X_{i,1}, X_{i,2})), \\
\pi_{\text{slight}}(\mathbf{X_i}) &= 0.05 + 0.9 \cdot \text{BetaCDF}_{2,4}(\min(X_{i,1}, X_{i,2})).
\end{aligned}
$$

and the considered configurations of the different DGPs are: DGP1 $(b(\mathbf{X}); \pi_{\text{extreme}}(\mathbf{X}))$ having an extreme target selection, DGP2 $(b(\mathbf{X}); \pi_{\text{moderate}}(\mathbf{X}))$ having a moderate target selection, and DGP3 $(b(\mathbf{X}); \pi_{\text{slight}}(\mathbf{X}))$ having a slight target selection. For each DGP, we considered the sample size of $n = 250$ as this was the sample size used in Hahn, Murray and Carvalho (2020) when presenting an example of how the use of BCF $(\hat{\pi}(\mathbf{X}))$ instead of BCF (no $\hat{\pi}(\mathbf{X})$) mitigates RIC. Anew, the negative impact of RIC on treatment effect estimation is greater as the extent to which the propensity score is a monotone function of the prognostic function increases (i.e., the target selection is more pronounced) (HAHN; MURRAY; CARVALHO, 2020). A visual illustration of the extent to which the different $\pi(\mathbf{X})$ configurations are a monotone function of the prognostic function can be found in Figure 2.

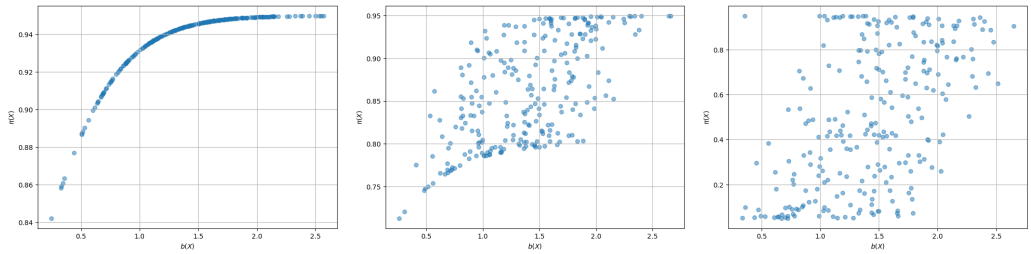

Figure 2 – Scatter plot of $\pi(\mathbf{X})$ by $b(\mathbf{X})$ for different configurations of $\pi(X)$

For each DGP, we considered the sample size of $n = 250$ as this was the sample size used in Hahn, Murray and Carvalho (2020) when presenting an example of how the use of BCF $(\hat{\pi}(\mathbf{X}))$ instead of BCF (no $\hat{\pi}(\mathbf{X})$) mitigates RIC.

Nevertheless, to ensure the sparsity of this paper, we focus on the most challenging scenarios in the main text. Specifically, we present the results for $alpha = 4$ of all DGPs. These DGPs were selected because are the contexts in which the use of BCF $(\hat{\pi}(\mathbf{X}))$ instead of BCF (no $\hat{\pi}(\mathbf{X})$) would be the most beneficial according to Hahn, Murray and Carvalho (2020). Yet, the results for $\alpha = 1$ and $\alpha = 2$ are provided in the Appendix B. Importantly, the findings across all α values lead to the same conclusions; thus, independently of the choice of α value for the main text, the findings of this ablation study would remain the same.

To further scrutinize the role of $\hat{\pi}(\mathbf{X})$ and the robustness of our initial findings, we extend our analysis through a series of conditional ablation studies. While the term "conditional ablation" is prominent in biological sciences for achieving spatial and temporal specificity in gene or cell inactivation (CORNEJO *et al.*, 2024; ZHANG *et al.*, 2012; ELLMAN *et al.*, 2020), its conceptual parallel in machine learning—assessing component impact given specific states or configurations of other features or hyperparameters (KöNIG *et al.*, 2021; EWALD *et al.*, 2024)—is highly relevant here. Our conditional ablation studies, therefore, involve repeating the primary ablation analysis while systematically varying key aspects of the DGP or the BCF model's hyperparameters. The aim is to identify specific conditions, if any, under which the inclusion of $\hat{\pi}(\mathbf{X})$ demonstrates a clear advantage, particularly in scenarios where RIC might be more pronounced or where the model's intrinsic capacity to adjust for confounding might be challenged. For these conditional analyses, and to maintain focus and ensure sparsity of this paper, all conditional ablation studies were conducted exclusively using DGP1 with $\alpha = 4$. This specific setting was chosen because, according to the arguments presented by Hahn, Murray and Carvalho (2020), it represents an extreme scenario where RIC is expected to be strongest. Consequently, if the benefits of including $\hat{\pi}(\mathbf{X})$ are to be observed, they should be most apparent under these conditions.

We explored three main conditional variations. Firstly, we investigated the impact of sample size on the relative performance of BCF with and without $\hat{\pi}(\mathbf{X})$ by setting $n \in [100, 500, 1000, 10000]$. This explores if the propensity score's utility becomes more evident in settings with little or great availability of data. For instance, with very small sample sizes (e.g., $n = 100$), the model estimation is inherently more challenging, and the explicit guidance from $\hat{\pi}(\mathbf{X})$ could theoretically offer stability or improved confounding adjustment. Conversely, at very large sample sizes (e.g., $n = 10000$), where asymptotic properties become more relevant and nuisance parameters like $\hat{\pi}(\mathbf{X})$ can be estimated with high precision, any subtle benefits or detriments of its inclusion might become more clearly distinguishable.

Secondly, we examined the effect of altering the complexity of the prognostic component of the BCF model by varying the number of trees used for the $\mu(\mathbf{X})$ function (mu_trees), considering values from $[5, 25, 100, 400]$. By significantly reducing the number of mu_trees (e.g., to 5 trees), we simplify the baseline function $\mu(\mathbf{X})$, potentially curtailing its inherent flexibility to capture complex relationships and adjust for confounding. In such a scenario with a less flexible prognostic component, it could be posited that the incorporation of the estimated propensity score may be more beneficial, as the model might need to rely more heavily on the explicit information provided by $\hat{\pi}(\mathbf{X})$ to mitigate RIC.

Thirdly, we explored the impact of covariate space dimensionality. Our initial

DGP1 used for the conditional studies, featured a prognostic function $b(\mathbf{X})$ and a treatment effect $\tau(\mathbf{X})$ both dependent on a fixed set of five covariates, $X_i \sim \text{Uniform}(0,1)^5$. For this conditional ablation study, we modified DGP1 to vary the total number of influential covariates p, considering $p \in [10, 50, 100, 200]$. The covariates $\mathbf{X}$ were drawn from $\text{Uniform}(0,1)^p$. To increase p while attempting to maintain a comparable overall magnitude for the prognostic function $b(\mathbf{X})$ and a consistent definition for the treatment effect $\tau(\mathbf{X})$, we adopted the following generalization: the p covariates were divided into $p/5$ blocks. The original $b(\mathbf{X_i}) = \sin(\pi X_{i,1} X_{i,2}) + 2(X_{i,3} - 0.5)^2 + X_{i,4} + 0.5X_{i,5}$ structure was calculated for each block using its respective five covariates, and the final $b(\mathbf{X_i})$ was the sum of these block-wise contributions, averaged by the number of blocks. The propensity score function remained $\pi_{\text{extreme}}(\mathbf{X_i}) = 0.05 + 0.9 \cdot \text{BetaCDF}_{2,4}(\text{sigmoid}(b(\mathbf{X_i})))$, now using the generalized $b(X_i)$. Crucially, the treatment effect function $\tau(X_i)$ was kept dependent only on the first two covariates, $\tau(\mathbf{X_i}) = (X_{i,1} + X_{i,2})/(2\alpha)$, with $\alpha = 4$, irrespective of the total number of covariates p. The outcome Y_i was then generated as $Y_i = b(\mathbf{X_i}) + (D_i - 0.5) \cdot \tau(\mathbf{X_i}) + \varepsilon_i$.

One could argue that the original 5-covariate setting is not a regime where regularization has a large effect. By increasing the number of covariates p upon which $b(\mathbf{X})$ depends, especially relative to a fixed sample size (e.g., $n = 250$ for the conditional studies focused on DGP1 with $\alpha = 4$), the model would presumably need to regularize more to avoid overfitting the prognostic function. This increased regularization pressure could, in turn, exacerbate RIC. In such higher-dimensional settings, it can be hypothesized that the propensity function $\hat{\pi}(\mathbf{X})$ would become more important as a targeted mechanism to counteract confounding induced by the stronger regularization, which might otherwise aggressively shrink the estimated effects of true confounders relevant to the prognostic score.

Lastly, all simulations, including these conditional ablation studies, were repeated 100 times to account for variability due to random sampling and to provide robust estimates of the evaluation metrics.

2.3.2 Evaluation Methods

The performance of the BCF model, both with and without the inclusion of the estimated propensity score, was evaluated using metrics consistent with those used in Souto and Neto (2024a), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coverage (Cover), and Length (Len).

All these metrics were computed for both ATE and CATE while only RMSE and MAE are used for $\hat{\pi}(\mathbf{X})$ to show that the used propensity scores are indeed different for each model (as it could theoretically be the case that BCF ($\hat{\pi}(\mathbf{X})$) $\approx$ BCF ($\pi(\mathbf{X})$) or BCF (no $\hat{\pi}(\mathbf{X})$) if the estimated propensity score was remarkably accurate or innacurate

respectively). Besides, the evaluation metrics results of the 100 different simulations of each DGP are presented in the form of their means and standard deviations.

Additionally, to determine whether differences in the evaluation metrics between the models with and without the estimated propensity score were statistically significant, we employed Welch's t-test (WELCH, 1947) as suggested by Souto and Neto (2024b). Different from the ordinary t-test, Welch's t-test (WELCH, 1947) does not assume equal variance, which is usually not the case for the compared model as one can see in the results of this paper, while achieving similar power even when variances are actually equal, especially for groups with balanced and considerable sample sizes like the 100 simulations we compare (SOUTO; NETO, 2024b; RUXTON, 2006; FAGERLAND, 2012). The only assumption Welch's t-test (WELCH, 1947) does is the normality assumption, but as discussed in Souto and Neto (2024b), for comparing the aforementioned model evaluation metrics for 100 simulations, this assumption is reasonable due to the Central Limit Theorem. We do not test this assumption as performing such a "pre-testing" has been widely criticized and not recommended in the statistical literature (ZIMMERMAN, 2004; ROCHON; GONDAN; KIESER, 2012; SCHUCANY; NG, 2006).

2.3.3 Programming Language and Libraries

For all ablation studies, the programming language R and the R library stochtree were used for the employement of the BCF model and its variations.

2.4 Results and Analysis: Main Ablation Studies

Before going to the results of the specifics DGPs, it is worth mentioning that all analyses were performed using the same computational environment to maintain consistency. The computational time for fitting each model was recorded to assess the impact of including the estimated propensity score on computational efficiency. The inclusion of $\hat{\boldsymbol{\pi}}(\mathbf{X})$ increased the computational time by approximately 21% (with BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) taking 27.27 minutes per DGP simulation study while BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) taking 33.33 minutes), demonstrating a trade-off between model complexity and computational resources. As we will see below, the additional computational costs do not lead to a better performance at CATE and ATE estimation nor uncertainty quantification.

Table 1 presents the results of the evaluation metrics for DGP1 ($b(\mathbf{X}); \boldsymbol{\pi}_{\text{extreme}}(\mathbf{X})$), DGP2 ($b(\mathbf{X}); \boldsymbol{\pi}_{\text{moderate}}(\mathbf{X})$), and DGP3 ($b(\mathbf{X}); \boldsymbol{\pi}_{\text{slight}}(\mathbf{X})$). It can be seen that all models perform relatively similar considering all evaluation measures and DGP1 and DGP2, while the metrics RMSE_{pi} and MAE_{pi} demonstrate that the models are indeed different (i.e., the used propensity scores are different per model). For DGP3, the performance of BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) is lower than the other models for the pointwise evaluation metrics, albeit

statistical tests are needed to determine whether the difference in performance between BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) and BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) are statistically significantly. Besides, the results are a bit contradicting to the affirmations of Hahn, Murray and Carvalho (2020) as we would expect that the negative impact of RIC would be more pronunced in DGP1 and DGP2, but here it appears to be present only in DGP3.

Moving to the statistical tests, Figure 3 presents the p-values for the statistical tests of the evaluation measures for the comparison of BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) and BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$). It can be seen that the used propensity scores for BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) and BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) are indeed statistically different, and even without using any estimation of the propensity score, the BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) model achieves a performance that is statistically significantly similar to the BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) model for all evaluation metrics. The only exception here would be the mean value of Len_{ATE} for DG1 and DGP2 and MAPE_{CATE} for DGP3, yet given the similarity in coverage for these models, such statistically significant difference is not important or indicative of a superiority of one model over the other in uncertainty quantification.

Hence, it can be concluded that the estimation of the propensity score and its use is not only not necessary for the considerable performance of the BCF model in treatment effect estimation (even in cases of target selection and the outcome variable being largely determined by the covariates), but also leads to additional computational costs, increasing the amount of time needed to fit the model by roughly 21%. Consequently, the BCF model ought to simply not estimate the propensity score function when estimating CATE and ATE (in any case not seperately as advocated by Hahn, Murray and Carvalho (2020) as perhaps the model can already estimate it directly thanks to its flexibility, albeit more research would be needed to explore this hypothesis), unless the researchers and/practitioners are interested in the estimation of the propensity score function.

Table 1 – Mean and Standard Deviation for Different Metrics

DGP	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
DGP1	RMSE_{CATE}	0.157 ± 0.119	0.155 ± 0.112	$\mathbf{0.155 \pm 0.106}$
	MAE_{CATE}	0.146 ± 0.121	0.146 ± 0.114	$\mathbf{0.145 \pm 0.106}$
	MAPE_{CATE}	1.570 ± 1.370	1.540 ± 1.300	$\mathbf{1.540 \pm 1.240}$
	Cover_{CATE}	$\mathbf{0.992 \pm 0.041}$	0.996 ± 0.024	0.998 ± 0.015
	Len_{CATE}	1.160 ± 0.170	$\mathbf{1.100 \pm 0.153}$	1.190 ± 0.176
	RMSE_{ATE}	0.139 ± 0.127	0.139 ± 0.120	$\mathbf{0.138 \pm 0.112}$
	MAE_{ATE}	0.139 ± 0.127	0.139 ± 0.120	$\mathbf{0.138 \pm 0.112}$
	MAPE_{ATE}	1.100 ± 1.000	1.110 ± 0.950	$\mathbf{1.100 \pm 0.898}$
	Cover_{ATE}	$\mathbf{0.960 \pm 0.197}$	0.980 ± 0.141	0.990 ± 0.100
	Len_{ATE}	0.869 ± 0.132	$\mathbf{0.863 \pm 0.130}$	0.914 ± 0.140
	RMSE_{π}	0.438 ± 0.00104	$\mathbf{0 \pm 0}$	0.045 ± 0.008
	MAE_{π}	0.438 ± 0.00110	$\mathbf{0 \pm 0}$	0.037 ± 0.006
DGP2	RMSE_{CATE}	0.130 ± 0.083	$\mathbf{0.129 \pm 0.081}$	0.134 ± 0.090
	MAE_{CATE}	0.119 ± 0.083	$\mathbf{0.117 \pm 0.082}$	0.122 ± 0.090
	MAPE_{CATE}	1.300 ± 0.955	$\mathbf{1.250 \pm 0.907}$	1.290 ± 1.010
	Cover_{CATE}	0.998 ± 0.008	0.997 ± 0.017	$\mathbf{0.997 \pm 0.016}$
	Len_{CATE}	0.930 ± 0.142	$\mathbf{0.901 \pm 0.132}$	0.959 ± 0.143
	RMSE_{ATE}	0.110 ± 0.088	$\mathbf{0.108 \pm 0.089}$	0.113 ± 0.096
	MAE_{ATE}	0.110 ± 0.088	$\mathbf{0.108 \pm 0.089}$	0.113 ± 0.096
	MAPE_{ATE}	0.877 ± 0.697	$\mathbf{0.865 \pm 0.705}$	0.904 ± 0.757
	Cover_{ATE}	$\mathbf{0.950 \pm 0.219}$	0.970 ± 0.171	0.970 ± 0.171
	Len_{ATE}	0.638 ± 0.077	$\mathbf{0.628 \pm 0.077}$	0.676 ± 0.088
	RMSE_{π}	0.366 ± 0.00355	$\mathbf{0 \pm 0}$	0.079 ± 0.012
	MAE_{π}	0.362 ± 0.00357	$\mathbf{0 \pm 0}$	0.062 ± 0.010
DGP3	RMSE_{CATE}	0.147 ± 0.099	0.131 ± 0.087	$\mathbf{0.129 \pm 0.083}$
	MAE_{CATE}	0.132 ± 0.097	0.117 ± 0.084	$\mathbf{0.115 \pm 0.080}$
	MAPE_{CATE}	1.650 ± 1.330	1.340 ± 1.180	$\mathbf{1.280 \pm 1.050}$
	Cover_{CATE}	$\mathbf{0.983 \pm 0.065}$	0.984 ± 0.077	0.989 ± 0.056
	Len_{CATE}	0.851 ± 0.147	$\mathbf{0.818 \pm 0.151}$	0.835 ± 0.146
	RMSE_{ATE}	0.117 ± 0.105	0.100 ± 0.094	$\mathbf{0.100 \pm 0.087}$
	MAE_{ATE}	0.117 ± 0.105	0.100 ± 0.094	$\mathbf{0.100 \pm 0.087}$
	MAPE_{ATE}	0.940 ± 0.852	0.800 ± 0.758	$\mathbf{0.799 \pm 0.705}$
	Cover_{ATE}	0.910 ± 0.288	$\mathbf{0.940 \pm 0.239}$	0.970 ± 0.171
	Len_{ATE}	0.555 ± 0.059	$\mathbf{0.542 \pm 0.062}$	0.564 ± 0.067
	RMSE_{π}	0.312 ± 0.00659	$\mathbf{0 \pm 0}$	0.119 ± 0.015
	MAE_{π}	0.280 ± 0.00822	$\mathbf{0 \pm 0}$	0.093 ± 0.011

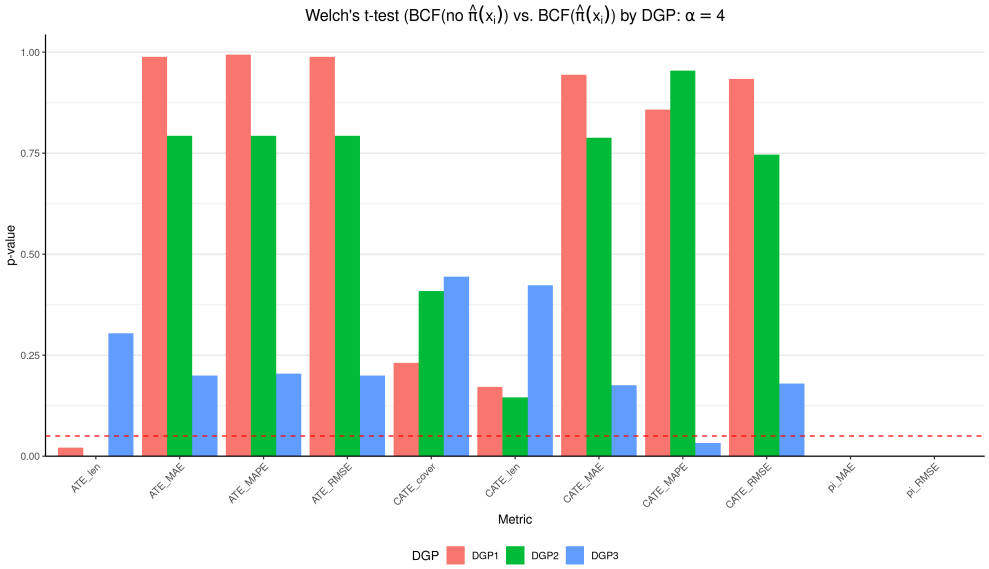

Figure 3 – Welch t-test p-values for BCF (no $\hat{\pi}(\mathbf{X})$) vs BCF ($\hat{\pi}(\mathbf{X})$) for Main Ablation Study and $\alpha = 4$

2.5 Results and Analysis: Conditional Ablation Studies

2.5.1 Varying Number of Samples

The first set of conditional ablation studies explores the impact of sample size ($n \in [100, 500, 1000, 10000]$) on model performance. The detailed results are presented in Table 2, with corresponding Welch's t-test p-values comparing BCF (no $\hat{\pi}(\mathbf{X})$) and BCF ($\hat{\pi}(\mathbf{X})$) shown in Figure 4. Across all tested sample sizes, BCF (no $\hat{\pi}(\mathbf{X})$) performed comparably to BCF ($\hat{\pi}(\mathbf{X})$) and BCF ($\pi(\mathbf{X})$), in terms of all evaluation metrics. The p-values for these estimation metrics are generally high, indicating no statistically significant difference in point estimation and uncertainty accuracy. While some statistically significant differences were observed for interval lengths (Len_{CATE} and Len_{ATE}) at for $n = 100$ and only Len_{ATE} for $n = 500\&10000$, this is more likely due to randomness than a superiority of BCF (no $\hat{\pi}(\mathbf{X})$). These results suggest that varying the sample size, even to extremes, does not reveal a clear practical or statistically significant advantage for including the estimated propensity score for treatment effect estimation accuracy.

Table 2 – Mean and Standard Deviation for Different Metrics with Varying Sample Sizes (DGP1)

N	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
100	RMSE_{CATE}	0.133 ± 0.081	$\mathbf{0.129 \pm 0.077}$	0.133 ± 0.074
	MAE_{CATE}	0.124 ± 0.083	$\mathbf{0.120 \pm 0.079}$	0.123 ± 0.076
	MAPE_{CATE}	1.321 ± 0.948	$\mathbf{1.256 \pm 0.875}$	1.292 ± 0.871
	Cover_{CATE}	$\mathbf{0.999 \pm 0.008}$	0.999 ± 0.012	1.000 ± 0.004

Continued on next page

Table 2 – continued from previous page

N	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
	Len_{CATE}	0.914 ± 0.179	$\mathbf{0.864 \pm 0.171}$	0.970 ± 0.165
	RMSE_{ATE}	0.119 ± 0.088	$\mathbf{0.115 \pm 0.083}$	0.118 ± 0.081
	MAE_{ATE}	0.119 ± 0.088	$\mathbf{0.115 \pm 0.083}$	0.118 ± 0.081
	MAPE_{ATE}	0.948 ± 0.696	$\mathbf{0.917 \pm 0.665}$	0.941 ± 0.650
	Cover_{ATE}	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	Len_{ATE}	0.738 ± 0.156	$\mathbf{0.718 \pm 0.156}$	0.798 ± 0.151
	RMSE_{π}	0.438 ± 0.002	$\mathbf{0 \pm 0}$	0.054 ± 0.012
	MAE_{π}	0.438 ± 0.002	$\mathbf{0 \pm 0}$	0.044 ± 0.009
500	RMSE_{CATE}	0.096 ± 0.046	$\mathbf{0.095 \pm 0.043}$	0.104 ± 0.041
	MAE_{CATE}	0.084 ± 0.045	$\mathbf{0.084 \pm 0.043}$	0.092 ± 0.041
	MAPE_{CATE}	0.876 ± 0.510	$\mathbf{0.848 \pm 0.479}$	0.895 ± 0.412
	Cover_{CATE}	0.976 ± 0.091	$\mathbf{0.972 \pm 0.097}$	0.974 ± 0.094
	Len_{CATE}	0.525 ± 0.098	$\mathbf{0.489 \pm 0.082}$	0.540 ± 0.096
	RMSE_{ATE}	$\mathbf{0.072 \pm 0.052}$	0.075 ± 0.050	0.083 ± 0.049
	MAE_{ATE}	$\mathbf{0.072 \pm 0.052}$	0.075 ± 0.050	0.083 ± 0.049
	MAPE_{ATE}	$\mathbf{0.578 \pm 0.416}$	0.598 ± 0.397	0.663 ± 0.394
	Cover_{ATE}	$\mathbf{0.950 \pm 0.219}$	0.960 ± 0.197	0.920 ± 0.273
	Len_{ATE}	0.336 ± 0.044	$\mathbf{0.327 \pm 0.041}$	0.353 ± 0.045
	RMSE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.040 ± 0.006
	MAE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.032 ± 0.004
1000	RMSE_{CATE}	0.084 ± 0.030	$\mathbf{0.083 \pm 0.028}$	0.094 ± 0.034
	MAE_{CATE}	0.072 ± 0.030	$\mathbf{0.071 \pm 0.027}$	0.082 ± 0.033
	MAPE_{CATE}	0.724 ± 0.290	$\mathbf{0.712 \pm 0.274}$	0.775 ± 0.314
	Cover_{CATE}	0.969 ± 0.084	0.972 ± 0.074	$\mathbf{0.950 \pm 0.113}$
	Len_{CATE}	0.427 ± 0.083	$\mathbf{0.410 \pm 0.076}$	0.432 ± 0.090
	RMSE_{ATE}	$\mathbf{0.058 \pm 0.038}$	0.061 ± 0.034	0.070 ± 0.042
	MAE_{ATE}	$\mathbf{0.058 \pm 0.038}$	0.061 ± 0.034	0.070 ± 0.042
	MAPE_{ATE}	$\mathbf{0.467 \pm 0.305}$	0.489 ± 0.276	0.557 ± 0.339
	Cover_{ATE}	0.910 ± 0.288	$\mathbf{0.920 \pm 0.273}$	0.850 ± 0.359
	Len_{ATE}	0.248 ± 0.028	$\mathbf{0.246 \pm 0.028}$	0.252 ± 0.029
	RMSE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.035 ± 0.004
	MAE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.028 ± 0.003
10000	RMSE_{CATE}	0.047 ± 0.011	$\mathbf{0.046 \pm 0.011}$	0.048 ± 0.012
	MAE_{CATE}	$\mathbf{0.038 \pm 0.009}$	$\mathbf{0.038 \pm 0.009}$	0.039 ± 0.010
	MAPE_{CATE}	0.454 ± 0.115	0.452 ± 0.115	$\mathbf{0.448 \pm 0.108}$

Continued on next page

Table 2 – continued from previous page

N	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
	Cover_{CATE}	0.968 ± 0.044	0.972 ± 0.042	$\mathbf{0.959 \pm 0.073}$
	Len_{CATE}	0.215 ± 0.034	$\mathbf{0.213 \pm 0.034}$	$\mathbf{0.213 \pm 0.034}$
	RMSE_{ATE}	$\mathbf{0.017 \pm 0.012}$	0.017 ± 0.013	0.019 ± 0.014
	MAE_{ATE}	$\mathbf{0.017 \pm 0.012}$	0.017 ± 0.013	0.019 ± 0.014
	MAPE_{ATE}	$\mathbf{0.136 \pm 0.098}$	0.136 ± 0.101	0.149 ± 0.113
	Cover_{ATE}	$\mathbf{0.960 \pm 0.197}$	0.930 ± 0.256	0.930 ± 0.256
	Len_{ATE}	0.083 ± 0.004	$\mathbf{0.082 \pm 0.003}$	0.084 ± 0.004
	RMSE_{π}	0.438 ± 0.000	$\mathbf{0 \pm 0}$	0.021 ± 0.002
	MAE_{π}	0.438 ± 0.000	$\mathbf{0 \pm 0}$	0.017 ± 0.001

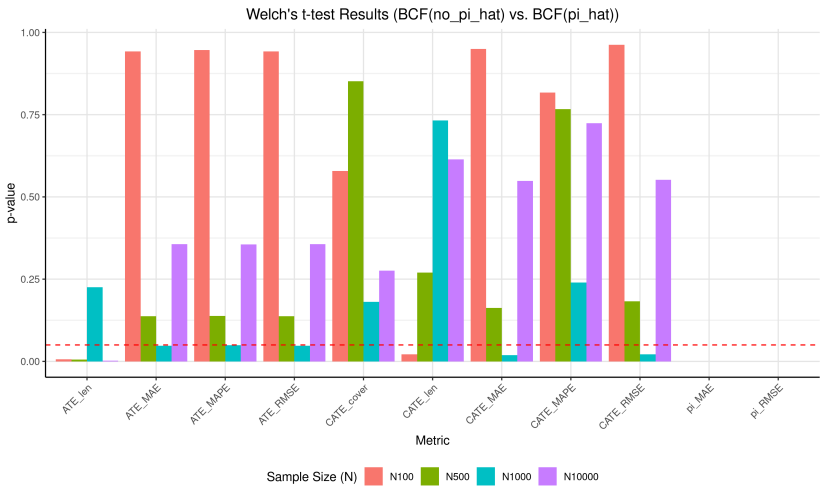

Figure 4 – Welch t-test p-values for BCF (no $\hat{\pi}(\mathbf{X})$) vs BCF ($\hat{\pi}(\mathbf{X})$) for Samples Conditional Ablation Study

2.5.2 Varying Number of μ Trees

Next, we assess the impact of the BCF model's complexity by varying the number of trees in the prognostic component (μ trees $\in [5, 25, 100, 400]$), with results in Table 3 and Figure 5. It was hypothesized that a simpler prognostic model (e.g., 5 μ trees) might benefit more from the explicit inclusion of $\hat{\pi}(\mathbf{X})$. Although this is not the case from the explicit inclusion of $\hat{\pi}(\mathbf{X})$, it is true from the explicit inclusion of $\pi(\mathbf{X})$. Even though BCF ($\hat{\pi}(\mathbf{X})$) only has a small π estimation error of roughly 4% on average, this error is already enough to take any advantage from including $\hat{\pi}(\mathbf{X})$ to the model. Yet, the advantage of BCF ($\pi(\mathbf{X})$) rapidly disappears as we move further from the extreme case of 5 μ trees and go closer to the default value of 200 μ trees. Hence, it could be affirmed that only in the extreme case of 5 μ trees, the inclusion of a nearly perfectly estimated $\hat{\pi}(\mathbf{X})$ would be beneficial.

Moving to the Welch's t-tests results, the p-values largely indicate no statistically significant differences in estimation accuracy between BCF (no $\hat{\pi}(\mathbf{X})$) and BCF ($\hat{\pi}(\mathbf{X})$) across the different numbers of μ trees. Overall, the hypothesis that reducing the number μ trees would make $\hat{\pi}(\mathbf{X})$ more beneficial was not supported and the results of this conditional ablation studies further evidence that the incorporation of $\hat{\pi}(\mathbf{X})$ to the BCF model is unnecessary.

Table 3 – Mean and Standard Deviation for Different Metrics with Varying Number of μ Trees (DGP1)

# μ Trees	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
5	RMSE_{CATE}	0.162 ± 0.083	$\mathbf{0.111 \pm 0.057}$	0.170 ± 0.075
	MAE_{CATE}	0.138 ± 0.072	$\mathbf{0.101 \pm 0.057}$	0.145 ± 0.065
	MAPE_{CATE}	1.474 ± 0.795	$\mathbf{1.012 \pm 0.632}$	1.563 ± 0.750
	Cover_{CATE}	0.959 ± 0.087	0.980 ± 0.082	$\mathbf{0.957 \pm 0.082}$
	Len_{CATE}	0.770 ± 0.158	$\mathbf{0.600 \pm 0.109}$	0.785 ± 0.160
	RMSE_{ATE}	0.101 ± 0.074	$\mathbf{0.094 \pm 0.063}$	0.109 ± 0.072
	MAE_{ATE}	0.101 ± 0.074	$\mathbf{0.094 \pm 0.063}$	0.109 ± 0.072
	MAPE_{ATE}	0.805 ± 0.593	$\mathbf{0.750 \pm 0.506}$	0.874 ± 0.572
	Cover_{ATE}	0.910 ± 0.288	$\mathbf{0.950 \pm 0.219}$	0.880 ± 0.327
	Len_{ATE}	0.461 ± 0.092	$\mathbf{0.452 \pm 0.074}$	0.480 ± 0.097
	RMSE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.045 ± 0.007
	MAE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.037 ± 0.005
25	RMSE_{CATE}	0.115 ± 0.061	$\mathbf{0.107 \pm 0.057}$	0.125 ± 0.070
	MAE_{CATE}	0.104 ± 0.060	$\mathbf{0.097 \pm 0.057}$	0.114 ± 0.070
	MAPE_{CATE}	1.052 ± 0.594	$\mathbf{0.972 \pm 0.593}$	1.111 ± 0.707
	Cover_{CATE}	0.977 ± 0.101	0.980 ± 0.090	$\mathbf{0.970 \pm 0.107}$
	Len_{CATE}	0.671 ± 0.130	$\mathbf{0.619 \pm 0.117}$	0.710 ± 0.134
	RMSE_{ATE}	0.095 ± 0.067	$\mathbf{0.089 \pm 0.065}$	0.105 ± 0.077
	MAE_{ATE}	0.095 ± 0.067	$\mathbf{0.089 \pm 0.065}$	0.105 ± 0.077
	MAPE_{ATE}	0.762 ± 0.531	$\mathbf{0.707 \pm 0.518}$	0.843 ± 0.612
	Cover_{ATE}	$\mathbf{0.940 \pm 0.239}$	0.920 ± 0.273	0.900 ± 0.302
	Len_{ATE}	0.475 ± 0.078	$\mathbf{0.463 \pm 0.075}$	0.516 ± 0.082
	RMSE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.045 ± 0.007
	MAE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.037 ± 0.005
100	RMSE_{CATE}	0.114 ± 0.064	$\mathbf{0.110 \pm 0.057}$	0.124 ± 0.066
	MAE_{CATE}	0.103 ± 0.063	$\mathbf{0.100 \pm 0.057}$	0.113 ± 0.068
	MAPE_{CATE}	1.044 ± 0.620	$\mathbf{1.003 \pm 0.588}$	1.102 ± 0.689

Continued on next page

Table 3 – continued from previous page

# μ Trees	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
	Cover_{CATE}	0.983 ± 0.066	0.978 ± 0.091	$\mathbf{0.976 \pm 0.082}$
	Len_{CATE}	0.668 ± 0.135	$\mathbf{0.629 \pm 0.115}$	0.703 ± 0.129
	RMSE_{ATE}	0.091 ± 0.070	$\mathbf{0.091 \pm 0.065}$	0.106 ± 0.074
	MAE_{ATE}	0.091 ± 0.070	$\mathbf{0.091 \pm 0.065}$	0.106 ± 0.074
	MAPE_{ATE}	0.730 ± 0.556	$\mathbf{0.726 \pm 0.522}$	0.849 ± 0.592
	Cover_{ATE}	$\mathbf{0.950 \pm 0.219}$	0.940 ± 0.239	0.920 ± 0.273
	Len_{ATE}	0.473 ± 0.078	$\mathbf{0.465 \pm 0.071}$	0.520 ± 0.088
	RMSE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.045 ± 0.007
	MAE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.037 ± 0.005
400	RMSE_{CATE}	0.116 ± 0.062	$\mathbf{0.109 \pm 0.057}$	0.130 ± 0.066
	MAE_{CATE}	0.104 ± 0.061	$\mathbf{0.099 \pm 0.057}$	0.116 ± 0.064
	MAPE_{CATE}	1.052 ± 0.624	$\mathbf{0.996 \pm 0.583}$	1.143 ± 0.673
	Cover_{CATE}	0.983 ± 0.066	0.985 ± 0.062	$\mathbf{0.977 \pm 0.078}$
	Len_{CATE}	0.680 ± 0.137	$\mathbf{0.635 \pm 0.122}$	0.731 ± 0.142
	RMSE_{ATE}	0.093 ± 0.068	$\mathbf{0.090 \pm 0.065}$	0.105 ± 0.072
	MAE_{ATE}	0.093 ± 0.068	$\mathbf{0.090 \pm 0.065}$	0.105 ± 0.072
	MAPE_{ATE}	0.741 ± 0.545	$\mathbf{0.716 \pm 0.521}$	0.838 ± 0.572
	Cover_{ATE}	$\mathbf{0.950 \pm 0.219}$	$\mathbf{0.950 \pm 0.219}$	0.910 ± 0.288
	Len_{ATE}	0.472 ± 0.077	$\mathbf{0.464 \pm 0.081}$	0.516 ± 0.078
	RMSE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.045 ± 0.007
	MAE_{π}	0.438 ± 0.001	$\mathbf{0 \pm 0}$	0.037 ± 0.005

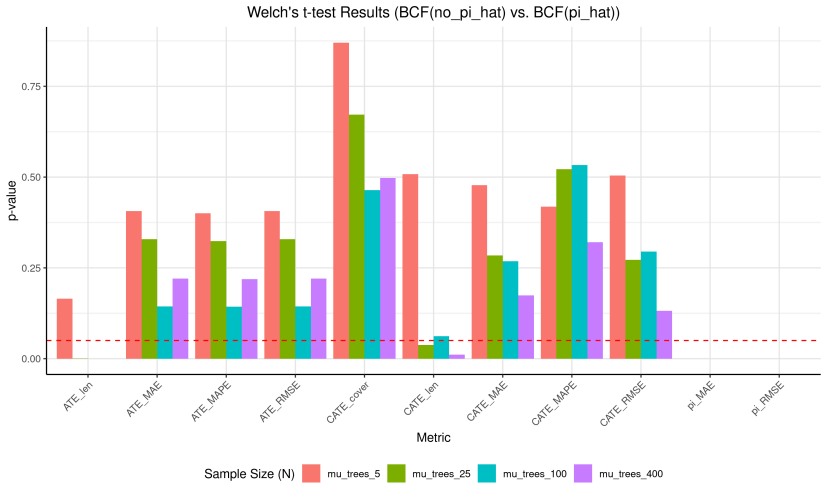

Figure 5 – Welch t-test p-values for BCF (no $\hat{\pi}(\mathbf{X})$) vs BCF ($\hat{\pi}(\mathbf{X})$) for μ Trees Conditional Ablation Study

2.5.3 Varying Number of Covariates

Finally, we investigate whether increasing the dimensionality of the covariate space ($p \in [10, 50, 100, 200]$ influential covariates) would make the inclusion of $\hat{\pi}(\mathbf{X})$ more critical due to increased regularization pressure. Results are presented in Table 4 and Figure 6. One could be quite intrigued with the fact that the treatment effect estimation precision and uncertainty quantification of all BCF models remain relatively equal as p increases. Nonetheless, this has already been shown in the original BCF paper of Hahn, Murray and Carvalho (2020) in the contributed discussion of Prado *et al.* (2020), where they explored the robustness of the BCF model against benchmark models using the same DGPs as Hahn, Murray and Carvalho (2020), but considering the following number of covariates $p \in [5, 50, 100, 500]$. The results of this exploration can be found in Figure 7. While we only varied p in the baseline function (given the original goal of our conditional ablation study), Prado *et al.* (2020) also varied it in the treatment effect function. Thus, their closest setting to our ablation study is the down right quadrant of Figure 7, and similar to our resutls, the treatment effect estimation precision and uncertainty quantification of the BCF model remains relatively equal as p increases. Thus, the robustness of the BCF model as seen in Table 4 converges to past literature and thus was already expected.

Now coming back to the ablation studies core objective: across all tested covariate dimensions, the performance of BCF (no $\hat{\pi}(\mathbf{X})$) and BCF ($\hat{\pi}(\mathbf{X})$) in estimating ATE and CATE remain remarkably similar. The p-values from the Welch's t-tests for CATE and ATE estimation metrics are overwhelmingly high, indicating no statistically significant difference between the two model variants regardless of the number of covariates. While the p-value for Len_{CATE} for $p = 10$ indicate statistically significant difference, since this difference disappears in the other values for p, we can safely assume that this significant p-value was presumably due to randomness. Thus, the findings do not support the hypothesis that $\hat{\pi}(X)$ becomes more important for mitigating RIC in higher-dimensional covariate settings, at least within the tested range. The BCF model's inherent flexibility appears sufficient to handle these scenarios without the explicit inclusion of an estimated propensity score.

Table 4 – Mean and Standard Deviation for Different Metrics with Varying Number of Covariates (DGP1)

p	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
10	RMSE_{CATE}	0.106 ± 0.053	$\mathbf{0.103 \pm 0.051}$	0.106 ± 0.053
	MAE_{CATE}	0.095 ± 0.054	$\mathbf{0.093 \pm 0.053}$	0.095 ± 0.055
	MAPE_{CATE}	0.986 ± 0.585	$\mathbf{0.944 \pm 0.585}$	0.945 ± 0.568
	Cover_{CATE}	0.989 ± 0.046	$\mathbf{0.984 \pm 0.062}$	0.992 ± 0.040

Continued on next page

Table 4 – continued from previous page

p	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
	Len_{CATE}	0.658 ± 0.113	$\mathbf{0.599 \pm 0.086}$	0.681 ± 0.105
	RMSE_{ATE}	0.085 ± 0.062	$\mathbf{0.084 \pm 0.060}$	0.085 ± 0.063
	MAE_{ATE}	0.085 ± 0.062	$\mathbf{0.084 \pm 0.060}$	0.085 ± 0.063
	MAPE_{ATE}	0.677 ± 0.496	$\mathbf{0.672 \pm 0.487}$	0.677 ± 0.507
	Cover_{ATE}	0.990 ± 0.100	$\mathbf{0.980 \pm 0.141}$	$\mathbf{0.980 \pm 0.141}$
	Len_{ATE}	0.473 ± 0.069	$\mathbf{0.459 \pm 0.069}$	0.508 ± 0.071
	RMSE_{π}	0.441 ± 0.001	$\mathbf{0 \pm 0}$	0.047 ± 0.008
	MAE_{π}	0.441 ± 0.001	$\mathbf{0 \pm 0}$	0.039 ± 0.005
50	RMSE_{CATE}	$\mathbf{0.117 \pm 0.061}$	$\mathbf{0.117 \pm 0.061}$	0.117 ± 0.062
	MAE_{CATE}	0.107 ± 0.063	$\mathbf{0.106 \pm 0.063}$	0.107 ± 0.063
	MAPE_{CATE}	1.070 ± 0.803	1.066 ± 0.793	$\mathbf{1.063 \pm 0.805}$
	Cover_{CATE}	0.974 ± 0.103	$\mathbf{0.970 \pm 0.105}$	0.976 ± 0.103
	Len_{CATE}	0.577 ± 0.089	$\mathbf{0.572 \pm 0.087}$	0.584 ± 0.090
	RMSE_{ATE}	0.101 ± 0.069	$\mathbf{0.100 \pm 0.069}$	0.100 ± 0.070
	MAE_{ATE}	0.101 ± 0.069	$\mathbf{0.100 \pm 0.069}$	0.100 ± 0.070
	MAPE_{ATE}	0.803 ± 0.557	$\mathbf{0.797 \pm 0.557}$	0.799 ± 0.560
	Cover_{ATE}	$\mathbf{0.950 \pm 0.219}$	$\mathbf{0.950 \pm 0.219}$	$\mathbf{0.950 \pm 0.219}$
	Len_{ATE}	$\mathbf{0.448 \pm 0.075}$	0.450 ± 0.074	0.458 ± 0.073
	RMSE_{π}	0.444 ± 0.000	$\mathbf{0 \pm 0}$	0.049 ± 0.004
	MAE_{π}	0.444 ± 0.000	$\mathbf{0 \pm 0}$	0.043 ± 0.003
100	RMSE_{CATE}	0.110 ± 0.047	0.110 ± 0.050	$\mathbf{0.109 \pm 0.048}$
	MAE_{CATE}	$\mathbf{0.099 \pm 0.049}$	0.100 ± 0.052	$\mathbf{0.099 \pm 0.049}$
	MAPE_{CATE}	0.937 ± 0.562	0.952 ± 0.584	$\mathbf{0.941 \pm 0.555}$
	Cover_{CATE}	0.980 ± 0.058	$\mathbf{0.975 \pm 0.066}$	0.984 ± 0.046
	Len_{CATE}	0.565 ± 0.093	$\mathbf{0.562 \pm 0.093}$	0.574 ± 0.094
	RMSE_{ATE}	0.093 ± 0.055	0.094 ± 0.058	$\mathbf{0.092 \pm 0.056}$
	MAE_{ATE}	0.093 ± 0.055	0.094 ± 0.058	$\mathbf{0.092 \pm 0.056}$
	MAPE_{ATE}	0.746 ± 0.440	0.749 ± 0.464	$\mathbf{0.740 \pm 0.446}$
	Cover_{ATE}	0.970 ± 0.171	$\mathbf{0.950 \pm 0.219}$	0.980 ± 0.141
	Len_{ATE}	$\mathbf{0.441 \pm 0.087}$	0.443 ± 0.084	0.448 ± 0.087
	RMSE_{π}	0.444 ± 0.000	$\mathbf{0 \pm 0}$	0.051 ± 0.003
	MAE_{π}	0.444 ± 0.000	$\mathbf{0 \pm 0}$	0.045 ± 0.004
200	RMSE_{CATE}	$\mathbf{0.110 \pm 0.048}$	0.111 ± 0.047	0.111 ± 0.047
	MAE_{CATE}	$\mathbf{0.099 \pm 0.049}$	0.100 ± 0.048	0.100 ± 0.048
	MAPE_{CATE}	$\mathbf{1.008 \pm 0.558}$	1.020 ± 0.557	1.012 ± 0.554

Continued on next page

Table 4 – continued from previous page

p	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
	Cover_{CATE}	0.983 ± 0.056	0.983 ± 0.052	$\mathbf{0.980 \pm 0.064}$
	Len_{CATE}	$\mathbf{0.569 \pm 0.078}$	0.572 ± 0.078	0.574 ± 0.075
	RMSE_{ATE}	$\mathbf{0.092 \pm 0.057}$	0.093 ± 0.056	0.093 ± 0.056
	MAE_{ATE}	$\mathbf{0.092 \pm 0.057}$	0.093 ± 0.056	0.093 ± 0.056
	MAPE_{ATE}	$\mathbf{0.732 \pm 0.455}$	0.745 ± 0.444	0.745 ± 0.443
	Cover_{ATE}	$\mathbf{0.950 \pm 0.219}$	0.970 ± 0.171	0.960 ± 0.197
	Len_{ATE}	$\mathbf{0.436 \pm 0.069}$	0.440 ± 0.069	0.443 ± 0.071
	RMSE_{π}	0.444 ± 0.000	$\mathbf{0 \pm 0}$	0.052 ± 0.003
	MAE_{π}	0.444 ± 0.000	$\mathbf{0 \pm 0}$	0.046 ± 0.003

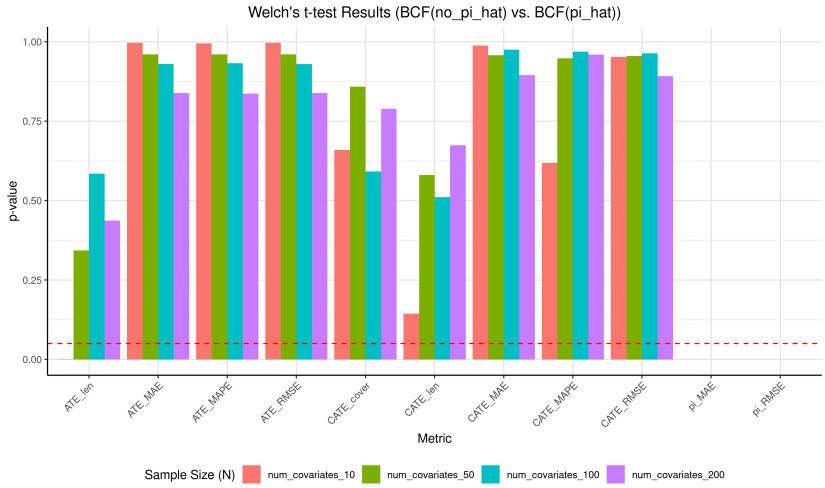

Figure 6 – Welch t-test p-values for BCF (no $\hat{\pi}(\mathbf{X})$) vs BCF ($\hat{\pi}(\mathbf{X})$) for Covariates Conditional Ablation Study

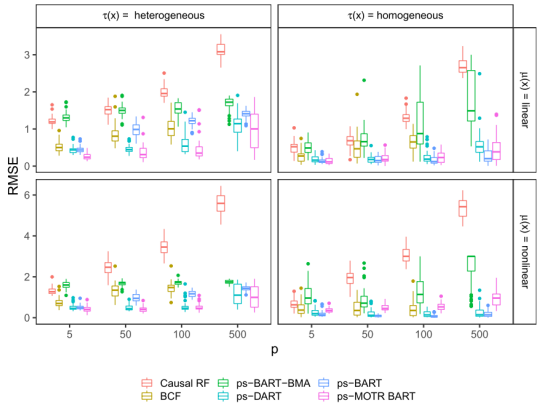

Figure 1: Simulation study results of RMSE for Conditional Average Treatment Effect (CATE).

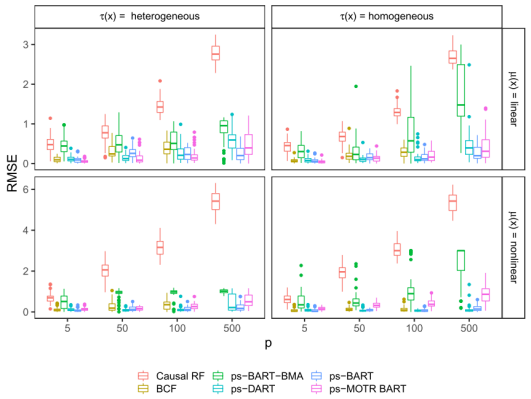

Figure 2: Simulation study results of RMSE for Average Treatment Effect (ATE).

Figure 7 – Results of Prado *et al.* (2020). BCF results for $p = 500$ are not shown as back when this study was done, the BCF code had not been optimized and numerical errors would occurs in such a high dimension (PRADO *et al.*, 2020).

2.6 Conclusion and Recommendations

In this paper, we have highlighted the importance of conducting ablation studies when proposing complex nonparametric models for treatment effect estimation, such as the BCF model. Ablation studies, a staple in the machine learning literature allow for a systematic assessment of individual model components to determine their necessity and impact on performance. Despite their utility, such practices are still virtually non-existent in the treatment effect estimation literature, which can lead to the adoption of unnecessarily complex models without empirical justification. This paper sheds light to this gap in the literature and aims to motivate researchers to adopt the use of ablation studies for their future papers.

Our investigation focused on the role of the estimated propensity score $\hat{\boldsymbol{\pi}}(\mathbf{X})$ within the BCF model, which is included to mitigate RIC Hahn, Murray and Carvalho (2020). We initially conducted a partial ablation study using synthetic data generated from three different DGPs, varying the hyperparameter α (responsible for the extent to which the

outcome variable is determined by the covariates rather than the treatment). The results of this primary study indicated that incorporating the estimated propensity score into the BCF model does not lead to improved performance in estimating the ATE or CATE, nor in uncertainty quantification. This finding was further reinforced by a series of conditional ablation studies where we systematically varied the sample size (from $N = 100$ to $N = 10000$), the number of prognostic (μ) trees in the BCF model (from 5 to 400), and the number of influential covariates in the DGP (from 10 to 200). Even under these diverse conditions, some of which were designed to create scenarios where RIC might be more pronounced (e.g., small sample sizes, smaller number of μ trees, or higher covariate dimensionality), the inclusion of $\hat{\pi}(\mathbf{X})$ did not yield discernible improvements in ATE/-CATE estimation accuracy or the quality of uncertainty quantification. Moreover, it is consistently observed that including $\hat{\pi}(\mathbf{X})$ in the BCF model resulted in a substantial increase in computational time by approximately 21%, while this additional computational cost translating into any additional model performance.

Our comprehensive findings suggest that the BCF model's inherent flexibility likely allows it to capture the necessary relationships between covariates, treatment assignment, and outcomes to adjust for confounding without explicitly including the estimated propensity score. This held true even when the model's complexity was reduced (via fewer μ-trees) or when the data presented higher dimensionality. Alternatively, our findings could indicate that, despite being argued by Hahn, Murray and Carvalho (2020), RIC may not be a significant issue for flexible nonlinear models like BCF in a wide array of scenarios, including those with varying sample sizes and covariate structures. In any case, our results, strengthened by the conditional ablation studies, robustly challenge the default assumption that incorporating $\hat{\pi}(\mathbf{X})$ in the BCF model is necessary for accurate treatment effect estimation.

The implications of this study are twofold. First, it emphasizes the need for ablation studies in the development and evaluation of complex models for treatment effect estimation. By systematically assessing the contributions of individual components, researchers can avoid unnecessary complexity and improve computational efficiency without compromising performance. Second, it calls into question the necessity of including the estimated propensity score in the BCF model for the purpose of treatment effect estimation, suggesting that its inclusion should be reconsidered unless there is a specific interest in estimating the propensity score itself.

In conclusion, our study underscores the critical role of ablation studies in advancing the field of causal inference. By questioning and empirically testing the necessity of model components, researchers can develop more parsimonious, efficient, and interpretable models. We advocate for the adoption of ablation studies as a standard practice in the evaluation of new models for treatment effect estimation, ensuring that each component

contributes meaningfully to the model's performance.

Supplementary Research Material and Code

The code used in this paper can be found in the following GitHub repository: Repository.

Acknowledgments

The authors would like to express heartfelt gratitude to the São Paulo Research Foundation (FAPESP) for their financial support through the Master's scholarship (Processo 2024/06274-0) of the first author of this paper and the associated project (Processo Vinculado 2013/07375-0).

CHAPTER

3

BEYOND ARBITRARY REPLICATIONS: A PRINCIPLED APPROACH TO SIMULATION DESIGN IN CAUSAL INFERENCE

Abstract

Evaluation of novel treatment effect estimators frequently relies on simulation studies lacking formal statistical comparisons and using arbitrary numbers of replications (J). This hinders reproducibility and efficiency. We propose the Test-Informed Simulation Count Algorithm (TISCA) to address these shortcomings. TISCA integrates Welch's t-tests with power analysis, iteratively running simulations until a pre-specified power (e.g., 0.8) is achieved for detecting a user-defined minimum detectable effect size (MDE) at a given significance level (α). This yields a statistically justified simulation count (J) and rigorous model comparisons. Our bibliometric study confirms the heterogeneity of current practices regarding J. A case study revisiting MCJAMES *et al.* (2024) demonstrates TISCA identifies sufficient simulations ($J = 500$ vs. original $J = 1000$), saving computational resources while providing statistically sound evidence. TISCA promotes rigorous, efficient, and sustainable simulation practices in causal inference and beyond.

Key Words: Treatment Effect Estimation, Simulation Studies, Statistical Power, Model Evaluation, Computational Efficiency

3.1 Introduction

In the field of causal inference, estimating the effect of a treatment or intervention is of paramount importance across a variety of disciplines, including economics (BACH; CHERNOZHUKOV; SPINDLER, 2018; VARIAN, 2016; ABADIE; DIAMOND;

HAINMUELLER, 2010; CARD, 1999), epidemiology (ROTHMAN; GREENLAND, 2005; OHLSSON; KENDLER, 2020; VANDENBROUCKE; BROADBENT; PEARCE, 2016), and social sciences (YEAGER *et al.*, 2019; YEAGER *et al.*, 2022; BAIL *et al.*, 2019). Two fundamental quantities of interest in this context are the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE). These metrics provide critical insights into the impact of a treatment on a population, both at an aggregate level and within specific subgroups defined by covariates.

Common approaches to estimating both CATE and ATE include tree-based methods such as Causal Regression Forests (WAGER; ATHEY, 2018), Bayesian Additive Regression Trees (BART) (CHIPMAN; GEORGE; MCCULLOCH, 2010), Bayesian Causal Forest (BCF) (HAHN; MURRAY; CARVALHO, 2020) and other machine learning techniques, such as the neural network models TNet (CURTH; SCHAAR, 2021), TARNet (SHALIT; JOHANSSON; SONTAG, 2017), DragonNet (SHI; BLEI; VEITCH, 2019), DR-CFR (HASSANPOUR; GREINER, 2020), and SNet (CURTH; SCHAAR, 2021). These nonparametric models can capture complex interactions between covariates and even estimate the treatment assignment mechanism/distribution. Almost every month, novel and state-of-the-art models are proposed in the literature, advancing the field of treatment effect estimation. Nonetheless, the current practices in models evaluation and comparisons are far from perfect and could make use of more statistical rigours.

Due to the inherent challenges in causal inference, particularly the unobservability of counterfactual outcomes, researchers resort to using synthetic and semi-synthetic data for model evaluation (CURTH *et al.*, 2021). These data allow for controlled experiments where the true treatment effects are known, thus enabling the assessment of model performance in estimating the ATE and CATE (CURTH *et al.*, 2021). Synthetic data are fully simulated datasets where both the covariates $\mathbf{X}$ and the potential outcomes $Y(1)$ and $Y(0)$ are generated according to pre-specified distributions and relationships. The advantage of using synthetic data lies in the precise control it offers over the data generating process (DGP), allowing researchers to test models under various controlled scenarios (CURTH *et al.*, 2021). Formally, let the DGP be defined as follows:

$$Y_i(\mathbf{X_i}, D_i) = f(\mathbf{X_i}) + \tau(\mathbf{X_i}) \cdot D_i + \varepsilon_i \tag{3.1}$$

where $f(\mathbf{X_i})$ is the baseline outcome model, D_i is the treatment indicator, $\tau(\mathbf{X_i})$ is the true CATE, and ε_i is the error term. By specifying $f(\mathbf{X})$, $\tau(\mathbf{X})$, and the distribution of ε, researchers can generate the potential outcomes $Y(1)$ and $Y(0)$ for any given covariate vector $\mathbf{X}$.

Semi-synthetic data, on the other hand, combine real-world covariates $\mathbf{X}$ with synthetic potential outcomes generated under a specified DGP. This approach retains the

realistic covariate distribution of actual datasets while allowing for controlled experimentation with the outcome-generating process (CURTH *et al.*, 2021). The use of semi-synthetic data is particularly appealing when real-world covariates exhibit complex dependencies that are difficult to capture through fully synthetic simulations.

Nevertheless, it is worth mentioning that both the use of synthetic and semi-synthetic data when evaluation novel models can be problematic if not done meticulously and considering the DGPs biases and limitations (CURTH *et al.*, 2021). In their paper, CURTH *et al.* (2021) address these problems found in many studies in the literature of models for treatment effect estimation and propose concrete actions that researchers ought to undertake to prevent such issues. Similar to the paper of CURTH *et al.* (2021), this paper puts light on problems found in the current analysis practices of model evaluation for treatment effect estimation and proposes solutions to avoid such issues. But before stating these problems, we must first understand the current analysis practices of model evaluation for treatment effect estimation.

The evaluation of model performance in the context of CATE and ATE estimation typically involves metrics such as the Root Mean Squared Error (RMSE), Coverage, and Confidence Interval Length (CIL) (HAHN; MURRAY; CARVALHO, 2020; HILL, 2011). For frequentist models, RMSE is the predominant metric used to assess the accuracy of the estimated treatment effects (CURTH; SCHAAR, 2021; HILL, 2011). RMSE for CATE is defined as:

$$\text{RMSE}_{CATE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2} \tag{3.2}$$

where $\hat{\tau}_i$ is the estimated treatment effect for unit i, and τ_i is the true treatment effect. It is worth mentioning that RMSE_{CATE} is more commonly referred to as Precision in Estimation of Heterogeneous Effect (PEHE) in the literature. The RMSE for ATE, on the other hand, is given as:

$$\text{RMSE}_{ATE} = \sqrt{\frac{1}{J}\sum_{i=j}^{J}(\hat{\tau}_j - \tau_j)^2} \tag{3.3}$$

where $\hat{\tau}_j$ is the estimated average treatment effect for simulation/dataset j, and τ_j is the true average treatment effect.

For Uncertainty quantification models, especially Bayesian models, additional metrics such as Coverage and CIL are commonly used (HAHN; MURRAY; CARVALHO, 2020; HILL, 2011). For CATE, coverage refers to the proportion of times the true treatment effect τ_i falls within the estimated $100(1-\alpha)\%$ credible interval:

$$\text{Coverage}_{CATE} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\tau_i \in \left[\hat{\tau}_i^{\text{lower}}, \hat{\tau}_i^{\text{upper}}\right]\right) \tag{3.4}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and $\left[\hat{\tau}_i^{\text{lower}}, \hat{\tau}_i^{\text{upper}}\right]$ is the credible interval for τ_i. For ATE, coverage refers to the proportion of times the true average treatment effect τ_j for simulation/dataset j falls within the estimated $100(1-\alpha)\%$ credible interval:

$$\text{Coverage}_{ATE} = \frac{1}{J}\sum_{i=j}^{J} \mathbf{1}\left(\tau_j \in \left[\hat{\tau}_j^{\text{lower}}, \hat{\tau}_j^{\text{upper}}\right]\right) \tag{3.5}$$

CIL, on the other hand, measures the width of these credible intervals for ATE and for CATE, the average width:

$$\text{CIL}_{CATE} = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\tau}_i^{\text{upper}} - \hat{\tau}_i^{\text{lower}}\right) \tag{3.6}$$

$$\text{CIL}_{ATE} = \left(\hat{\tau}_j^{\text{upper}} - \hat{\tau}_j^{\text{lower}}\right) \tag{3.7}$$

A shorter CIL indicates a more precise estimate, but this must be balanced against coverage to ensure that the intervals are not too narrow.

While these metrics provide quantitative assessments of performance on a *single* simulation run or dataset, the critical step involves aggregating results across multiple Monte Carlo simulations (indexed by $j = 1, \ldots, J$ in the definitions above) to compare the overall performance of different models, particularly when evaluating a newly proposed estimator against existing benchmarks. It is precisely within this aggregation and comparison process that two significant methodological issues arise in the current literature on treatment effect estimation models.

First, there is a prevalent lack of formal statistical testing when comparing the performance of a novel proposed model against established benchmarks using these aggregated metrics. Often, conclusions about model superiority are drawn based on observing seemingly lower average RMSE or better average coverage across the J simulations, without statistically assessing whether these observed differences are likely due to genuine performance advantages or simply random variation inherent in the Monte Carlo process. For instance, a new model might show a slightly lower average PEHE than a benchmark across J simulations, but without a statistical test (e.g., a t-test comparing the distributions of PEHE values obtained from the two models across the simulations), it remains uncertain whether this difference is statistically significant or could have occurred by chance (SOUTO; MORADI, 2024b; HANSEN; LUNDE ASGER AMD NASON, 2011; HOLLANDER; WOLFE; CHICKEN, 2015; WITT; SONG; LOUVIERIS, 2003).

Second, and closely related to the first issue, is the often arbitrary selection of the number of simulation replications, J. As our bibliometric analysis in Subchapter 3.2 will demonstrate, the choice of J varies considerably across studies, frequently ranging from as low as 5 simulations to several thousand, often based on convention, available computational resources, or an 'educated guess' rather than a principled, statistically motivated approach. This heterogeneity makes cross-study comparisons challenging and raises concerns about the reliability and efficiency of the findings (SOUTO; MORADI, 2024b; HANSEN; LUNDE ASGER AMD NASON, 2011; HOLLANDER; WOLFE; CHICKEN, 2015; WITT; SONG; LOUVIERIS, 2003). An insufficient number of simulations (J) may lead to underpowered comparisons, failing to detect genuine, meaningful differences between models (a Type II error), even if a statistical test were applied (HOLLANDER; WOLFE; CHICKEN, 2015; PREL *et al.*, 2010; GREENLAND *et al.*, 2016; HEDGES; PIGOTT, 2001). Conversely, performing an unnecessarily large number of simulations wastes valuable computational resources and time, hindering research productivity and increasing the environmental footprint of computational research.

These methodological shortcomings are particularly concerning in the current era, characterized by a rapid acceleration in research output, especially thanks to artificial intelligence (AI) and large language model (LLM) applications (GOLDKUHLE *et al.*, 2024; VALE *et al.*, 2024; WYNTER, 2024; YU *et al.*, 2024; PENG *et al.*, 2022). The pressure to publish quickly, coupled with the ease of generating seemingly novel model variations, can lead to a proliferation of studies. Without rigorous evaluation standards, including formal statistical testing and justified simulation designs, we risk saturating the literature with papers claiming state-of-the-art performance based on statistically insignificant or underpowered findings (GOLDKUHLE *et al.*, 2024; VALE *et al.*, 2024; WYNTER, 2024; YU *et al.*, 2024; PENG *et al.*, 2022). This not only hinders true scientific progress by creating noise and propagating potentially false conclusions, but it also erodes trust in the research community and misdirects efforts towards models that may not offer genuine improvements. Emphasizing methodological rigor, ensuring that claims of superiority are statistically sound, and promoting efficient use of computational resources are crucial steps to maintain research quality and integrity amidst this rapid expansion.

This paper addresses these two fundamental shortcomings in the evaluation methodology for treatment effect estimators. We propose a novel algorithm, named **T**est-**I**nformed **S**imulation **C**ount **A**lgorithm (TISCA), designed to integrate formal statistical hypothesis testing directly into the Monte Carlo simulation framework while simultaneously determining the necessary number of simulations required to achieve adequate statistical power. Our approach is grounded in the principles of power analysis, utilizing the Welch's t-test to compare model performance metrics (like RMSE) between a proposed model and its competitors.

Specifically, our algorithm requires the researcher to pre-specify three key parameters standard in hypothesis testing:

1. The desired statistical power $(1-\beta)$, typically set at 0.80, representing the probability of detecting a true difference when it exists.

2. The significance level (α), usually set at 0.05, controlling the probability of a Type I error (falsely claiming a difference).

3. The Minimum Detectable Effect size (MDE), representing the smallest difference in the chosen performance metric (e.g., difference in mean RMSE between the proposed model and the best benchmark) that the researcher deems practically meaningful and wishes to detect. This MDE can be informed by preliminary runs (e.g., 50 simulations) to gauge typical performance differences.

Given these inputs, TISCA iteratively performs simulations, calculating the chosen performance metric for the models under comparison at each step. After each batch of simulations, it conducts a Welch's t-test and estimates the current statistical power based on the observed data and the specified MDE. The simulations continue until the estimated power reaches or exceeds the pre-specified target (e.g., 0.80).

Consequently, researchers employing our algorithm will not only obtain a statistically rigorous assessment of their model's relative performance against benchmarks but will also arrive at a justified and efficient number of simulation replications (J) tailored to their specific research question and desired level of certainty. This data-driven approach replaces arbitrary choices with a principled determination of simulation effort, thereby enhancing the credibility, reproducibility, and efficiency of model evaluation studies in treatment effect estimation.

The remainder of this paper is structured as follows: Subchapter 3.2 presents a bibliometric study quantifying the heterogeneity in the number of simulations used in recent literature, empirically motivating the need for standardization. Subchapter 3.3 details the proposed algorithm, discusses its theoretical underpinnings based on the Welch's t-test and power analysis, explores its strengths and limitations, and introduces a companion website with an interactive fine-tuned LLM tool designed to assist researchers in implementing the proposed algorithm (see <https://tisca-llm-app.streamlit.app/>). Subchapter 3.4 provides a practical demonstration by revisiting the simulation study of MCJAMES *et al.* (2024), illustrating how our algorithm could have determined the necessary number of simulations, saving significant computational resources while ensuring statistical validity. Finally, Subchapter 3.5 concludes with a summary of our contributions and discusses potential avenues for future research.

3.2 Bibliometric Study

3.2.1 Study Design and Data Collection

To empirically investigate the current practices regarding the number of simulation replications used in evaluating treatment effect estimation models, we conducted a targeted bibliometric analysis analysing a sample of 100 papers. The primary objective was to quantify the heterogeneity in the reported number of simulations (J) across relevant studies.

Search Strategy

We utilized Google Scholar as the primary search database due to its broad coverage of academic literature, including peer-reviewed articles and preprints. The search was conducted using a combination of keywords relevant to the field of treatment effect estimation, specifically:

- "treatment effect estimation"
- "heterogeneous treatment effect estimation"
- "inverse probability weighting treatment effect estimation"
- "nonparametric treatment effect estimation"
- "matching treatment effect estimation"

The search focused on identifying papers published in recent years to capture contemporary practices (with 57% of the papers being published in the last 5 years and 88% in the last 10 years).

Selection Criteria

From the initial pool of search results, studies were selected for inclusion based on the following criteria:

1. The study must propose either a novel model for treatment effect estimation, a significant modification of an existing model aimed at improvement, or a novel methodological approach for estimating treatment effects (ATE or CATE).
2. The study must include simulation experiments (using synthetic or semi-synthetic data) as part of its model evaluation or comparison methodology.
3. The study must report the number of simulation replications (J) performed for these experiments.

Studies that solely applied existing methods without proposing novelty, focused purely on theoretical aspects without simulations, or did not clearly report the number of simulations were excluded (with the later criterium culminating in only one exclusion, something positive from the considered literature).

Handling of Preprints

Our sample included studies published in peer-reviewed journals as well as preprints hosted on arXiv. Preprints constitute a significant channel for disseminating cutting-edge research in machine learning and related fields, including treatment effect estimation. To ensure relevance and potential impact, arXiv preprints were included only if they met at least one of the following conditions:

- The preprint had accrued at least one citation from a published paper (according to Google Scholar), suggesting some level of peer recognition or influence.
- The preprint had been posted within the last year prior to our data collection cutoff date, ensuring recency and allowing the inclusion of relevant papers that are presumably under review currently.

Approximately 10% of our final sample consisted of such arXiv preprints. To assess the potential impact of including these non-peer-reviewed works, we performed a sensitivity analysis. We compared the distribution and summary statistics of the number of simulations (J) for the full sample with the results obtained when excluding the arXiv preprints. This comparison revealed that the overall findings and conclusions regarding the heterogeneity in simulation counts remained unchanged, confirming the robustness of our results irrespective of the inclusion of these preprints, which for some could be a controversial decision.

Final Sample

Following the application of these search and selection criteria, a final sample of $N = 100$ studies was compiled for analysis. These studies represent a cross-section of recent work in the field that utilizes simulation studies for evaluating novel contributions to treatment effect estimation. The distribution and characteristics of the reported number of simulations (J) across these studies are presented in the following subsection.

3.2.2 Results

The results of the bibliometric analysis, conducted as described in Subchapter 3.2.1, are summarized in Figures 8, 9, and 10.

Figure 8 illustrates the distribution of the 100 sampled studies across various publishers and publication outlets. The wide array of sources, ranging from top-tier statistics and machine learning journals (e.g., PMLR, Statistica Sinica) and conferences (e.g., ICLR, AAAI) to broader scientific venues (e.g., PNAS) and preprint servers (arXiv, representing 11.11%), demonstrates that our sample is not confined to a narrow niche within the literature.

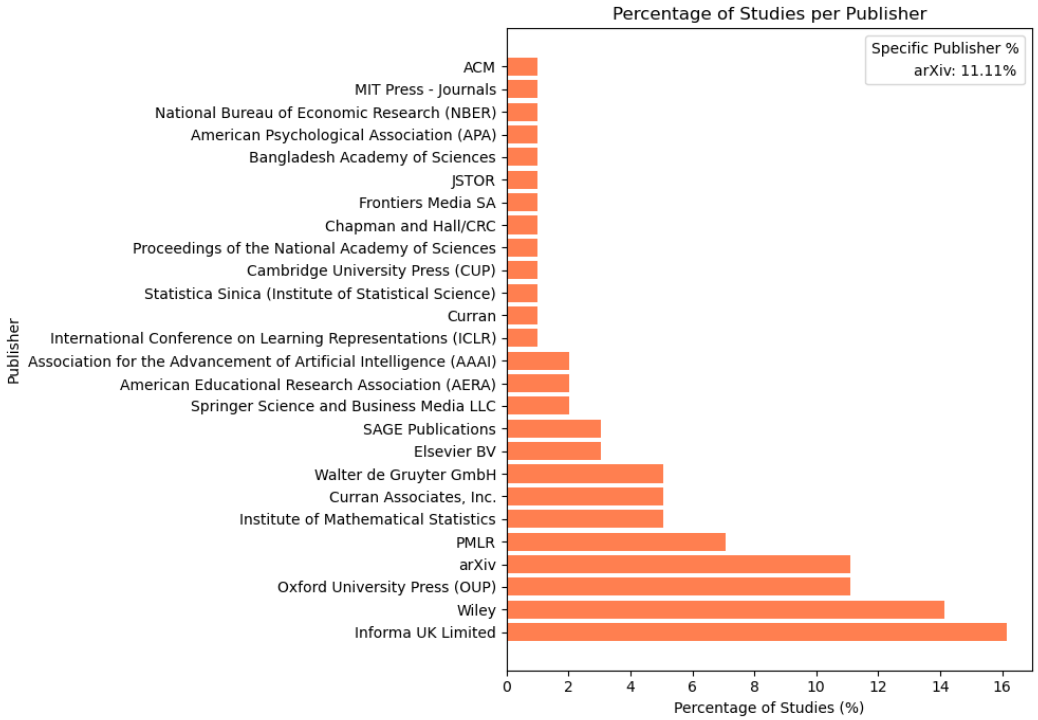

Figure 8 – Percentage of Studies per Publisher

Similarly, Figure 9 shows the distribution of studies by publication year. While the sample spans over more than two decades, there is a clear concentration in recent years, with 56.57% of studies published between 2021 and 2025 and 87.88% published between 2016 and 2025. This temporal distribution ensures that our analysis reflects contemporary practices. Together, the diversity of publishers and the temporal spread underscore the representativeness and relevance of our sample, lending credibility to the findings regarding simulation practices.

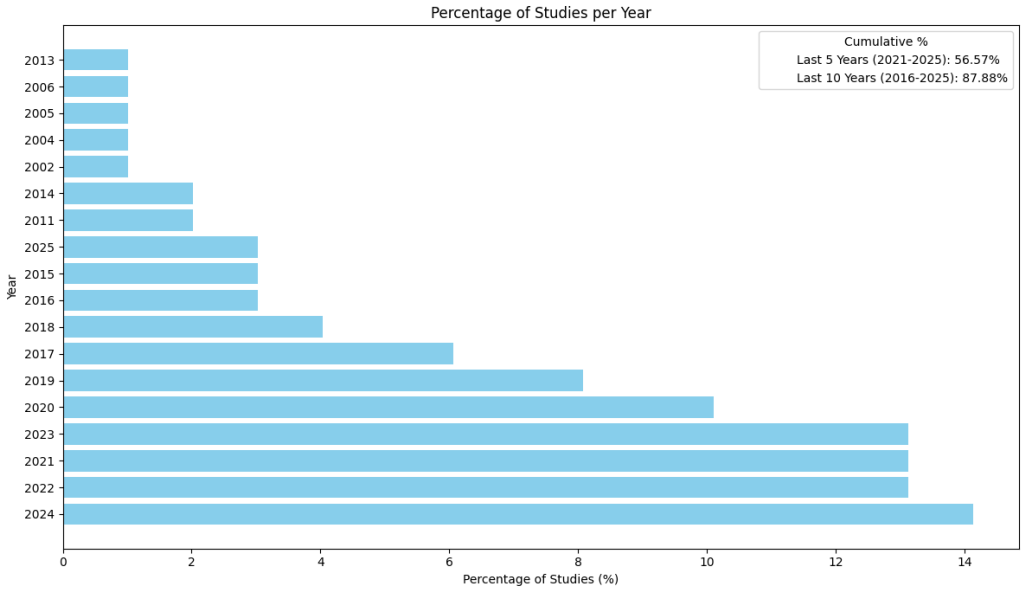

Figure 9 – Percentage of Studies per Year

The core finding of our bibliometric study is presented in Figure 10, which displays the distribution of the number of simulation replications (J) reported in the sampled studies. The top panel shows the distribution for the full sample ($N = 100$), while the bottom panel shows the distribution after excluding arXiv preprints ($N = 89$). Both plots reveal striking heterogeneity in the choice of J. The number of simulations ranges dramatically, from as few as 10 or even 5 replications in some studies to as many as 100,000 in others.

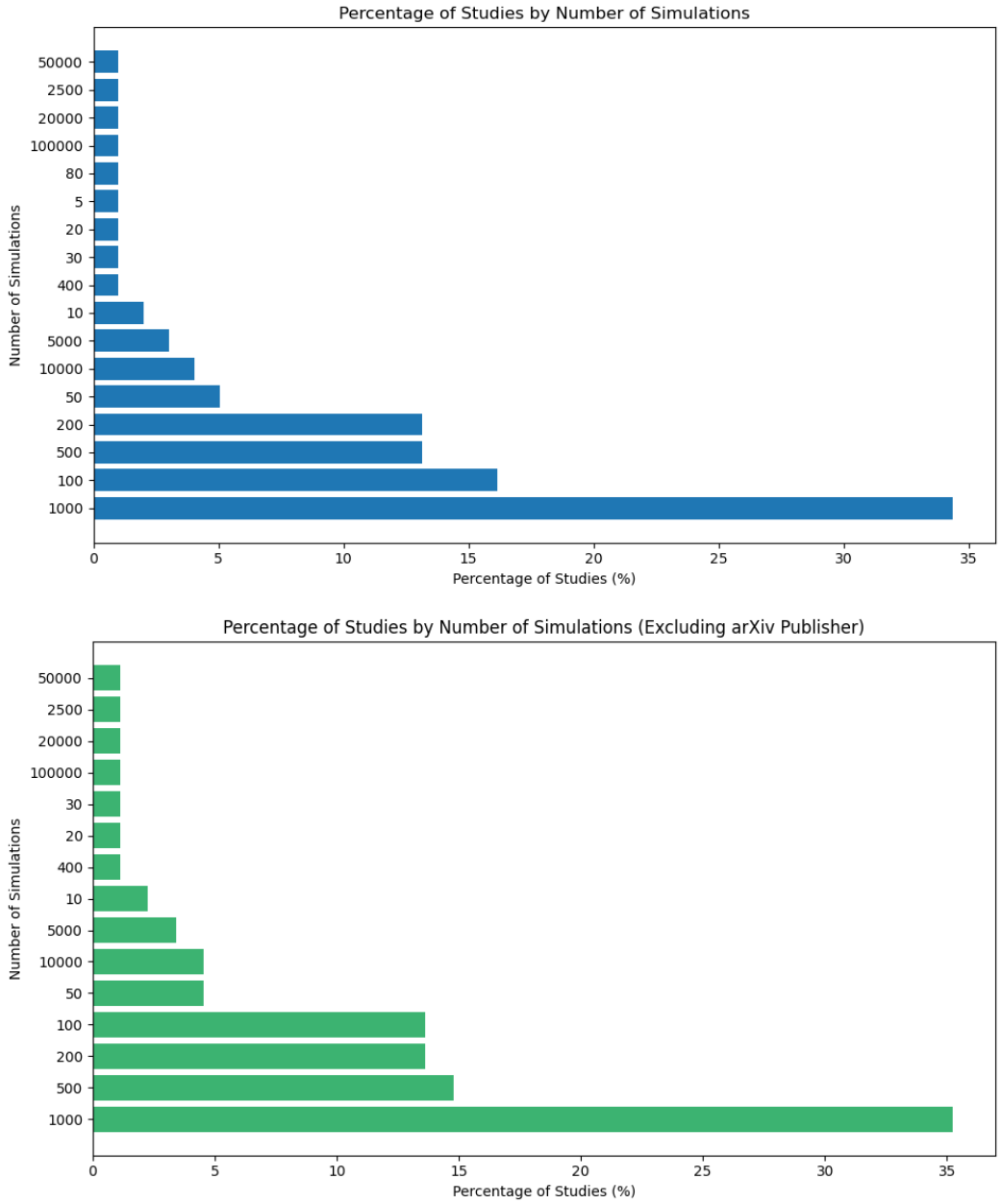

Figure 10 – Percentage of Studies by Number of Simulations

Despite this wide variation, a clear mode exists: $J = 1000$ simulations is the most frequent choice, accounting for approximately 35% of studies in both the full sample and the sample excluding arXiv publications. This confirms the robustness of this observation and highlights a common, albeit potentially unfounded, convention in the field. The pronounced heterogeneity, however, strongly indicates a lack of consensus or standardized, evidence-based guidelines for determining the appropriate simulation effort.

Crucially, simply adopting the most common practice ($J = 1000$) as a de facto standard, which would be an alternative for the solution proposed in this paper, would be highly problematic and lack rigor. As indicated by the distributions in Figure 10, a substantial portion of the literature utilizes significantly fewer simulations. Specifically,

53.54% of all studies (and 53.41% of non-arXiv studies) in our sample used 500 or fewer simulations – meaning that adopting $J = 1000$ as a standard would implicitly classify over half of the recent simulation studies analyzed as potentially insufficient, employing at least two times fewer replications than this arbitrary benchmark.

More fundamentally, selecting $J = 1000$ merely based on its popularity is not rooted in any statistical or scientific principle related to the goals of the simulation study, such as achieving adequate statistical power for model comparison. It reflects convention rather than a rigorous assessment of need. This is precisely where the algorithm proposed in this paper offers a significant improvement. Instead of relying on arbitrary conventions or educated guesses, our approach provides a statistically principled method to determine the necessary number of simulations (J) based on researcher-specified parameters for significance (α), power ($1-\beta$), and the minimum detectable effect size (MDE) for the comparison of interest. This ensures that the simulation effort is both justified and sufficient, promoting higher quality, more reliable, and more efficient research in the field of treatment effect estimation.

3.3 The Proposed Algorithm: TISCA

Addressing the identified methodological gaps in the evaluation of treatment effect estimators—namely, the prevalent lack of formal statistical testing for model comparisons and the arbitrary selection of simulation replications—requires a systematic and statistically grounded approach. To this end, we introduce the **T**est-**I**nformed **S**imulation **C**ount **A**lgorithm (TISCA). TISCA is designed to integrate rigorous hypothesis testing directly into the Monte Carlo simulation workflow, simultaneously determining the minimum number of simulations (J) required to achieve a pre-specified statistical power for detecting meaningful differences in model performance.

3.3.1 Methodological Foundation: The Welch's t-test

At the core of TISCA lies the comparison of performance metrics (e.g., PEHE, RMSE_{ATE}, Coverage, CIL) between the proposed model and the most performing benchmark model across J simulation runs, which consequently means that if the proposed model is statistically significantly superior to the most performing benchmark model for a certain performance metric, it must also be superior to the other benchmark models. A natural choice for comparing the means of these metrics between two groups (e.g., performance of model A vs. model B across simulations) is the t-test (WELCH, 1947). However, a standard Student's t-test relies on the assumption of equal variances between the two groups being compared (WELCH, 1947).

During our bibliometric analysis, alongside general observations in simulation stud-

ies, we noticed that that the variability (standard deviation or variance) of performance metrics across Monte Carlo runs often differs between models, especially models that are highly different in nature (e.g., OLS linear model and Bayesian Causal Forest (HAHN; MURRAY; CARVALHO, 2020)). For instance, a more complex model might exhibit higher variance in its PEHE estimates across simulations compared to a simpler, more stable benchmark. Violating the homogeneity of variance assumption can lead to unreliable p-values and potentially incorrect conclusions if a standard t-test is employed (ZIMMERMAN, 2004).

Therefore, TISCA utilizes the **Welch's t-test** (WELCH, 1947) as its default comparison method. The Welch's t-test does not assume equal variances and adjusts its degrees of freedom accordingly using the Welch-Satterthwaite equation, providing a more robust comparison when the variances of the performance metrics between models are unequal.

A second assumption for the t-test (both Student's and Welch's) is the normality of the underlying data within each group. In our context, this refers to the distribution of a specific performance metric (e.g., PEHE values) obtained from the J simulation runs for a given model. While true normality is seldom guaranteed, we leverage the **Central Limit Theorem (CLT)**. Many key performance metrics, such as RMSE_{ATE} (an average error) and PEHE (average squared error), are derived from averaging or summing operations across units within each simulation. As the number of simulation replications (J) increases, the distribution of the mean of these metrics tends towards normality due to the CLT (BROSAMLER, 1988). We consider this assumption reasonably met, particularly for the sample sizes (J) typically encountered or targeted in simulation studies (e.g., $J >> 30$), which aligns with common heuristics for the applicability of the CLT (KWIATKOWSKI *et al.*, 1992).

Crucially, TISCA deliberately *avoids* performing formal statistical tests for normality (e.g., Shapiro-Wilk test) prior to conducting the Welch's t-test. The practice of "pre-testing" assumptions like normality has been widely criticized in the statistical literature (ROCHON; GONDAN; KIESER, 2012; ZIMMERMAN, 2004; RASCH; KUBINGER; MODER, 2009; SCHUCANY; NG, 2006; DEAN; VOSS, 1999). Such pre-testing conditions the subsequent primary test (the Welch's t-test in our case) on the outcome of the pre-test, altering the true Type I error rate and distorting the statistical inference. Given the known robustness of the t-test (especially Welch's) to moderate deviations from normality, particularly with reasonable sample sizes (J) (WELCH, 1947; ZIMMERMAN, 2004), and the established issues with pre-testing, we proceed directly with the Welch's t-test, relying on the CLT for justification.

3.3.2 The TISCA Algorithm Workflow

TISCA operates iteratively, executing batches of simulations and evaluating statistical power until a desired threshold is met. The researcher initiates the process by providing the following inputs:

- **Simulation Function/Code:** A user-provided function or script that executes one full simulation run. This function must take the DGP settings as input and return a data frame containing the calculated performance metrics for all models being compared in that single simulation.
- **Performance Metrics:** A list of column names from the simulation output corresponding to the performance metrics to be formally compared (e.g., c("pehe_modelA", "pehe_modelB", "rmse_ate_modelA", "rmse_ate_modelB")). Comparisons are typically set up between the proposed model and one or more benchmarks for each metric.
- **Target Statistical Power** $(1-\beta)$: The desired probability of detecting a true difference if it exists (the standard in the scientific community is 0.80). Needless to say, the smaller the value for β, the higher the number of required simulations will be needed when holding the other inputs the same. Yet, the smaller the value for β, the more certain the researcher will be about the superiority of their proposed model, if it is indeed superior.
- **Significance Level**(α): The threshold for statistical significance, controlling the Type I error rate (the standard in the scientific community is 0.05). Needless to say, the smaller the value for α, the higher the number of required simulations will be needed when holding the other inputs the same. Yet, the smaller the value for α, the more certain the researcher will be about the superiority of their proposed model, if it is indeed superior.
- **Minimum Detectable Effect Sizes** (MDEs, δ): A vector specifying the smallest true difference in the mean of each performance metric (Proposed Model Mean - Benchmark Model Mean) that the researcher wants to be able to detect with the target power. This is a crucial parameter reflecting practical significance. For instance, $\delta_{PEHE} = -0.1$ might indicate the goal is to detect if the proposed model's average PEHE is at least 0.1 units lower than the benchmark's. These values can be informed by pilot runs (e.g., 50 simulations) or domain knowledge.
- **Comparison Pairs:** Specification of which model metrics to compare against each other (e.g., compare 'pehe_modelA' vs 'pehe_modelB', 'rmse_ate_modelA' vs 'rmse_ate_modelB').

- **Simulation Batch Size:** (B) The number of new simulations to run in each iteration before recalculating power (e.g., $B = 50$).

- **Initial Simulation Count** (J_0): An optional starting number of simulations (e.g., $J_0 = 50$) to perform before the first power check. This ensures sufficient data for initial variance estimates. Defaults to ‘batch_size‘.

- **Multiple Testing Correction Method:** The chosen method to adjust p-values if multiple hypotheses are tested simultaneously (options: ‘”none”‘, ‘”bonferroni”‘, ‘”holm”‘, ‘”BH”‘ (Benjamini-Hochberg)). Default is ‘”none”‘.

TISCA operates iteratively, executing batches of simulations and evaluating statistical power until a desired threshold is met. The algorithm requires several inputs specifying the simulation setup, the desired statistical guarantees, and the comparisons of interest. It then proceeds according to the pseudo-code outlined below:

Algorithm 1: TISCA: Test-Informed Simulation Count Algorithm (Part 1 of 2)

Input : User simulation function $SimFunc(seed)$ returning performance metrics;
List of K comparison pairs
$Comparisons = \{(metric_k, modelP_k, modelB_k)\}_{k=1}^{K}$;
Vector of K Minimum Detectable Effect Sizes $MDEs = \{\delta_k\}_{k=1}^{K}$;
Target statistical power P_{target} (e.g., 0.80);
Significance level α (e.g., 0.05);
Simulation batch size B;
Initial simulation count J_0 (optional, e.g., 50);
Multiple testing correction method
$CorrMethod \in$ {"none", "bonferroni", "holm", "BH"};

// Initialization Phase
$J \leftarrow 0$;
$results\_agg \leftarrow \emptyset$; // e.g., an empty dataframe
$P_{current} \leftarrow \mathbf{0}_K$; // vector of K zeros
// Optional Initial Run
if $J_0 > 0$ and $J_0 \geq B$ then
  for $i = 1$ to J_0 do
    $metrics\_run \leftarrow SimFunc(seed = i)$;
    Append $metrics\_run$ to $results\_agg$;
  $J \leftarrow J_0$;
  // Perform initial power calculation before main loop
  $p_{values\_raw\_init} \leftarrow \mathbf{0}_K$;
  $test\_stats\_init \leftarrow \mathbf{0}_K$;
  for $k = 1$ to K do
    if $sd_P > 0$ and $sd_B > 0$ then
      $P_{current}[k] \leftarrow$ EstimateWelchPower$(J, J, sd_P, sd_B, \delta_k, \alpha)$;
    else $P_{current}[k] \leftarrow 0$;

Algorithm 1: TISCA: Test-Informed Simulation Count Algorithm (Part 2 of 2)

Output: Final required simulation count J_{final};
Vector of final raw p-values $P_{raw} = \{p_{raw,k}\}_{k=1}^{K}$;
Vector of final adjusted p-values $P_{adj} = \{p_{adj,k}\}_{k=1}^{K}$;
Vector of final test statistics $Stats = \{stat_k\}_{k=1}^{K}$;
Vector of final estimated achieved powers $P_{achieved} = \{p_{achieved,k}\}_{k=1}^{K}$;

// Iterative Simulation and Power Check Loop
1 while $\min(P_{current}) < P_{target}$ do
    // Run a new batch of simulations
2   $start\_seed \leftarrow J + 1$;
3   $end\_seed \leftarrow J + B$;
4   for $i = start\_seed$ to end_seed do
5     $metrics\_run \leftarrow SimFunc(seed = i)$;
6   Append $metrics\_run$ to $results\_agg$;
7   $J \leftarrow J + B$;
    // Perform tests and estimate power on current J simulations
8   for $k = 1$ to K do
9   if $sd_P > 0$ and $sd_B > 0$ then // Need valid SDs
10      $P_{current}[k] \leftarrow \text{EstimateWelchPower}(J, J, sd_P, sd_B, \delta_k, \alpha)$;
11   else $P_{current}[k] \leftarrow 0$ // Cannot estimate power yet
12   ;

// Output Phase
13 $J_{final} \leftarrow J$;
14 $P_{raw} \leftarrow p_{values\_raw}$; // Final raw p-values
15 $P_{adj} \leftarrow p_{values\_adj}$; // Final adjusted p-values
16 $Stats \leftarrow test\_stats$; // Final test statistics
17 $P_{achieved} \leftarrow P_{current}$; // Final estimated powers
18 return $J_{final}, P_{raw}, P_{adj}, Stats, P_{achieved}$;

This pseudo-code describes the iterative process: initializing, optionally running a pilot batch, then looping through batches of simulations, performing Welch's t-tests on the accumulated data, applying multiple testing corrections, estimating the current power for each comparison based on the observed variances and the target MDE, and stopping only when the power target is achieved for all comparisons. Helper functions like 'SimFunc', 'Welch_t_test', 'AdjustPValues', 'StandardDeviation', and 'EstimateWelchPower' represent the underlying necessary computational steps detailed elsewhere in the text or implemented in the actual code.

This iterative process ensures that simulations continue precisely until the study has sufficient power to detect the pre-defined minimally important differences, providing a statistically justified stopping point and simulation count (J).

3.3.3 Addressing Multiple Comparisons

Researchers often compare a new model against benchmarks across multiple performance metrics (e.g., PEHE, RMSE_{ATE}, Coverage) or against multiple different benchmarks simultaneously. Performing multiple hypothesis tests increases the probability of making at least one Type I error (a false positive finding) across the family of tests – the "multiple comparisons problem" (STREINER; NORMAN, 2011; MENON, 2019).

TISCA acknowledges this issue by offering optional p-value adjustment procedures as an input parameter. The available methods include:

- **Bonferroni correction:** A simple but often overly conservative method that controls the Family-Wise Error Rate (FWER) by multiplying each p-value by the number of tests (or equivalently, dividing α by the number of tests) (DUNN, 1961).
- **Holm's method (Holm-Bonferroni):** A step-down procedure that also controls the FWER but is uniformly more powerful than the standard Bonferroni correction (HOLM, 1979).
- **Benjamini-Hochberg (BH) procedure:** Controls the False Discovery Rate (FDR) – the expected proportion of rejected null hypotheses that are actually true (BENJAMINI; HOCHBERG, 1995). This is generally less conservative and more powerful than FWER-controlling methods, making it suitable when controlling the proportion of false positives among the significant findings is the primary goal.

The choice of correction method depends on the researcher's specific goals and tolerance for Type I versus Type II errors. While these methods provide established ways to handle multiple tests, they are not perfect solutions; FWER methods can be overly strict, potentially masking true effects, while FDR control allows for some false positives among the declared significant results (MENON, 2019). The default setting of ‘"none"‘ assumes that only a few tests will be performed (say two or three), yet if more tests are performed, then we advise researchers choose one of the adjustment procedures considering their trade-offs.

3.3.4 Implementation and Practical Use

TISCA is designed to be implemented as a flexible function, envisioned primarily within the R statistical environment, leveraging existing packages for Welch's t-tests (‘stats::t.test‘) and p-value adjustments (‘stats::p.adjust‘).

To facilitate the adoption and use of TISCA, particularly for researchers who may be less familiar with power analysis or R programming, we have developed a companion web-based tool. Hosted on <https://tisca-llm-app.streamlit.app/>, this tool features an interactive interface powered by a fine-tuned LLM, namey Gemini 2.5 Flash. The LLM

is fine-tuned on the TISCA methodology and its R implementation details, allowing researchers to describe their simulation study setup and receive guidance on structuring their simulation code, choosing appropriate parameters (like MDEs based on pilot data), and interpreting the TISCA output. This aims to lower the barrier to entry for adopting more rigorous simulation practices.

3.3.5 Advantages and Limitations

This approach offers statistical rigor by replacing arbitrary choices of J with a statistically principled method grounded in power analysis and formal hypothesis testing (Welch's t-test). It provides a justified simulation count, offering a clear, data-driven rationale for the number of simulations, thereby enhancing reproducibility and credibility. The method promotes efficiency by avoiding unnecessary computational costs; simulations stop once sufficient power is achieved, saving resources and time compared to using an overly large, fixed J. It also prevents underpowered studies by ensuring simulations run until the desired power is reached. Furthermore, it fosters a focus on effect size, requiring researchers to explicitly define the MDE and encouraging the consideration of practical significance alongside statistical significance. Finally, it offers flexibility, accommodating various performance metrics, multiple comparisons (with appropriate corrections), and user-defined simulation code.

Despite its advantages, this method has a few limitations. The resulting J is highly dependent on the MDE chosen; specifying an extremely small MDE can lead to computationally prohibitive simulation counts, making careful consideration and justification of the MDE crucial. While potentially more efficient than over-simulation, the iterative power calculation itself adds some computational overhead, especially if power estimation involves many simulations. The overall cost still heavily depends on the complexity of a single simulation run. It also relies on certain assumptions, specifically the appropriateness of Welch's t-test and the adequacy of the CLT approximation for the distributions of performance metrics. While robust, severe violations of these assumptions could affect results. Lastly, the sequential testing nature, which involves repeated checks on accumulating data, means the stopping rule itself is data-dependent. While the final inference uses the full J dataset and standard tests/corrections, and the implications for error rates are generally considered minor in the context of achieving a target power for a frequentist test based on pre-specified MDE and alpha (JENNISON; TURNBULL, 2000), it does represent a departure from fixed-sample designs.

Despite these limitations, TISCA offers a substantial improvement over current common practices by embedding statistical rigor and efficiency directly into the design and execution of Monte Carlo simulation studies for treatment effect estimation model evaluation.

3.4 Real Life Example: Revisiting MCJAMES *et al.* (2024)

To illustrate the practical application and utility of TISCA, we revisit the simulation study presented in MCJAMES *et al.* (2024). This recent work introduced the Multivariate Bayesian Causal Forest (MVBCF) model for estimating treatment effects for multiple outcomes simultaneously. Their evaluation included extensive Monte Carlo simulations, namely three different DGPs and 1000 replications for each, comparing MVBCF against relevant benchmarks, including standard Bayesian Causal Forests (BCF) (HAHN; MURRAY; CARVALHO, 2020) applied separately to each outcome (denoted here as 'wsBCF' or simply 'BCF' where context is clear), Bayesian Additive Regression Trees (BART) (CHIPMAN; GEORGE; MCCULLOCH, 2010), and a multivariate version of BART (MVBART) (UM *et al.*, 2022). By applying TISCA while replicating their a part of simulation studies, we aim to demonstrate how our algorithm could have provided formal statistical evidence for model comparisons and determined a sufficient, potentially more efficient, number of simulation replications.

It is worth mentioning that for this study case, the programming laguage R was utilized, and the following R libraries were used: 1. progress (for the iterations), 2. dbarts (for the BART model), 3. stochtree (for the BCF model), 4. mvbcf (for the MVBCF model), 5. skewBART (for the MVBART model), and 6. mvtnorm (for the DGPs creation).

Our analysis specifically focuses on the outcomes reported for the first DGP (DGP1) in MCJAMES *et al.* (2024) with a training sample size of n=500 observations and test size of 1000 observations. This focused scope is chosen deliberately. DGP1 was constructed by MCJAMES *et al.* (2024) to represent conditions where the assumptions behind the MVBCF model are met, namely where the prognostic baseline (μ) and the treatment effect modification (τ) components for the two outcome variables, Y_1 and Y_2, share influential predictors and exhibit similar functional dependencies. Evaluating performance in this "ideal" scenario provides a critical test case for the primary claims of the model's superiority. Furthermore, the detailed numerical results for the DGPs of the paper are directly available in the main text of the original publication (their Table 2) only for n=500, motivating us to use n=500, albeit MCJAMES *et al.* (2024) have also explored training sample sizes of 100 and 1000 (in all cases the test size remains 1000 observations). While MCJAMES *et al.* (2024) presented two additional DGPs, our purpose here is illustrative—to showcase TISCA's utility—rather than performing an exhaustive replication. Thus, concentrating on DGP1 allows for a clear demonstration while maintaining conciseness.

In DGP1, MCJAMES *et al.* (2024) generated ten covariates $(X_1, \ldots, X_{10})$ with a mix of distributions intended to mimic their real-world Trends in International Mathematics and Science Study (TIMSS) data (to which they applied their proposed model in their paper): $X_1, \ldots, X_5 \sim U(0,1)$; $X_6, \ldots, X_8 \sim \text{Bernoulli}(0.5)$; and X_9, X_{10} as ordinal vari-

ables with five equally likely levels $\{0,1,2,3,4\}$. The treatment assignment D_i followed $\pi(\mathbf{X_i}) = X_{4,i}$, making X_4 an observed confounder. The outcomes $Y_{1,i}$ and $Y_{2,i}$ were generated according to the following functional forms (adapted from their Table 1 (MCJAMES *et al.*, 2024)):

$$Y_{1,i} = \underbrace{\left(300 + 110\sin(\pi X_{1,i}X_{2,i}) + 180(X_{3,i} - 0.5)^2 + 100X_{4,i} + 120X_{6,i} + 10X_{9,i}\right)}_{\mu_1(\mathbf{X_i})} + \underbrace{\left(20X_{4,i} + 20X_{5,i}\right)}_{\tau_1(\mathbf{X}_i)} D_i + \varepsilon_{1,i} \tag{3.8}$$

$$Y_{2,i} = \underbrace{\left(300 + 90\sin(\pi X_{1,i}X_{2,i}) + 220(X_{3,i} - 0.5)^2 + 140X_{4,i} + 80X_{6,i} + 10X_{9,i}\right)}_{\mu_2(\mathbf{X_i})} + \underbrace{\left(10X_{4,i} + 30X_{5,i}\right)}_{\tau_2(\mathbf{X}_i)} D_i + \varepsilon_{2,i} \tag{3.9}$$

where the error terms $(\varepsilon_{1,i}, \varepsilon_{2,i})$ were drawn from a multivariate normal distribution $MVN(\mathbf{0}, \Sigma)$ with $\Sigma = 50^2\mathbf{I}$, calibrating the signal-to-noise ratio based on residual variance observed in their target TIMSS dataset.

The subsequent DGPs served essentially as sensitivity analyses. DGP2 assessed robustness to unobserved confounding by modifying DGP1: the effect of X_4 on μ_2 was removed, and X_4 was excluded from the available covariates. This setup induced confounding only for Y_1, allowing the authors to investigate if performance degradation was isolated, which was in their presented results (MCJAMES *et al.*, 2024). On the other hand, DGP3 examined model flexibility by altering the functional forms within DGP1: different interaction terms were used in μ_1, and different linear combinations of covariates were used for τ_1 and τ_2. This tested whether MVBCF's structure could adapt when the two outcomes necessitated distinct tree representations for their prognostic and treatment effect components, which it could, though not perfectly (MCJAMES *et al.*, 2024). Given that DGPs 2 and 3 probe robustness rather than baseline performance under favorable conditions, focusing our TISCA demonstration on DGP1 remains the most direct way to illustrate its utility in rigorously comparing models and determining simulation sufficiency for core performance claims.

Table 5 reproduces the key performance metrics for DGP1 with n=500 from Table 2 in MCJAMES *et al.* (2024), focusing on the comparison between MVBCF, BCF (wsBCF), and MVBART for PEHE and 95% coverage of the true treatment effect ($\tau(\mathbf{X})$).

The original study reported the following key results (Mean ± SE across 1000 simulations) for DGP1, n=500:

Based on these results, the authors concluded: "Looking at the PEHE results from

Table 5 – Selected Simulation Results from MCJAMES *et al.* (Table 2, DGP1, n=500). Values are Mean ± SE.

Metric	Outcome	MVBCF	BCF	MVBART
PEHE on τ	Y_1	$\mathbf{9.05 \pm 0.16}$	9.63 ± 0.16	10.29 ± 0.19
	Y_2	$\mathbf{9.40 \pm 0.16}$	9.96 ± 0.16	10.83 ± 0.19
τ 95% Coverage	Y_1	$\mathbf{0.96 \pm 0.00}$	0.97 ± 0.00	0.98 ± 0.00
	Y_2	$\mathbf{0.95 \pm 0.00}$	0.96 ± 0.00	0.98 ± 0.00

DGP1... multivariate BCF clearly outperforms the other three methods when tasked with accurately predicting heterogeneity in the treatment effect τ... The PEHE, bias, and coverage results from DGP1 in Table 2 tell a very similar story... multivariate BCF shows minimal bias... and the 95% coverage rate is close to ideal." (MCJAMES *et al.*, 2024).

While the point estimates suggest MVBCF performs favorably regarding PEHE and coverage, the differences are not remarkable, especially in comparison to the BCF model given that the actual ATE for this DGP is around 20 units. Without formal statistical testing, it remains ambiguous whether these observed differences, based on 1000 simulations, reflect genuine superiority or could be attributed to Monte Carlo variability. This ambiguity motivates the application of TISCA to formally test these comparisons.

For the replication of the simulation study for DGP1 (n=500), we focused on the following six key comparisons motivated by the original paper's discussion and results:

1. MVBCF vs. BCF: 95% τ Coverage for Y_1 ('mvbcf_tau_951' vs 'wsbcf_tau_951')
2. MVBCF vs. BCF: 95% τ Coverage for Y_2 ('mvbcf_tau_952' vs 'wsbcf_tau_952')
3. MVBCF vs. BCF: PEHE for Y_1 ('mvbcf_pehe1' vs 'bcf_pehe1')
4. MVBCF vs. BCF: PEHE for Y_2 ('mvbcf_pehe2' vs 'bcf_pehe2')
5. MVBCF vs. MVBART: 95% τ Coverage for Y_1 ('mvbcf_tau_951' vs 'mvbart_tau_-951')
6. MVBCF vs. MVBART: 95% τ Coverage for Y_2 ('mvbcf_tau_952' vs 'mvbart_tau_-952')

We configured TISCA with standard parameters: a significance level $\alpha = 0.05$ and a target statistical power $1 - \beta = 0.80$. Based on the observed differences in the original study and representing plausible thresholds for practical significance, we set the Minimum Detectable Effect Sizes (MDEs, denoted δ) as follows:

- For PEHE comparisons (MVBCF vs. BCF): $\delta = 0.5$ (reflecting the approximate observed difference, testing if MVBCF PEHE is lower).

- For Coverage comparisons (MVBCF vs. BCF/MVBART): $\delta = 0.015$ (slightly larger than the observed 0.01 difference vs BCF, smaller than the average 0.025 difference vs MVBART, testing if MVBCF coverage is closer to 0.95, hence potentially lower than BCF/MVBART's coverage which exceeded 0.95).

We performed TISCA's iterative process using a batch size $B = 100$.

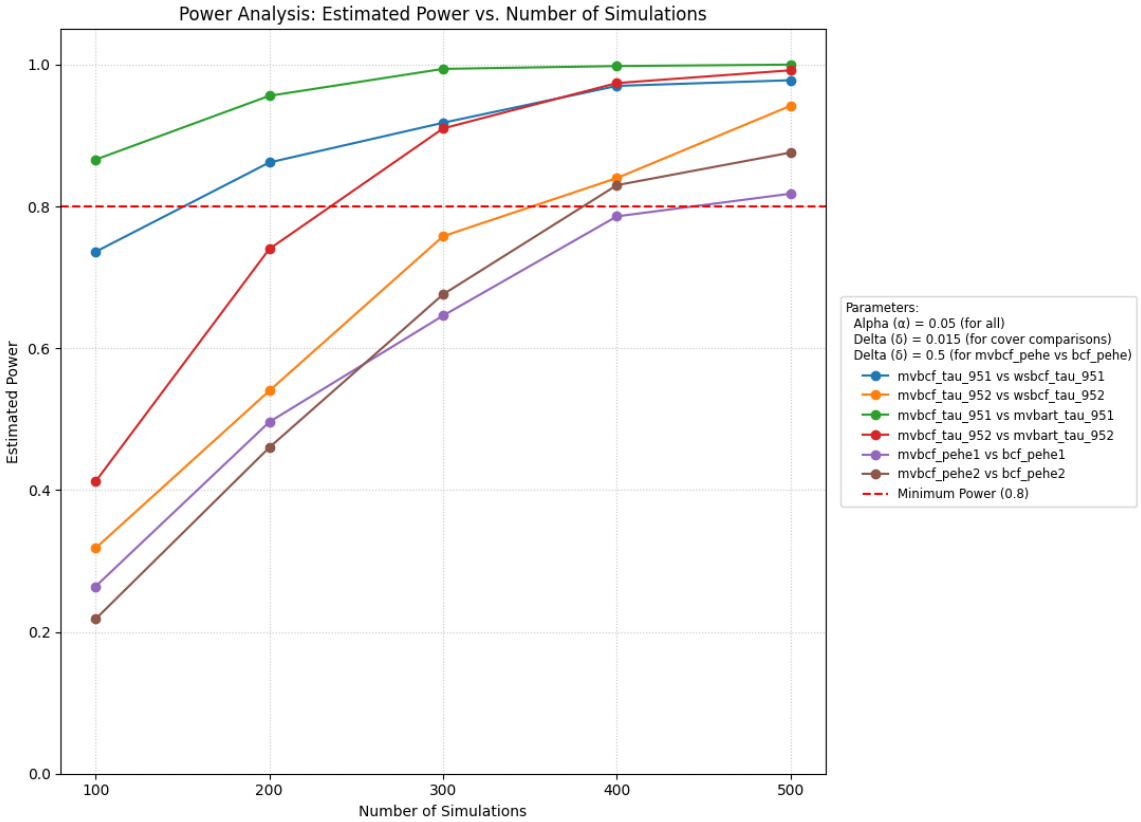

Figure 11 – TISCA Power Analysis for MCJAMES *et al.* Simulation (DGP1, n=500). Estimated statistical power for detecting specified MDEs (δ) for key comparisons as a function of the number of simulation replications (J). The red dashed line indicates the target power of 0.8.

Figure 11 shows the estimated statistical power for each of the six comparisons as TISCA accumulates simulation results. The power curves demonstrate how the ability to detect the specified MDE increases with the number of simulations J. Critically, for all six hypotheses, the estimated power reaches or exceeds the target threshold of 0.80 at or before $J = 500$ simulations. This provides strong evidence that, for the chosen MDEs and significance level, performing 500 simulations would have been sufficient to achieve the desired statistical power for these specific comparisons.

Having determined that $J = 500$ simulations suffice based on the power analysis, we examine the results of the Welch's t-tests performed by TISCA after these 500 runs. The raw p-values and p-values adjusted for multiple comparisons (using Bonferroni, Holm, and

Benjamini-Hochberg methods for the family of 6 tests) are presented in Table 6. Statistical significance is indicated using asterisks. Figure 12 provides a visual representation of these results on a -log10 scale.

Comparison Test	Raw p-value	Adj. (Bonf.)	Adj. (Holm)	Adj. (BH)
MVBCF vs BCF (τ Cov, Y_1)	3.40×10^{-1}	1.00	3.40×10^{-1}	3.40×10^{-1}
MVBCF vs BCF (τ Cov, Y_2)	2.28×10^{-2}**	1.37×10^{-1}	4.55×10^{-2}**	2.73×10^{-2}**
MVBCF vs BCF (PEHE, Y_1)	3.88×10^{-34}***	2.33×10^{-33}***	1.94×10^{-33}***	1.16×10^{-33}***
MVBCF vs BCF (PEHE, Y_2)	2.36×10^{-35}***	1.42×10^{-34}***	1.42×10^{-34}***	1.42×10^{-34}***
MVBCF vs MVBART (τ Cov, Y_1)	1.51×10^{-9}***	9.06×10^{-9}***	6.04×10^{-9}***	3.02×10^{-9}***
MVBCF vs MVBART (τ Cov, Y_2)	3.42×10^{-9}***	2.05×10^{-8}***	1.03×10^{-8}***	5.14×10^{-9}***

Table 6 – Welch's t-test Results for Key Comparisons after $J = 500$ Simulations (DGP1, n=500). Significance levels: ** $p < 0.05$, *** $p < 0.01$.

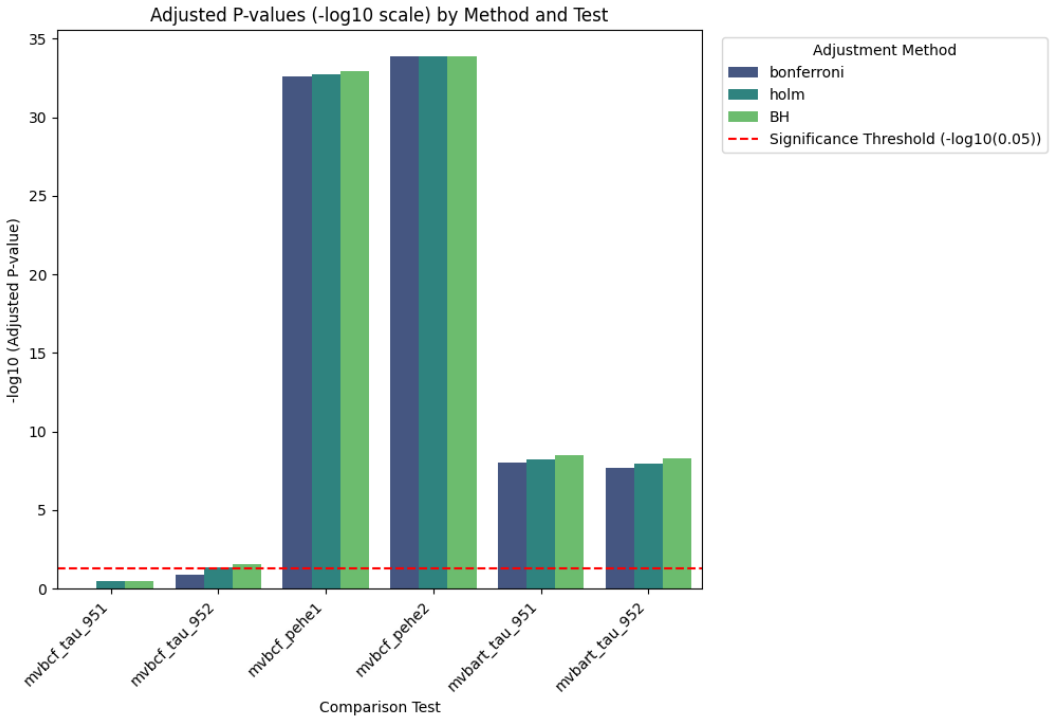

Figure 12 – Original vs. Adjusted P-values from Welch's t-tests after $J = 500$ simulations for MCJAMES *et al.* comparisons (DGP1, n=500). Values are shown on a -log10 scale (higher bars indicate smaller p-values/stronger evidence against the null hypothesis of no difference). The red dashed line represents the significance threshold ($\alpha = 0.05$, i.e., -log10(0.05) $\approx$ 1.3).

The p-values, both unadjusted and adjusted, provide a nuanced picture. The extremely small p-values, both raw and adjusted, for the PEHE comparisons (mvbcf_pehe1 vs bcf_pehe1 and mvbcf_pehe2 vs bcf_pehe2) offer overwhelming statistical evidence that MVBCF achieves significantly lower PEHE than the standard BCF implementation in this scenario. This strongly supports the authors' claims regarding MVBCF's accuracy

in estimating heterogeneous effects. Comparisons of τ coverage between MVBCF and MVBART (mvbart_tau_951 and mvbart_tau_952) also yield highly significant p-values across all adjustment methods, indicating a statistically significant difference in coverage properties between these two multivariate approaches in this setting. However, the results for τ coverage comparison between MVBCF and BCF are mixed. For Y_1 (mvbcf_tau_951), the difference is not statistically significant ($p > 0.05$) regardless of adjustment. For Y_2 (mvbcf_tau_952), the raw p-value (0.023) is below 0.05 and remains significant after Holm and BH correction, but not after the stricter Bonferroni correction, suggesting weaker evidence for a difference in coverage between MVBCF and BCF compared to the other tested differences. Overall, TISCA's analysis largely substantiates the authors' claims regarding MVBCF's superior PEHE performance in DGP1, but adds nuance by highlighting that the evidence for superior or different coverage performance compared to standard BCF is less conclusive, particularly for outcome Y_1.

Nonetheless, perhaps the most salient finding from this retrospective TISCA application is the determination that $J = 500$ simulations were sufficient to achieve 80% power for detecting the specified effect sizes across all six key comparisons. The original study performed $J = 1000$ simulations. While performing additional simulations generally increases power and precision, TISCA demonstrates that, for the stated goals (detecting MDEs of 0.5 for PEHE and 0.015 for coverage with 80% power), the final 500 simulations performed by MCJAMES *et al.* (2024) were potentially redundant.

This redundancy has significant practical implications. For our simulation studies replication, it took on average 10 hours to run 100 simulations using an Intel(R) Xeon(R) CPU @ 2.20GHz on Google Colab; thus, running their full set of simulations would require approximately 100 hours of computational time. Halving the number of simulations for DGP1 (and potentially for other DGPs, had TISCA been applied there) could have saved roughly **50 hours of computation**. This translates directly to increased research efficiency, as the saved computational resources and researcher time could be redirected towards other valuable activities, such as exploring additional DGPs, testing sensitivity to hyperparameters, analyzing more real-world data, or developing further methodological improvements. Additionally, it contributes to enhanced sustainability, as reducing unnecessary computation lowers the energy consumption and associated environmental footprint of research activities, supporting more sustainable scientific practices—a growing concern in computationally intensive fields (LANNELONGUE; GREALEY; INOUYE, 2021). Finally, this approach leads to improved reliability and comparability; by providing a statistically justified simulation count, TISCA enhances the reliability of the study's conclusions regarding statistical power. Furthermore, widespread adoption of such principled approaches, rather than arbitrary choices of J, would significantly improve the comparability of simulation results across different studies in the treatment effect estimation literature.

In summary, this case study demonstrates how TISCA can be applied to real-world simulation studies to provide formal statistical comparisons and determine an adequate number of simulations. It confirms key findings of the original study (MCJAMES *et al.*) (2024) regarding PEHE while adding statistical nuance to coverage comparisons, and critically highlights the potential for significant gains in computational efficiency and research sustainability without sacrificing statistical rigor.

3.5 Conclusion

This paper addressed critical methodological shortcomings prevalent in the evaluation of treatment effect estimators via Monte Carlo simulation studies in the current literature: the widespread absence of formal statistical hypothesis testing for model comparisons and the often arbitrary selection of the number of simulation replications (J). As evidenced by our bibliometric analysis, current practices exhibit significant heterogeneity in the choice of J, frequently relying on convention (such as $J = 1000$) rather than statistical principles. This lack of rigor can lead to underpowered studies, unsubstantiated claims of model superiority, inefficient use of computational resources, and difficulties in comparing findings across the literature—concerns that are amplified in the current era of rapidly proliferating AI-driven research where robust validation is paramount (GOLDKUHLE *et al.*, 2024; VALE *et al.*, 2024).

As a solution, we introduced the Test-Informed Simulation Count Algorithm (TISCA). TISCA provides a systematic and statistically grounded framework designed to integrate rigorous hypothesis testing directly into the simulation workflow. By leveraging the robust Welch's t-test and principles of statistical power analysis, TISCA iteratively performs simulations until a user-specified power level (e.g., 0.80) is achieved for detecting pre-defined Minimum Detectable Effect Sizes (MDEs) at a given significance level (α, e.g., 0.05). It explicitly accounts for potential unequal variances between comparison groups and incorporates options for multiple testing corrections (Bonferroni Correction (DUNN, 1961), Holm's method (HOLM, 1979), and Benjamini-Hochberg procedure (BENJAMINI; HOCHBERG, 1995)) when necessary.

The practical utility and benefits of TISCA were demonstrated through a case study revisiting the simulation experiments of MCJAMES *et al.* (2024). Our retrospective application revealed that for their primary simulation scenario, achieving 80% power for detecting meaningful differences in PEHE and coverage required only half the number of simulations ($J = 500$) originally performed ($J = 1000$). This finding highlights the potential for substantial gains in computational efficiency—saving valuable research time and resources—and contributes to more environmentally sustainable research practices (LANNELONGUE; GREALEY; INOUYE, 2021). Furthermore, the TISCA analysis pro-

vided formal statistical evidence that largely supported the original study's conclusions regarding the superior PEHE of their proposed MVBCF model, while adding important nuance regarding the statistical significance of observed differences in coverage metrics compared to benchmarks.

By adopting TISCA, researchers can move beyond arbitrary choices for J, ensuring their simulation studies are adequately powered to detect effects deemed practically significant, while simultaneously avoiding wasteful over-simulation. This methodology fosters greater transparency, reproducibility, and comparability within the field of treatment effect estimation. The provision of a statistically justified simulation count strengthens the credibility of research findings and facilitates more reliable assessments of novel estimators. To aid adoption, we have also introduced plans for a companion web-based tool leveraging a fine-tuned LLM to guide researchers in applying TISCA to their specific simulation setups.

While TISCA offers significant advantages, we acknowledge its limitations, primarily the crucial dependence on the thoughtful specification of the MDE and the computational overhead associated with iterative power estimation. Future research could explore extensions of TISCA, such as incorporating alternative robust statistical tests suitable for different types of performance metrics or comparison scenarios (e.g., non-inferiority testing, multiple comparisons with a control). Further investigation into optimizing the power estimation step through the use of C++ code instead of R code and refining the guidance provided by the LLM tool also represent valuable avenues for development.

In conclusion, TISCA provides a needed methodological advancement for conducting simulation studies in causal inference and beyond. By embedding statistical rigor and efficiency into the core of the evaluation process, it offers a pathway towards more reliable, resource-conscious, and comparable research, ultimately strengthening the foundation upon which new methods for treatment effect estimation are developed, validated, and compared.

Supplementary Research Material and Code

The code for TISCA as well as the code used Section 2.4 can be found in the following GitHub repository: Repository. Additionally, the website to facilitate the use of TISCA by researchers can be found here: <https://tisca-llm-app.streamlit.app/>.

CHAPTER

4

FORESTS FOR DIFFERENCES: ROBUST CAUSAL INFERENCE BEYOND PARAMETRIC DID

Abstract

This paper introduces the Difference-in-Differences Bayesian Causal Forest (DiD-BCF), a novel non-parametric model addressing key challenges in DiD estimation, such as staggered adoption and heterogeneous treatment effects. DiD-BCF provides a unified framework for estimating Average (ATE), Group-Average (GATE), and Conditional Average Treatment Effects (CATE). A core innovation, its Parallel Trends Assumption (PTA)-based reparameterization, enhances estimation accuracy and stability in complex panel data settings. Extensive simulations demonstrate DiD-BCF's superior performance over established benchmarks, particularly under non-linearity, selection biases, and effect heterogeneity. Applied to U.S. minimum wage policy, the model uncovers significant conditional treatment effect heterogeneity related to county population, insights obscured by traditional methods. DiD-BCF offers a robust and versatile tool for more nuanced causal inference in modern DiD applications.

4.1 Introduction

Estimating causal effects lies at the heart of scientific inquiry across disciplines such as economics (VARIAN, 2016; HOOVER, 1990; CAUSAL…, 2024), epidemiology (ROTHMAN; GREENLAND, 2005; OHLSSON; KENDLER, 2020; VANDENBROUCKE; BROADBENT; PEARCE, 2016), and social sciences (IMBENS, 2024; GANGL, 2010; GRIMMER, 2014). Yet, the task of empirically identifying causal relationships remains highly challenging, particularly when relying on observational data where confounding

and selection bias are prevalent (TCHETGEN; PARK; RICHARDSON, 2023; HAMMERTON; MUNAFò, 2021; NICHOLS, 2007; FENG *et al.*, 2023; BOYER; DAHABREH; STEINGRIMSSON, 2023; KEOGH; GELOVEN, 2024; DOUTRELIGNE; VAROQUAUX, 2025). The absence of random assignment in non-experimental settings means that simple outcome comparisons between treated and untreated groups are often misleading, since these groups may differ systematically in both observed and unobserved characteristics.

To address these limitations, researchers employ quasi-experimental methods that build strong identification assumptions into their designs. Among such tools, matching (STUART, 2010; KING *et al.*, 2011; KALLUS, 2020; IMAI; KIM; WANG, 2021), regression discontinuity (BOR *et al.*, 2014; BOR; MOSCOE; BäRNIGHAUSEN, 2015; CATTANEO; TITIUNIK, 2022; LINDEN; ADAMS, 2012; OLDENBURG; MOSCOE; BäRNIGHAUSEN, 2016), instrumental variables strategies (ANGRIST; IMBENS; RUBIN, 1996; HERNáN; ROBINS, 2006; TAN, 2006; BOWDEN *et al.*, 2021; BAIOCCHI; CHENG; SMALL, 2014), and, notably, Difference-in-Differences (DiD) are widely used to emulate experimental settings as closely as possible.

DiD in particular has become a workhorse approach for estimating causal impacts of time series relational data, like discrete shocks, policy changes, or interventions (ROTH, 2024; ROTH, 2022; FREYALDENHOVEN *et al.*, 2021; ATHEY; IMBENS, 2022; CALLAWAY; SANT'ANNA, 2021; GARDNER, 2022; SANT'ANNA; ZHAO, 2020). By leveraging panel or repeated cross-sectional data for treated and control groups observed before and after an intervention, DiD seeks to construct the counterfactual evolution of outcomes for the treated group. Its historical roots reach back to John Snow's study of the London cholera outbreak (CANIGLIA; MURRAY, 2020), and today, DiD is widely applied to evaluate the effects of regulatory changes, policy implementations, and social programs (LEER, 2016; SALINAS; SOLé-OLLé, 2018; LOWENSTEIN *et al.*, 2019; YEON *et al.*, 2020; GROENIGER *et al.*, 2021; ZHOU *et al.*, 2021).

Despite its popularity and intuitive appeal, the credibility of DiD hinges crucially on the validity of a set of key identification assumptions. Foremost among these is the *parallel trends assumption* (PTA): in the absence of treatment, the average temporal trend in outcomes for treated and control groups would have been the same (ROTH, 2024; ROTH, 2022; FREYALDENHOVEN *et al.*, 2021). Formally, if $Y_{it}(0)$ denotes the potential outcome for unit i at time t without treatment, the PTA can be expressed as:

$$E\left[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 1\right] = E\left[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 0\right]$$

for all relevant t, where D_i indicates treatment assignment (i.e., if the i-th observation is in the treated group or control group). This condition guarantees that, absent the intervention, treated and control units would have experienced equivalent changes in the outcome variable across time.

The canonical two-way fixed effects (TWFE) DiD regression model implements this logic empirically:

$$Y_{it} = \alpha + \eta D_i + \theta_t + \tau D_{it} + \varepsilon_{it},$$

where Y_{it} is the observed outcome, η and θ_t denote unit and time fixed effects, D_{it} is a treatment indicator (equal to 1 for treated units post-intervention, 0 otherwise), τ is the treatment effect, ε_{it} is the gaussian error term with mean 0, and α is interpreted as the average treatment effect on the treated (ATT) under the aforementioned PTA.

In addition to the parallel trends assumption, the DiD framework has a few other assumptions, from which most are shared common to causal inference methods more broadly:

- **Stable Unit Treatment Value Assumption (SUTVA)**: There is no interference between units (the potential outcome of each unit depends solely on its own treatment status), and there is only one version of treatment and control.
- **No Anticipation**: Units do not change their behavior in anticipation of future treatment; potential outcomes prior to treatment are unaffected by eventual treatment assignment.
- **Ignorability/Unconfoundedness of Treatment Timing**: Conditional on observed (and in basic DiD, time-invariant) characteristics and group/time assignment, treatment is as good as randomly assigned with respect to potential outcome trends.
- **Consistency**: The observed outcome equals the potential outcome under the treatment actually received.

While the classical PTA is "unconditional," i.e., assumes comparability of outcome trends across groups as a whole, it is often empirically more credible and theoretically flexible to impose a *conditional parallel trends* assumption, allowing outcome trends to be similar only *within* strata of observed covariates $\mathbf{X_{i,t}}$. This generalization parallels the conditional ignorability assumption familiar from matching methods and propensity score-based estimators:

$$E\left[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 1, \mathbf{X_{i,t}}\right] = E\left[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 0, \mathbf{X_{i,t}}\right] \quad \forall \mathbf{X_{i,t}}, t.$$

That is, after conditioning on relevant observed covariates $\mathbf{X_{i,t}}$, the evolution of untreated potential outcomes is assumed to be the same for treated and control units. This extension is particularly valuable in settings where assignment to treatment is related

to observable characteristics. Many recent DiD estimators—notably, those employing regression adjustment, inverse probability weighting, or doubly robust machine learning methods—explicitly rely on or estimate effects under a conditional parallel trends framework (CALLAWAY; SANT'ANNA, 2021; SANT'ANNA; ZHAO, 2020).

Since the PTA is inherently untestable (as it refers to a counterfactual), empirical practice typically probes its plausibility by examining *pre-treatment trends* in the outcome variable, e.g., through event-study regression estimates and associated pre-trend ("placebo") coefficients (ROTH, 2022). The most common check tests the (non-)significance of these pre-treatment coefficients (i.e., $\tau_k \ \forall k < 0$) interpreting significant pre-treatment effects as evidence against PTA:

$$Y_{it} = \alpha + \eta D_i + \theta_t + \mathbf{X_{it}}'\boldsymbol{\gamma} + \sum_{k \neq 0} \tau_k D_i + \varepsilon_{it},$$

where $\mathbf{X_{it}}'\boldsymbol{\gamma}$ captures the contribution of the observed covariates, and $k = 0$ being the exact period where treatment is received. However, this common pre-testing approach is known to suffer from low power, vulnerability to multiple hypotheses statistical testing issues, and commonly introduces distortions in estimation and inference if used as a selection criterion (ROTH, 2022).

While the original DiD setup assumed simultaneous adoption of the intervention by a single treated group, modern empirical settings typically feature richer panel data structures. Observational units may be treated at different times (*staggered adoption*), in varying intensities, or there may be multiple control groups. Extending DiD to such non-classical settings raises new identification issues and has revealed critical shortcomings of the TWFE regression estimator. Under staggered adoption and heterogeneous treatment effects, the TWFE parameter α can become a non-convex, and sometimes negatively weighted, average of many possible group-by-time comparisons—leading to potentially biased, misleading, or even sign-reversed estimates relative to the true average causal effect (GOODMAN-BACON, 2021).

A rapidly growing literature has developed improved estimators that directly address these challenges. Methods from Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Chaisemartin and D'Haultfœuille (2020), among others, define and estimate group-time or cohort-specific average treatment effects using appropriately selected control groups and robust aggregation schemes, thereby avoiding the "bad comparisons" inherent in the conventional TWFE approach. Additional methodological advances, such as synthetic DiD (ARKHANGELSKY *et al.*, 2021), further bolster the credibility of causal inference when treatment timing is non-uniform and effects are heterogeneous across subgroups or over time (ARKHANGELSKY *et al.*, 2021).

Most of these methodological advances have focused on estimating overall ATT or, in staggered settings, group-by-time average effects. Yet, in many substantive applications,

it is of central interest to uncover and characterize *heterogeneity* in treatment effects across observable characteristics. This motivates estimation of the Conditional Average Treatment Effect on the Treated (CATT): the average treatment effect for the treated, conditional on baseline covariates. Identifying such heterogeneity can shed light on effect mechanisms and improve the targeting of policy interventions (CINTRON *et al.*, 2022; REHILL; BIDDLE, 2023; HITSCH; MISRA; ZHANG, 2024).

Recent years have seen a surge in methods to estimate heterogeneous treatment effects in DiD settings, notably leveraging advances in machine learning (ATHEY; WAGER, 2019; KATTENBERG; SCHEER; THIEL, 2023; HATAMYAR *et al.*, 2023). Among these, Causal Forests and other flexible estimators, when adapted to account for fixed effects and staggered adoption, enable the recovery of dynamic, covariate-specific treatment effects post-intervention. These approaches marry the robustness of modern DiD identification strategies with the flexibility of machine learning tools, allowing nuanced exploration of treatment effect heterogeneity in complex policy environments.

Building on these foundations, this paper introduces a novel Bayesian machine learning approach for causal inference within the DiD framework that achieves unified and flexible estimation of ATT, group-average treatment effects for units treated at the same time (GATT), and CATT, accommodating both traditional non-staggered and modern staggered adoption settings. Our method generalizes the Bayesian Causal Forest (BCF) model (HAHN; MURRAY; CARVALHO, 2020) to panel data and DiD designs, enabling robust recovery of heterogeneous and dynamic treatment effects while flexibly modeling unit, time, and covariate interactions. Furthermore, we develop an innovative bias-correction term that exploits the PTA, improving the accuracy of posterior treatment effect estimates.

Our proposed model, coined DiD-BCF model, delivers a unified and practical Bayesian framework for causal effect estimation across a wide range of modern policy evaluation settings. Extensive simulation studies demonstrate the notable advantages of our approach relative to existing DiD estimators.

In the remainder of this paper, Section 4.2 presents a brief summary of the current state of DiD estimators literature, while Section 4.3 explains the proposed model of this paper. The Monte Carlo simulation studies are explained in Section 4.4 and their results are discussed in Section 4.5. On the other hand, Section 4.6 illustrates the applicability of DiD-BCF by exploring the salient policy question of minimum wage effects on teen employment. Finally, Section 4.7 concludes the paper.

4.2 Related Work

A number of robust approaches have been advanced to address staggered rollout and heterogeneous effects. Callaway and Sant'Anna (2021) propose estimating group-time average treatment effects $ATT(g,t)$, defined as the average post-treatment effect for units first treated in time g observed at time t. Identification hinges on a conditional parallel trends assumption, either with never-treated or not-yet-treated units as the comparison group, and can flexibly incorporate covariate adjustment.

In their paper, Callaway and Sant'Anna (2021) devise three estimation strategies for $ATT(g,t)$, namely Outcome Regression (OR), Inverse Probability Weighting (IPW), and Doubly Robust (DR). By exploiting these estimation methods, aggregated estimands (by event time, group, or overall) can then be produced, with bootstrap methods providing valid inference.

Gardner (2022), in his "Two-stage differences in differences" paper and R library did2s (BUTTS *et al.*, 2023), offers an alternative framework to address the limitations of standard DiD regressions with staggered adoption and dynamic treatment effects. His approach involves a two-step estimation procedure. In the first stage, group and period fixed effects are estimated using only the subsample of untreated observations. The intuition is that these observations cleanly identify the baseline additive outcome structure under the PTA. In the second stage, these estimated group and period effects are subtracted from the observed outcomes for all units (both treated and untreated). The resulting "adjusted" outcomes are then regressed on the treatment status indicator. Gardner (2022) shows this two-stage method identifies the overall ATT and is robust to treatment effect heterogeneity across groups and time. The method is also presented as intuitive, easy to implement, and extendable to event-study analyses and various other average treatment effect measures.

Chaisemartin and D'Haultfœuille (2020), on the other hand, introduce estimators based on "switchers" (units changing treatment status between periods). Their DID_M estimator targets the average treatment effect among units whose status changes, relying on tailored trend assumptions among switchers. It avoids the negative weighting issues of TWFE and has extensions to "fuzzy" designs (incomplete or probabilistic assignment), using Wald ratios of DiDs.

Also exploring a different path, Arkhangelsky *et al.* (2021) synthesize ideas from synthetic control and DiD. Their Synthetic Difference-in-Differences (SDID) estimator constructs both unit and time weights in pre-treatment periods and combines these with traditional DiD adjustment, yielding a doubly robust estimator that is particularly credible when only a few units adopt treatment and standard parallel trends is suspect.

Moving to (dynamic) heterogeneous effects based on covariates, Kattenberg, Scheer

and Thiel (2023) introduce the Causal Forest with Fixed Effects (CFFE) using within-unit transformations to partial out fixed effects prior to estimating heterogeneity. Their method is in broad terms an adaptation of Causal Forests (ATHEY; WAGER, 2019) to DiD settings with fixed effects and staggered adoption. Similarly, Hatamyar *et al.* (2023) propose the MLDID estimator, combining machine learning (e.g., random forests, BART, etc) with the group-time DiD identification of Callaway and Sant'Anna (2021). MLDID deploys machine learning to flexibly estimate both propensity scores (treatment adoption probabilities) and counterfactual outcome means, and then integrates these via a doubly robust framework to recover dynamic, covariate-dependent CATT trajectories for each treated unit post-adoption.

Across these developments, research has progressively advanced from simple two-period/two-group settings to flexible frameworks that allow for staggered interventions, effect heterogeneity, covariate conditioning, and sophisticated machine learning estimation. As a result, contemporary practice in DiD is equipped to address the methodological challenges encountered in real-world policy analysis, and recent innovations continue to broaden the empirical applicability and interpretability of DiD-based causal inference, which is exactly what this paper aims to do by proposing the DiD-BCF model.

4.3 DiD-BCF Model

Generalizing the DiD Framework

The traditional DiD framework, while powerful, often relies on restrictive linear and additive assumptions. Consider the canonical dynamic DiD model with covariates shown in Section 4.1:

$$Y_{it} = \alpha + \eta D_i + \theta_t + \mathbf{X_{it}}'\boldsymbol{\gamma} + \sum_{k\neq 0} \tau_k D_i + \varepsilon_{it},$$

This model can be generalized by allowing for more flexible functional forms. First, the baseline outcome component, $\alpha + \eta D_i + \theta_t + \mathbf{X_{it}}'\boldsymbol{\gamma}$, can be conceived as a general function of unit identity, time, and covariates:

$$\mu(D_i, t, \mathbf{X_{it}})$$

This function $\mu(\cdot)$ captures the expected outcome trajectory for units in the absence of treatment (or for control units), potentially in a highly non-linear and interactive way, while allowing for static differences between the eventually-treated and control groups.

Second, the treatment effect component, $\sum_{k\neq 0} \tau_k D_i$, can also be generalized. Instead of constant τ_k coefficients, we can allow the treatment effect to be a flexible function of covariates $\mathbf{X_{it}}$ and event time k, $\tau(k, \mathbf{X_{it}})$. Additionally, we can even generalize this

modeling to staggered treatment setting if we consider k_{it} instead of k, where k_{it} is the event time for unit i at calendar time t (i.e., $t - G_i$, where G_i is the treatment start time for unit i, and $G_i = \infty$ for never-treated units). In this case the generalized DiD model then becomes:

$$Y_{it} = \mu(D_i, t, \mathbf{X_{it}}, \mathbf{X_i}) + \tau(\mathbf{X_{it}}, k_{it}) \cdot \mathbb{I}(G_i \neq \infty) + \varepsilon_{it}, \tag{4.1}$$

where $\mathbb{I}(G_i \neq \infty)$ is an indicator for unit i being an "ever-treated" unit. For control units ($G_i = \infty$), the $\tau(\cdot)$ term is absent. The function $\tau(\cdot)$ now explicitly models how the treatment effect evolves over event time k and varies across units with different covariate values $\mathbf{X_{it}}$. The key assumption is the additive separability between the baseline function $\mu(\cdot)$ and the treatment effect function $\tau(\cdot)$. Incidentally, $\mu(\cdot)$ and $\tau(\cdot)$ can be identified using one of the nonparametrically point-identified estimants proposed in the literature; e.g., the outcome regression approach of Heckman, Ichimura and Todd (1997), Heckman *et al.* (1998), the inverse probability weighting (IPW) approach of Abadie (2005), or the doubly robust (DR) of Sant'Anna and Zhao (2020).

Additionally, by exploiting the PTA, we can reparametrize Equation 4.1 as:

$$Y_{it} = \mu(D_i, t, \mathbf{X_{it}}, \mathbf{X_i}) + \tau(\mathbf{X_{it}}, k_{it}) \cdot \mathbb{I}(G_i \neq \infty) \cdot \mathbb{I}(k_{it} \geq 0) + \varepsilon_{it}, \tag{4.2}$$

Such a reparametrization decreases the complexity of causal regression task. This reduction in complexity can be understood by comparing the functional requirements placed upon the treatment effect estimator in different model specifications. Consider the formulation in Equation 4.1 and let μ_1 and τ_1 be the functions of this formulation; for this model to be consistent with the definition of k_{it} (time relative to treatment) and the PTA, the function $\tau_1 : \mathscr{X} \times \mathscr{K} \to \mathbb{R}$ must intrinsically satisfy the constraint:

$$\tau_1(\mathbf{X}, k) = 0 \quad \forall \mathbf{X} \in \mathscr{X}, \forall k \in \mathscr{K}_{<0} \tag{4.3}$$

where $\mathscr{X}$ is the covariate space, $\mathscr{K}$ is the space of relative time periods k_{it}, and $\mathscr{K}_{<0} = \{k \in \mathscr{K} \mid k < 0\}$. The estimation process for τ_1, typically involving flexible/nonparametric methods, must therefore learn a function that not only captures the potentially complex relationship between $\mathbf{X}_{it}$ and k_{it}, and the treatment effect magnitude for $k_{it} \geq 0$, but also perfectly adheres to the zero constraint (4.3) for $k_{it} < 0$. Let $\mathscr{F}_1$ denote the function class embodying this constraint. Estimating a function within $\mathscr{F}_1$ imposes a significant burden (HOROWITZ, 2009; DAI, 2024; NEWEY, 1990; KLAASSEN; PUTTER, 2005), requiring the model to capture a potentially sharp discontinuity or behavioral change precisely at $k = 0$ (DEB; MUKHERJEE, 2024).

Now, consider the alternative specification in Equation 4.2. Here, the treatment effect term, denoted as $\tau_2(\mathbf{X}_{it}, k_{it}) \cdot D_{it}$ (it is trivial to show that $\mathbb{I}(G_i \neq \infty) \cdot \mathbb{I}(k_{it} \geq 0) = D_{it}$), is structurally zero whenever $D_{it} = 0$, which includes all pre-treatment periods ($k_{it} < 0$).

This occurs by construction due to the multiplication by D_{it}, irrespective of the value of $\tau_2(\mathbf{X}_{it}, k_{it})$. Consequently, the function $\tau_2 : \mathscr{X} \times \mathscr{K} \to \mathbb{R}$ is not required to intrinsically satisfy the zero constraint for $k_{it} < 0$. Let $\mathscr{F}_2$ be the corresponding function class for τ_2, which does not necessarily impose the constraint (4.3). The estimation task for τ_2 effectively concentrates on the domain where $D_{it} = 1$, i.e., the post-treatment periods ($\mathscr{K}_{\geq 0} = \{k \in \mathscr{K} \mid k \geq 0\}$) (DEB; MUKHERJEE, 2024). The complexity of the target function τ_2 is reduced because it only needs to model the effect conditional on treatment activation, $\tau_2(\mathbf{X}, k | k \geq 0)$, rather than simultaneously modeling both the post-treatment effect and the pre-treatment null effect (DEB; MUKHERJEE, 2024; KLAASSEN; PUTTER, 2005; FRIEDRICH *et al.*, 2022).

This simplification of the target function's required behavior implies a reduction in the complexity of the estimation task itself (DEB; MUKHERJEE, 2024; KLAASSEN; PUTTER, 2005; FRIEDRICH *et al.*, 2022). Learning a function in $\mathscr{F}_1$, which must exhibit specific behavior (identically zero) on $\mathscr{K}_{<0}$ and potentially complex behavior on $\mathscr{K}_{\geq 0}$, is arguably more challenging than learning a function in $\mathscr{F}_2$ whose behavior on $\mathscr{K}_{<0}$ is rendered irrelevant by the structural zero $D_{it} = 0$ (DEB; MUKHERJEE, 2024). From the perspective of non-parametric estimation (e.g., minimizing a loss function $\mathscr{L}$ like $\sum(Y_{it} - \hat{Y}_{it})^2$ subject to regularization), the optimization problem associated with Equation 4.2 is presumably simpler or possess a more well-behaved solution space (DEB; MUKHERJEE, 2024; FRIEDRICH *et al.*, 2022). The requirement for $\hat{\tau}_1$ to sharply transition to zero at $k = 0$ might demand significant model capacity or lead to instability, particularly near the boundary (DEB; MUKHERJEE, 2024).

The reduced complexity associated with estimating τ_2 can facilitate better model convergence and stability (KLAASSEN; PUTTER, 2005; FRIEDRICH *et al.*, 2022). By removing the need for τ_2 to explicitly model the null pre-treatment period, the estimation algorithm can dedicate its resources more efficiently to capturing the potentially intricate variations in treatment effects conditional on covariates $\mathbf{X}_{it}$ and post-treatment time $k_{it} \geq 0$.

Consequently, for the devise of our proposed model, DiD-BCF, we choose to use the formulation in Equation 4.2 given the aforementioned advantages of this formulation and the fact that using the DiD framework only makes sense under the PTA, albeit there is some work on the modification of the DiD framework for cases when the PTA is not respected (RAMBACHAN; ROTH, 2023).

Estimating Flexible DiD Models with Bayesian Causal Forests

To estimate the flexible functions $\mu(\cdot)$ and $\tau(\cdot)$ in our generalized DiD model, we turn to non-parametric Bayesian methods, particularly Bayesian Additive Regression Trees (BART) and its extension, Bayesian Causal Forests (BCF).

Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees (BART), introduced by Chipman, George and McCulloch (2010), is a non-parametric method that models an unknown function $f(x) = E(Y \mid X = x)$ as a sum of multiple regression trees. The BART model for an observation i is typically expressed as:

$$Y_i = \sum_{l=1}^{M} g_l(\mathbf{X_i}; T_l, M_l) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Here, Y_j is the response, $\mathbf{X_i}$ is a vector of predictors. Each $g_l(\mathbf{X_i}; T_l, M_l)$ represents a single regression tree, where T_l defines the tree's structure (splitting rules) and M_l contains the parameter values at the terminal nodes (leaves) of that tree. The error term ε_i is assumed to be normally distributed, with σ^2 being "bayesianly" modeled with the data. The key idea is that each tree is a "weak learner," capturing a small part of the overall function $f(\mathbf{X})$, and their sum provides a robust and flexible fit. Incidentally, the default value for M is 200 (CHIPMAN; GEORGE; MCCULLOCH, 2010) (default here is meant to be the value that Chipman, George and McCulloch (2010) recommended for a general use of the model, that is, without hyperparameter optimization).

BART employs sophisticated prior distributions for the tree structures (T_l), terminal node parameters (M_l), and the error variance (σ^2). These priors are crucial for regularization, preventing overfitting and ensuring good predictive performance. Specifically, the tree prior $p(T_l)$ penalizes overly complex trees by making deeper nodes less likely to split, using a rule like $\eta(1+d)^{-\beta}$, where d is the node depth, $\eta \in (0,1)$, and $\beta \geq 0$, with default values for these priors being $\eta = 0.95$ and $\beta = 2$. Priors also govern the choice of splitting variables and split points. For more details on BART priors and its Bayesian backfitting Markov chain Monte Carlo (MCMC) estimation algorithm, see Chipman, George and McCulloch (2010).

Bayesian Causal Forests (BCF)

While BART can be used for causal inference by including the treatment indicator as a predictor in $\mathbf{X_i}$ (an "S-learner" approach, per Künzel *et al.* (2019)), this can obscure the treatment effect and its heterogeneity (CURTH; SCHAAR, 2021). Another approach is to fit a BART model for the observations with the treatment indicator being one and another BART model for the observations with the treatment indicator being zero and taking their difference as:

$$\hat{\tau}(\mathbf{X_i}) = \hat{Y}_i(1) - \hat{Y}_i(0).$$

where $\hat{\tau}(\mathbf{X_i})$ would be the estimated conditional treatment effect and $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ are the fitted BART models. Such a procedure is named T-learner approach (KüNZEL

et al., 2019). Nonetheless, the issue with the two-model T-learner approach is that it inherently applies less regularization to the treatment effect compared to each individual model (HAHN; MURRAY; CARVALHO, 2020; CURTH; SCHAAR, 2021). This is counterintuitive in many scenarios where treatment effects are anticipated to be minimal (ALCANTARA *et al.*, 2024). As a result, Hahn, Murray and Carvalho (2020) developed Bayesian Causal Forests (BCF) to directly model heterogeneous treatment effects. BCF reparameterizes the response function:

$$Y_i = \mu(\mathbf{X_i}, \hat{\pi}(\mathbf{X_i})) + \tau(\mathbf{X_i})D_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

In this formulation, D_i is the binary treatment indicator for unit i. $\mu(\cdot)$ is the "prognostic function," representing the baseline outcome. $\tau(\cdot)$ is the "treatment effect function", capturing how the treatment effect varies with $\mathbf{X_i}$. $\hat{\pi}(\mathbf{X_i})$, on the other hand, is the estimated propensity score (i.e., the probability of receiving the treatment given covariates), which is used in $\mu(\cdot)$ to allegedly mitigate the regularization-induced confounding (RIC) bias (HAHN *et al.*, 2018) when target treatment selection (i.e., $\pi(\cdot)$ depends on $\mu(\cdot)$) is present (HAHN; MURRAY; CARVALHO, 2020), albeit Souto and Louzada (2024) demonstrate that the inclusion is actually not necessary even in target treatment selection settings.

Both $\mu(\cdot)$ and $\tau(\cdot)$ are themselves modeled as sums of BART trees. Yet, it is worth mentioning that while the BART priors and default hyperparameters of $\mu(\cdot)$ are the same as recommend by Chipman, George and McCulloch (2010) (with one small modification made Hahn, Murray and Carvalho (2020), which places a half-Cauchy prior over the scale of the leaf parameters with prior median equal to twice the marginal standard deviation of Y), $\tau(\cdot)$ employes a stronger regularization, with 50 trees, $\beta = 3$, $\eta = 0.25$, and a half Normal prior over the scale of $\tau(\mathbf{X_i})$, pegging the prior median to the marginal standard deviation of Y (HAHN; MURRAY; CARVALHO, 2020).

This structure allows for separate regularization of the prognostic and treatment effect components, which is often desirable as treatment effects might be smoother or simpler than baseline outcome functions (HAHN; MURRAY; CARVALHO, 2020). (The original BCF model includes scaling parameters b_0, b_1 for $\tau(\mathbf{X_i})D_i$, which we omit here for simplicity but are part of the full specification, whose details can be found in Hahn, Murray and Carvalho (2020)).

Warm-Start Bayesian Causal Forests (ws-BCF)

Estimation in BART and BCF traditionally relies on backfitting MCMC algorithms, which can be computationally intensive and slow to converge (HE; HAHN, 2021), especially with large datasets, due to highly correlated tree samples (ALCANTARA *et al.*, 2024). He and Hahn (2021) proposed XBART (Accelerated BART), which uses a more efficient "Grow-From-Root" stochastic tree-fitting algorithm, described in Algorithm 2. This algorithm explores the tree space more rapidly.

Algorithm 1 – GrowFromRoot (as described in Alcantara *et al.* (2024))

procedure GrowFromRoot($y, \mathbf{X}, \Phi, \Psi, d, T,$ node) ▷ Modifies T by adding nodes and sampling associated leaf parameters μ.
  if the stopping conditions are met for node **then**
    $\mu_{\text{node}} \leftarrow$ SampleParameters($\emptyset$) ▷ Node becomes a leaf, update parameter
    return
  end if
  $s^{\emptyset} \leftarrow s(y, \mathbf{X}, \Psi, \mathscr{C}, \text{all})$
  for $c_{jk} \in \mathscr{C}$ **do**
    $s_{jk}^{(1)} \leftarrow s(y, \mathbf{X}, \Psi, \mathscr{C}, j, k, \text{left})$
    $s_{jk}^{(2)} \leftarrow s(y, \mathbf{X}, \Psi, \mathscr{C}, j, k, \text{right})$
    Calculate $L(c_{jk}) = m\left(s_{jk}^{(1)}; \Phi, \Psi\right) \times m\left(s_{jk}^{(2)}; \Phi, \Psi\right)$
  end for
  Calculate $L(\emptyset) = |\mathscr{C}|\left(\eta(1+d)^{-\beta} - 1\right) m\left(s^{\emptyset}; \Phi, \Psi\right)$
  Sample a cutpoint c_{jk}^* (from $\mathscr{C}$) or the null cutpoint $\emptyset$ with probabilities:
    $P(c_{jk}) = \frac{L(c_{jk})}{\sum_{c'_{j'k'} \in \mathscr{C}} L(c'_{j'k'}) + L(\emptyset)}$ for $c_{jk} \in \mathscr{C}$
    $P(\emptyset) = \frac{L(\emptyset)}{\sum_{c'_{j'k'} \in \mathscr{C}} L(c'_{j'k'}) + L(\emptyset)}$
  if the null cutpoint $\emptyset$ is selected (i.e., $c_{jk}^* = \emptyset$) **then**
    $\mu_{\text{node}} \leftarrow$ SampleParameters($\emptyset$)
    return
  else ▷ A non-null cutpoint c_{jk}^* was selected
    Create two new child nodes, **left_node** and **right_node**, for **node** in T
    Partition data $(y, \mathbf{X})$ at **node** into $(y_{\text{left}}, \mathbf{X}_{\text{left}})$ and $(y_{\text{right}}, \mathbf{X}_{\text{right}})$:
      $(y_{\text{left}}, \mathbf{X}_{\text{left}})$: data where $x_{ij'} \leq x_{jk}^{**}$
      $(y_{\text{right}}, \mathbf{X}_{\text{right}})$: data where $x_{ij'} > x_{jk}^{**}$
      (where x_{jk}^{**} is the value corresponding to the sampled cutpoint c_{jk}^*)
    Call GrowFromRoot($y_{\text{left}}, \mathbf{X}_{\text{left}}, \Phi, \Psi, d+1, T,$ **left_node**)
    Call GrowFromRoot($y_{\text{right}}, \mathbf{X}_{\text{right}}, \Phi, \Psi, d+1, T,$ **right_node**)
  end if
end procedure

Krantsevich, He and Hahn (2023) extended this to BCF, creating the XBCF algorithm. XBCF essentially applies the XBART fitting approach to the BCF model, often with a slight modification allowing for heteroskedastic errors by treatment status:

$$Y_i = a\mu(\mathbf{X_i}, \hat{\pi}(\mathbf{X_i})) + b_{D_i} D_i \tilde{\tau}(\mathbf{X_i}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$a \sim N(0,1), \quad b_0, b_1 \sim N(0, 1/2),$$

where $\mu(x)$ and $\tilde{\tau}(x)$ are XBART forests, the actual treatment effect is $\tau(x) = (b_1 - b_0)\tilde{\tau}(x)$ from the full parameterization, and a is an additional scaling factor, which enhances the learning of the prognostic term (KRANTSEVICH; HE; HAHN, 2023). The Grow-From-Root algorithm stochastically grows trees, offering faster posterior exploration. In our DiD

context, we can use XBCF to obtain efficient initial estimates (a "warm start") for the more standard BCF MCMC, potentially speeding up convergence to the target posterior, which has been proposed by Krantsevich, He and Hahn (2023).

For more information about the Grow-From-Root algorithm and warm-start BCF, please see He and Hahn (2021) and Krantsevich, He and Hahn (2023) respectively.

The DiD-BCF Model and PTA-Based Debiasing

We can now specify the ws-BCF model for our generalized DiD framework from Eq. (4.2) as:

$$Y_{it} = \mu(D_i, t, \mathbf{X}_{it}) + \tau(\mathbf{X}_{it}, k_{it}) \cdot D_{it} + \varepsilon_{it} \tag{4.4}$$

It is worth mentioning that we choose to not use $\hat{\pi}(\mathbf{X}_i, G_i)$ given its lack of importance for the BCF's model performance as shown in the ablation studies of Souto and Louzada (2024).

4.4 Simulation Studies Design

To evaluate the performance of our proposed estimator within the DiD framework, we conduct extensive simulation studies. These simulations rely on data generated from five distinct Data Generating Processes (DGPs), designed to mirror diverse empirical scenarios. Each DGP is examined under four different settings to probe the estimator's robustness to model misspecification (particularly regarding linearity assumptions) and dynamic treatment effects. The DGPs vary in treatment assignment (simultaneous vs. staggered), treatment effect heterogeneity (homogeneous vs. conditional), and selection into treatment (random vs. propensity score-based).

4.4.1 Benchmark Models

The benchmark models for our simulation studies were chosen based on two criteria: 1. their relevance in the literature and 2. code availability.

The first and presumably most straight-forward benchmark model is the TFWE with dynamic treatment effect and covariates:

$$Y_{it} = \alpha + \eta D_i + \theta_t + \mathbf{X_{it}}'\boldsymbol{\gamma} + \sum_k \tau_k D_i + \varepsilon_{it},$$

Its implementation in this study was performed in the programming language Python, using the statsmodels library. Another straight-forward benchmark models are

the models proposed by Sant'Anna and Zhao (2020) and Gardner (2022). Sant'Anna and Zhao (2020) model can be defined as:

$$Y_{it} = \alpha_{G_i} + \eta_{G_i} G^g \cdot \mathbb{I}(G_i \neq \infty) + \theta_t \cdot \mathbb{I}(G_i \neq \infty) + \mathbf{X_{it}}'\boldsymbol{\gamma} + \tau_{G_i,t} G^g \cdot \mathbb{I}(T = t) \cdot \mathbb{I}(G_i \neq \infty) + \varepsilon_{it}$$

where $G^g = 1$ if G_i of Y_{it} equals g. The parameters for this model are estimated using the DR approach as recommended by Sant'Anna and Zhao (2020) and the R package did is used (CALLAWAY; SANT'ANNA, 2022). Onwards, we refer to this model as DiD DR

Moving to Gardner (2022) model, the two-stage procedure is as follows:

1. **First Stage: Estimate Fixed Effects from Untreated Observations.** Estimate the group fixed effects and period fixed effects by regressing outcomes on group and period indicators using only the subsample of *yet untreated* observations (where $D_{it} = 0$):

$$Y_{it} = \alpha_{G_i} + \eta_{G_i} G^g \cdot \mathbb{I}(G_i \neq \infty) + \theta_t + \mathbf{X_{it}}'\boldsymbol{\gamma} + \varepsilon_{it}^{\text{first stage}} \quad \text{for observations where } D_{it} = 0$$

2. **Second Stage: Estimate ATT from Adjusted Outcomes.** Construct an adjusted outcome variable by subtracting the estimated fixed effects from the observed outcome for all observations: $\tilde{Y}_{it} = Y_{it} - \hat{\alpha}_{G_i} - \hat{\eta}_{G_i} G^g \cdot \mathbb{I}(G_i \neq \infty) - \hat{\theta}_t - \mathbf{X_{it}}'\hat{\boldsymbol{\gamma}}$. Then, estimate the GATT $\tau_{G_i,t}$, by regressing the adjusted outcome on the treatment indicator using the full sample:

$$\tilde{Y}_{it} = \tau_{G_i,t} G^g \cdot \mathbb{I}(T = t) \cdot \mathbb{I}(G_i \neq \infty) + \varepsilon_{it}^{\text{second stage}}$$

Gardner (2022) model parameters is estimated using a joint generalized method of moments and its code can be found in the R library did2s (BUTTS *et al.*, 2023). Onwards, we refer to this model as DiD2s.

Moving to a perhaps less well-known benchmark model, we have the SDID model of Arkhangelsky *et al.* (2021). Let $\tilde{Y}_{it}$ here be $\tilde{Y}_{it} = Y_{it} - \mathbf{X_{it}}'\hat{\boldsymbol{\gamma}}$, $\mathcal{N}_C$ be the set of control units and $\mathcal{N}_T$ be the set of treated units. Let T_{pre} be the set of pre-treatment periods and T_{post} be the set of post-treatment periods. The number of treated units is $N_T = |\mathcal{N}_T|$, control units $N_C = |\mathcal{N}_C|$, pre-treatment periods $T_{pre}^{num} = |T_{pre}|$, and post-treatment periods $T_{post}^{num} = |T_{post}|$.

The SDID estimate of the ATT, $\hat{\tau}_{sdid}$, is obtained from a weighted TWFE regression :

$$(\hat{\tau}_{sdid}, \hat{\alpha}, \hat{\eta}, \hat{\theta}_t) = \arg\min_{\tau,\alpha,\eta,\theta_t} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{Y}_{it} - \alpha - \eta D_i - \theta_t - D_{it}\beta_{ij})^2 \omega_i^{sdid} \lambda_t^{sdid} \tag{4.5}$$

The crucial components are the unit weights ω_i^{sdid} and time weights λ_t^{sdid}.

The unit weights ω_i^{sdid} are determined as follows: For control units $i \in \mathcal{N}_C$, the weights $\hat{\omega}_i$ (and an intercept $\hat{\omega}_0$) are chosen to minimize the sum of squared differences between the average outcome of treated units and the weighted average outcome of control units over pre-treatment periods, plus a regularization term:

$$(\hat{\omega}_0, \{\hat{\omega}_i\}_{i\in\mathcal{N}_C}) = \arg\min_{\omega_0, \{\omega_i\}_{i\in\mathcal{N}_C}} \sum_{t\in T_{pre}} \left(\omega_0 + \sum_{i\in\mathcal{N}_C} \omega_i \tilde{Y}_{it} - \frac{1}{N_T} \sum_{j\in\mathcal{N}_T} \tilde{Y}_{jt} \right)^2 + \zeta_\omega \sum_{i\in\mathcal{N}_C} \omega_i^2$$

subject to $\omega_i \geq 0$ for $i \in \mathcal{N}_C$ and $\sum_{i\in\mathcal{N}_C} \omega_i = 1$. For treated units $j \in \mathcal{N}_T$, weights are typically fixed, e.g., $\omega_j^{sdid} = 1/N_T$ for use in Eq. (4.5). The term ζ_ω is a regularization parameter. The intercept $\hat{\omega}_0$ allows for level differences between the synthetic control and the treated group average.

The time weights λ_t^{sdid} are determined similarly: For pre-treatment periods $t \in T_{pre}$, the weights $\hat{\lambda}_t$ (and an intercept $\hat{\lambda}_0$) are chosen to minimize the sum of squared differences between each control unit's average post-treatment outcome and its weighted average pre-treatment outcome, aggregated over control units, plus a regularization term:

$$(\hat{\lambda}_0, \{\hat{\lambda}_t\}_{t\in T_{pre}}) = \arg\min_{\lambda_0, \{\lambda_t\}_{t\in T_{pre}}} \sum_{i\in\mathcal{N}_C} \left(\lambda_0 + \sum_{t\in T_{pre}} \lambda_t \tilde{Y}_{it} - \frac{1}{T_{post}^{num}} \sum_{s\in T_{post}} \tilde{Y}_{is} \right)^2 + \zeta_\lambda \sum_{t\in T_{pre}} \lambda_t^2$$

subject to $\lambda_t \geq 0$ for $t \in T_{pre}$ and $\sum_{t\in T_{pre}} \lambda_t = 1$. For post-treatment periods $s \in T_{post}$, weights are fixed, e.g., $\lambda_s^{sdid} = 1/T_{post}^{num}$ for use in Eq. (4.5). The term ζ_λ is a (typically small) regularization parameter. The intercept $\hat{\lambda}_0$ allows for systematic differences over time for control units.

The final weights ω_i^{sdid} and λ_t^{sdid} used in Eq. (4.5) are thus composed of these estimated weights $(\hat{\omega}_i, \hat{\lambda}_t)$ and the pre-defined fixed weights. For its code in this study, we used the R library "synthdid" created by Arkhangelsky *et al.* (2021).

While the aforementioned benchmark models are based on linear parametric assumptions, we also consider nonparametric benchmark models. The first one considers the DR estimation approach of Sant'Anna and Zhao (2020) using the random forest algorithm (hence, this benchmark model being fully nonparametric) with 500 trees as recommended by Chang (2020). For this model, we use the R package DoubleML (BACH *et al.*, 2022) and we refer to it as DoubleML_did from now on.

The other two nonparametric benchmark models estimate not only ATT or GATT, as the previously mentioned benchmark models, but also the CATT (i.e., $\tau(\mathbf{X}_{it}, k_{it})$). These models are namely the CFFE (KATTENBERG; SCHEER; THIEL, 2023) and MLDID (HATAMYAR *et al.*, 2023) models explored in Section 4.2. Despite our desire and effort to include these benchmark models in our simulation studies, their respective GitHub repositories have open issues in the installation part (as shown in Figure 13 and 14); and

consequently not allowing their use by researchers. As a result, we were not able to add these nonparametric benchmark models to our simulation studies, but plan to do so if the open issues are resolved before the potential publication of this paper.

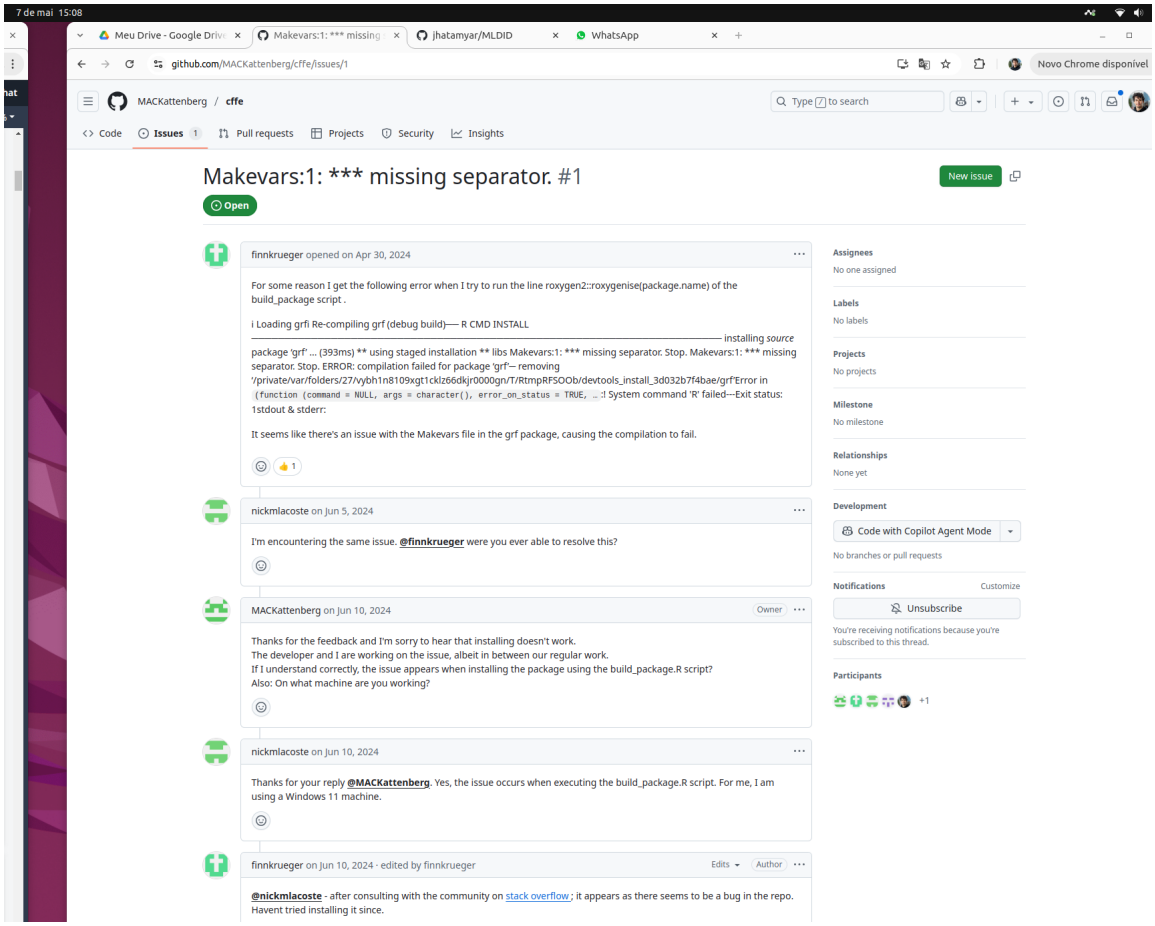

Figure 13 – Issue not allowing the use of Kattenberg, Scheer and Thiel (2023) model

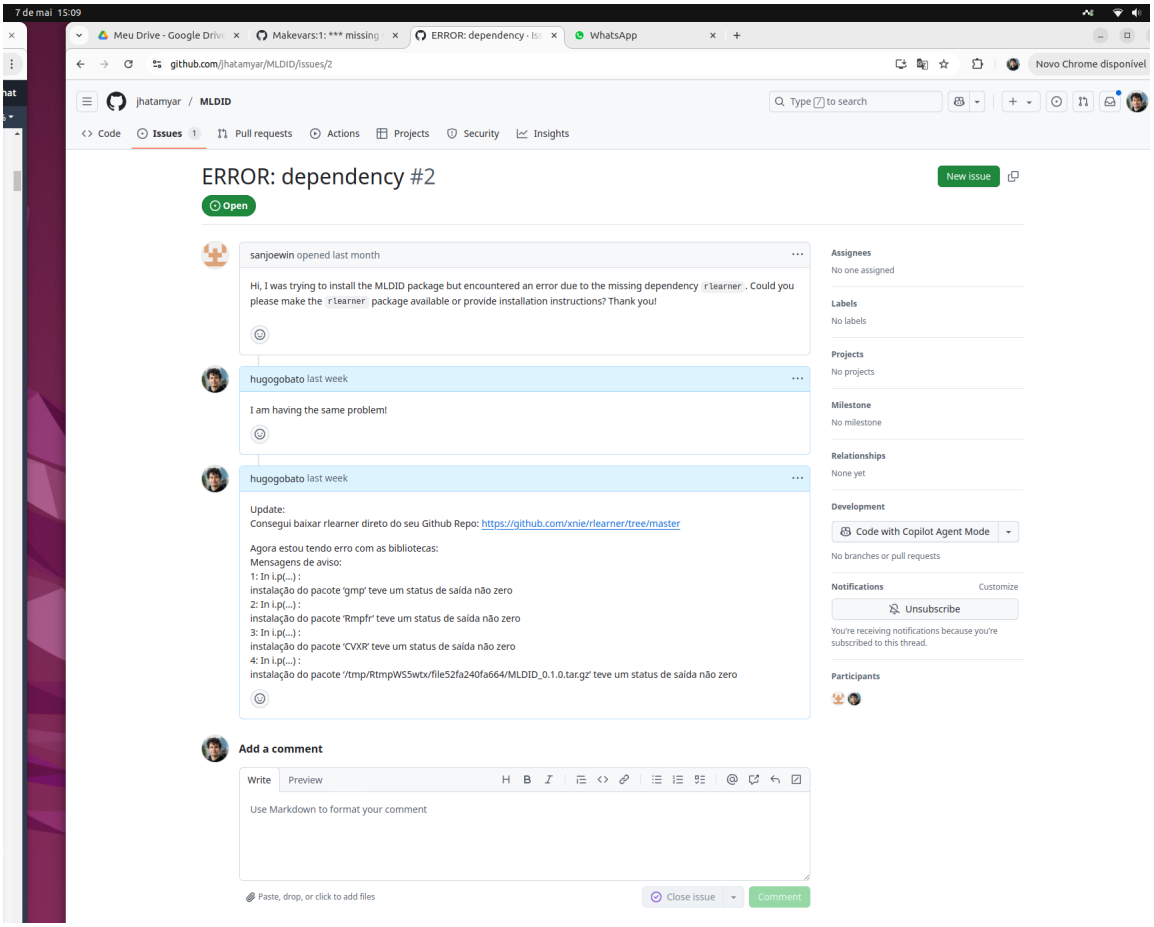

Figure 14 – Issue not allowing the use of Hatamyar *et al.* (2023) model

4.4.2 General Simulation Setup

Across all DGPs and settings, we simulate panel data. Each simulation run involves generating a dataset, applying our proposed estimator and relevant benchmarks, and storing the results. We perform $M = 100$ Monte Carlo repetitions for each DGP configuration to assess the estimators' sampling distributions. In each repetition, a unique random seed, which is equal to the iteration number of the respective iteration, is used for reproducibility.

The basic structure involves potential outcomes $Y_{it}(1)$ (if treated) and $Y_{it}(0)$ (if untreated) for unit i at time t. The observed outcome is $Y_{it} = Y_{it}(0)(1 - D_{it}) + Y_{it}(1)D_{it}$, where D_{it} is the treatment status indicator. The treatment effect is $\tau_{it} = Y_{it}(1) - Y_{it}(0)$. Our DGPs model $Y_{it}(0)$ based on unit fixed effects, time fixed effects, covariates, and an idiosyncratic error term. We assume $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, setting $\sigma_\varepsilon = 1$ in our implementation.

Simulations use $N = 200$ units observed over $T = 8$ time periods, with $T_{\text{pre}} = 4$ pre-treatment periods and $T_{\text{post}} = 4$ post-treatment periods relative to the earliest treatment start time.

The covariate vector $\mathbf{X}_{it}$ consists of $p = 7$ variables:

- One binary covariate $X_{1,it} \sim \text{Bernoulli}(0.66)$.
- Five continuous covariates $X_{2,it}, \ldots, X_{6,it} \sim \mathcal{N}(0,1)$.
- One categorical covariate $X_{7,it}$ taking values $\{1,2,3,4\}$ with probabilities $\{0.3, 0.1, 0.2, 0.4\}$.

These covariates are generated independently for each unit-time observation. The associated coefficient vector used in the linear component of the outcome model is $\boldsymbol{\beta}_x = [-0.75, 0.5, -0.5, -1.30, 1.8, 2.5, -1.0]'$.

For each DGP, we investigate four settings, corresponding to the `linearity_degree` parameter in our code:

1. **Setting 1 (Fully Linear):** $Y_{it}(0)$ depends linearly on covariates and time. The treatment effect is constant. Specifically, $E[Y_{it}] = -0.5 + 0.75 D_i + 0.2t + \mathbf{X_{it}}'\boldsymbol{\gamma} + \tau(\mathbf{X}_{it}, k_{it}) \cdot D_{it}$, where for the DGPs without covariate-heterogeneous treatment effects (CHTE), $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 3 \ \forall k_{it} \geq 0$. The linear DiD benchmark models are correctly specified in this setting (and thus are expected to perform well).

2. **Setting 2 (Partially Non-linear Covariates):** $Y_{it}(0)$ includes non-linear functions for approximately half of the covariates (i.e., $\lfloor p/2 \rfloor$), while the time trend remains linear. Specifically, $E[Y_{it}] = -0.5 + 0.75 D_i + 0.2t + \mathbf{exp}\{\tilde{\mathbf{X}}_{\mathbf{it}}\}'\tilde{\boldsymbol{\gamma}} + \dot{\mathbf{X}}^{\mathbf{2}'}_{\mathbf{it}}\dot{\boldsymbol{\gamma}} + \bar{\mathbf{X}}'_{\mathbf{it}}\bar{\boldsymbol{\gamma}} + \tau(\mathbf{X}_{it}, k_{it}) \cdot D_{it}$, where $\tilde{\mathbf{X}}$ and $\dot{\mathbf{X}}$ are respectively the first and second half of the first half of the covariates and $\bar{\mathbf{X}}_{\mathbf{it}}$ is the other half of the covariates. Anew, for the DGPs without CHTE, $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 3 \ \forall k_{it} \geq 0$. Linear benchmark models are partially misspecified regarding covariate relationships.

3. **Setting 3 (Fully Non-linear Covariates):** $Y_{it}(0)$ incorporates non-linear transformations for all covariates and now the time trend is quadratic. Specifically, $E[Y_{it}] = -0.5 + 0.75 D_i + 0.2t^2 + \mathbf{exp}\{\tilde{\mathbf{X}}_{\mathbf{it}}\}'\tilde{\boldsymbol{\gamma}} + \dot{\mathbf{X}}^{\mathbf{2}'}_{\mathbf{it}}\dot{\boldsymbol{\gamma}} + |\bar{\mathbf{X}}^{\text{first}}_{\mathbf{it}}|'\bar{\boldsymbol{\gamma}}^{\text{first}} + \sqrt{|\bar{\mathbf{X}}^{\text{second}}_{\mathbf{it}}|}'\bar{\boldsymbol{\gamma}}^{\text{second}} + \tau(\mathbf{X}_{it}, k_{it}) \cdot D_{it}$, where $\bar{\mathbf{X}}^{\text{first}}$ and $\bar{\mathbf{X}}^{\text{second}}$ are respectively the first and second half of the second half of the covariates. Now, for the DGPs without CHTE, $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 5 \ \forall k_{it} \geq 0$. Linear benchmark models are significantly misspecified regarding both covariate relationships and the time trend.

4.4.2.1 DGP 1: Canonical DiD with Homogeneous Effects (ATT Focus)

This DGP simulates the classic DiD setting extended to panel data, with two groups (treatment and control) and a single, simultaneous treatment adoption time $t_0 = T_{\text{pre}} = 4$. Treatment status D_i is assigned fully randomly, with $N/2 = 100$ units assigned to the treatment group and $N/2 = 100$ to the control group. The target estimand is the ATT.

For this DGP, $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 3 \ \forall k_{it} \geq 0$ for Setting 1 & 2, and $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 5 \ \forall k_{it} \geq 0$ for Setting 3.

4.4.2.1.1 Real-life Example:

Card and Krueger (1993)

Card and Krueger (1993) conducted one of the most famous difference-in-differences studies, examining the effect of a minimum wage increase in New Jersey on fast-food employment by comparing it with neighboring Pennsylvania (which maintained the same minimum wage). This study perfectly illustrates a canonical DiD setting with two groups and a single treatment adoption time. The study focused on the ATT, comparing outcomes before and after the wage increase across treated (New Jersey) and control (Pennsylvania) restaurants.

4.4.2.2 DGP 2: Staggered Adoption with Homogeneous Effects (GATT Focus)

This DGP models staggered treatment adoption, where different units adopt treatment at different times. Units are randomly fully assigned to one of the three treatment timing groups or a never-treated control group. Needless to say, the focus for this DGP is on estimating GATT.

We divide $N = 200$ units randomly into four groups of approximately equal size ($N/4 = 50$):

- Group 0: Never-treated control group ($G_i = \infty$).
- Group 1: Treatment starts at $t = T_{\text{pre}} = 4$ ($G_i = 4$).
- Group 2: Treatment starts at $t = T_{\text{pre}} + 1 = 5$ ($G_i = 5$).
- Group 3: Treatment starts at $t = T_{\text{pre}} + 2 = 6$ ($G_i = 6$).

Anew, for this DGP, $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 3 \ \forall k_{it} \geq 0$ for Setting 1 & 2, and $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 5 \ \forall k_{it} \geq 0$ for Setting 3. It is worth remembering that k_{it} is the event time for unit i at calendar time t (i.e., $t - G_i$)

4.4.2.2.1 Real-life Example:

Lindrooth *et al.* (2018).

A real-world example for DGP 2 is the analysis of the Affordable Care Act's Medicaid expansion, which different states adopted at different times. Lindrooth *et al.* (2018) have examined how this staggered adoption affected hospital financial stability

across states, with states adopting the expansion at different times (2014, 2015, 2016, and beyond).

DGPs with Selection on Observables (DGPs 3-5)

The following DGPs introduce selection into treatment based on observable covariates (i.e., a propensity score $\pi(\mathbf{X})$), a common challenge in empirical work. To implement this, we expand the covariate set to $p = 8$ variables for these DGPs, including two time-invariant (static) covariates used specifically in the selection mechanism (hence, $\pi(\mathbf{X})_{\text{static}}$), alongside six time-varying (dynamic) covariates. The coefficient vector $\boldsymbol{\beta}_x$ for the outcome model is expanded accordingly to $\boldsymbol{\beta}_x = [-0.75, 0.5, -0.5, -1.30, 1.8, 2.5, -1.0, 0.3]'$.

4.4.2.3 DGP 3: Staggered Adoption with Selection via Utility Maximization (GATT Focus)

This DGP modifies the staggered adoption scenario (DGP 2) by introducing selection on observables. Assignment to treatment timing groups (G_i) depends on predetermined, static unit characteristics via a utility maximization model.

4.4.2.3.1 Implementation Details:

- **Covariates:** $p = 8$.
  - X_{i1}: Static Bernoulli(0.66), influences assignment.
  - X_{i8}: Static Normal(0,1), influences assignment.
  - X_{i2}: Dynamic Bernoulli(0.45).
  - X_{i3}: Dynamic Categorical({1,2,3,4}, {0.3,0.1,0.2,0.4}).
  - $X_{i4} - X_{i7}$: Dynamic Normal(0,1).

  The full vector $\mathbf{X}_{it} = [X_{i1}, X_{i2t}, \ldots, X_{i7t}, X_{i8}]$ consists of these static and dynamic components.

- **Assignment Mechanism:** Staggered adoption times $G_i \in \{\infty, 4, 5, 6\}$ (corresponding to groups $g = 0, 1, 2, 3$) are determined by maximizing a latent utility $U_{ig} = V_g(X_{i1}, X_{i8}) + \psi_{ig}$. The systematic part V_g is a linear function of the static covariates X_{i1} and X_{i8} with group-specific coefficients. For group $g = 1$: $V_{i1} = 0.1 + 0.8X_{i1} + 0.6X_{i8}$, group $g = 2$: $V_{i1} = +-0.5X_{i1} + -0.7X_{i8}$, and group $g = 3$: $V_{i1} = 0.1 + 0.3X_{i1} + 0.4X_{i8}$. The random component $\psi_{ig} \sim \mathcal{N}(0, \sigma^2_{\text{prop}})$ with $\sigma_{\text{prop}} = 0.5$. Unit i is assigned the group g^* that maximizes U_{ig}.

- **Treatment Effect:** again, for this DGP, $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 3 \ \forall k_{it} \geq 0$ for Setting 1 & 2, and $\tau(\mathbf{X}_{it}, k_{it}) = \tau(k_{it}) = 5 \ \forall k_{it} \geq 0$ for Setting 3.

4.4.2.3.2 Real-life Example:

Chirinko and Wilson (2008)

A study by Chirinko and Wilson (2008) examined how U.S. states adopted investment tax credits (ITCs) for businesses at different times based on observable characteristics (like neighboring states' policies, economic conditions, and political factors). States did not randomly adopt these policies but rather made decisions based on utility maximization - specifically, states with weaker economies or those competing with neighbors who already had such incentives were more likely to adopt ITCs.

4.4.2.4 DGP 4: Non-Staggered Adoption with Propensity Score Assignment and CHTE (CATT Focus)

This DGP features simultaneous treatment adoption ($t_0 = 4$) with selection into treatment based on a propensity score derived from static covariates, and includes CATT depending on dynamic covariates.

4.4.2.4.1 Implementation Details:

- **Covariates:** $p = 8$.
  - X_{i1}: Static Bernoulli(0.66), influences assignment.
  - X_{i7}: Static Normal(0,1), influences assignment.
  - X_{i2}: Dynamic Bernoulli(0.45).
  - X_{i3}: Dynamic Normal(0,1), influences CHTE.
  - $X_{i4} - X_{i6}$: Dynamic Normal(0,1).
  - X_{i8}: Dynamic Categorical($\{1,2,3,4\}$, $\{0.3,0.1,0.2,0.4\}$), influences CHTE.

  The full vector $\mathbf{X}_{it} = [X_{i1}, X_{i2t}, \ldots, X_{i6t}, X_{i7}, X_{i8t}]$.
- **Assignment Mechanism:** Treatment status D_i (time-invariant) is determined via a propensity score $\pi(X_{i1}, X_{i7}) = P(D_i = 1 | X_{i1}, X_{i7}) = \sigma(\theta_0 + \theta_1 X_{i1} + \theta_7 X_{i7})$, where $\sigma(\cdot)$ is the logistic sigmoid function and coefficients are $\theta_0 = 0.0, \theta_1 = 0.5, \theta_7 = -0.5$. Unit i is treated ($D_i = 1$) if a draw $u_i \sim U(0,1)$ is less than $\pi(X_{i1}, X_{i7})$.
- **Treatment Effect:** $\tau(\mathbf{X}_{it}, k_{it})$ depends on dynamic covariates X_{i3t} and X_{i8t}. Let τ_{base} be the baseline effect size (3.0 for Settings 1, 2; 5.0 for Setting 3). The potential CATE is:

$$\tau(\mathbf{X}_{it}) = \begin{cases} \tau_{\text{base}} + 1.5\sqrt{|X_{i3t}|} & \text{if } X_{i8t} \in \{1,3\} \\ \tau_{\text{base}} & \text{if } X_{i8t} = 2 \\ \tau_{\text{base}} - 0.5\sqrt{|X_{i3t}|} & \text{if } X_{i8t} = 4 \end{cases}$$

4.4.2.4.2 Real-life Example:

FIGLIO (1998). A study by FIGLIO (1998) examined the effects of property tax limits on school district spending. While all districts in a state faced the same implementation date for these limits, the impact varied based on district characteristics. The effect of tax limits depended on dynamic factors like district wealth, student demographics, and prior spending levels - creating conditional heterogeneous treatment effects. The focus was on estimating the CATT for different types of school districts.

4.4.2.5 DGP 5: Staggered Adoption with Selection and CHTE (CATT & GATE Focus)

This DGP combines complexities: staggered adoption (G_i), selection into timing groups via utility maximization based on static covariates and treatment effect depending on dynamic covariates.

4.4.2.5.1 Implementation Details:

- **Covariates:** $p = 8$.
  - X_{i1}: Static Bernoulli(0.66), influences assignment.
  - X_{i8}: Static Normal(0,1), influences assignment.
  - X_{i2}: Dynamic Bernoulli(0.45).
  - X_{i3}: Dynamic Categorical($\{1,2,3,4\}$, $\{0.3,0.1,0.2,0.4\}$), influences CHTE.
  - X_{i4}: Dynamic Normal(0,1), influences CHTE modifier term.
  - $X_{i5} - X_{i7}$: Dynamic Normal(0,1).

  The full vector $\mathbf{X}_{it} = [X_{i1}, X_{i2t}, \ldots, X_{i7t}, X_{i8}]$. Note: While the draft text suggested X_{i1} modifies the CATE, the code uses the dynamic X_{i4t} for the modifier; we describe the code's implementation here.

- **Assignment Mechanism:** Staggered adoption times $G_i \in \{\infty, 4, 5, 6\}$ determined by utility maximization based on static X_{i1} and X_{i8}, identical to DGP 3.

- **Treatment Effect:** $\tau(\mathbf{X}_{it}, k_{it})$ depends on dynamic covariates X_{i3t} and X_{i4t}. Let τ_{base} be the baseline effect size (3.0 for Settings 1, 2; 5.0 for Setting 3). The potential CATE is:

$$\tau(\mathbf{X}_{it}) = \begin{cases} \tau_{\text{base}} + 1.5\sqrt{|X_{i4t}|} & \text{if } X_{i3t} \in \{1,3\} \\ \tau_{\text{base}} & \text{if } X_{i3t} = 2 \\ \tau_{\text{base}} - 0.5\sqrt{|X_{i4t}|} & \text{if } X_{i3t} = 4 \end{cases}$$

4.4.2.5.2 Real-life Example:

Greenstone (2002).

A study by Greenstone (2002) examined how the Clean Air Act Amendments affected manufacturing plants. Counties were designated as "non-attainment" (subject to stricter regulations) or "attainment" (not subject to these regulations) based on their pollution levels, and these designations changed over time.

This created a staggered adoption scenario where:

1. Counties entered treatment (non-attainment status) at different times
2. Selection into timing groups was based on observable air quality measures (though for our DGP, the covariates influencing the selection into timing groups are static)
3. Treatment effects varied by industry type, plant size, and other dynamic factors

Though for our DGP, the covariates influencing the selection into timing groups are static while in Greenstone (2002) they were dynamic, this example matches our DGP 5's complex structure with staggered adoption, selection based on covariates, and treatment effects dependent on dynamic characteristics.

4.4.3 Evaluation Metrics

To comprehensively assess the performance of various estimators across the different DGPs and settings, we employed three evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics quantify the discrepancy between the estimated treatment effects on the treated and the true, known treatment effects on the treated from our simulations (i.e., ATT, GATT, or CATT).

The RMSE measures the square root of the average of the squared differences between these estimated and true treatment effects. It is particularly sensitive to large errors due to the squaring term and is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{M_{T_{post}}} \sum_{t=1}^{T_{post}} \sum_{j=1}^{M_t} (\hat{\tau}(\mathbf{X}_{jt}, k_{jt}) - \tau(\mathbf{X}_{jt}, k_{jt}))^2},$$

where $M_{T_{post}}$ is the total number of treated observations for the post-treatment period and M_t is the number of treated observations at time t.

The MAE, on the other hand, calculates the average of the absolute differences between the estimated and true treatment effects, defined as

$$\text{MAE} = \sqrt{\frac{1}{M_{T_{post}}} \sum_{t=1}^{T_{post}} \sum_{j=1}^{M_t} |\hat{\tau}(\mathbf{X}_{jt}, k_{jt}) - \tau(\mathbf{X}_{jt}, k_{jt})|},$$

Unlike RMSE, MAE treats all errors with equal weight in the averaging process.

Lastly, the MAPE expresses the average absolute difference as a percentage of the true treatment effect, calculated as:

$$\text{MAPE} = \sqrt{\frac{1}{M_{T_{post}}} \sum_{t=1}^{T_{post}} \sum_{j=1}^{M_t} \frac{|\hat{\tau}(\mathbf{X}_{jt}, k_{jt}) - \tau(\mathbf{X}_{jt}, k_{jt})|}{|\tau(\mathbf{X}_{jt}, k_{jt})|}},$$

This metric is scale-independent, which can be useful for comparing performance across effects of different magnitudes. The nature of τ in the denominator varies with the DGP. For DGP 4 (non-staggered adoption with CHTE) and DGP 5 (staggered adoption with selection and CHTE), the treatment effect which can vary based on dynamic covariates. In these CHTE scenarios, RMSE and MAE inherently give more weight to higher (in magnitude) CATT values. This means that a large error in estimating a large CATT has a disproportionately greater impact on these metrics than a similar-sized error in estimating a small CATT. MAPE, however, normalizes the error by the true treatment effect before averaging. This means that a 10% error in estimating a CATT of 1 has the same impact on MAPE as a 10% error in estimating a CATT of 10. Therefore, MAPE provides a more balanced view of estimator performance, reflecting the percentage accuracy across the entire range of CATT values, regardless of their absolute magnitude. In essence, MAPE gives equal weight to the proportional error in estimating each individual treatment effect, making it more suitable when the goal is to assess the relative accuracy of the estimations and avoid being dominated by the magnitude of the estimated effects themselves. This is especially pertinent when the CATTs exhibit significant

In addition to the accuracy of point estimates, we evaluated the statistical inference properties of the models. We assessed the performance of the statistical tests for the null hypothesis of no treatment effect (i.e., $H_0 : \tau = 0$). First, to estimate the statistical power, we used the simulations where a true treatment effect exists and calculated the frequency with which each model correctly rejected the null hypothesis, using a standard significance threshold of p-value < 0.05. A higher frequency indicates greater power to detect a true effect. Second, to assess the validity of the tests, we conducted separate simulations where the true treatment effect was set to zero. For these null-effect scenarios, we measured the empirical size (Type I error rate) of each test, which is the frequency of incorrectly rejecting the null hypothesis. A well-calibrated test should have an empirical size close to the nominal level of 0.05.

While for the benchmark models, we use their standard treatment effect statistical testing methods (which are not described in this paper to ensure its sparsity, but can be found in their respective original papers and software (SANT'ANNA; ZHAO, 2020; GARDNER, 2022; ARKHANGELSKY *et al.*, 2021; BACH *et al.*, 2022)), the Bayesian nature of the proposed DiD-BCF model allows for a more direct and intuitive approach to inference.

For the DiD-BCF model, we leverage the full posterior distribution of the treatment effect parameter, τ, obtained from the MCMC sampling process. Let $\{\tau^{(s)}\}_{s=1}^{S}$ denote the set of S draws from the posterior distribution of the treatment effect. Instead of relying on a frequentist p-value, we can directly compute the posterior probability that the parameter lies on either side of zero. This quantity serves as a direct measure of evidence for or against a directional hypothesis. Additionally, different from the other benchmark models, now we have evidence for or against a directional hypothesis in the level of CATE (and not only ATE or GATE anymore). This is a great advantage of this study's proposed model.

The procedure is as follows: we first estimate the posterior probability of the effect being positive and the probability of it being negative using the MCMC samples:

$$P(\tau > 0|\text{data}) \approx \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(\tau^{(s)} > 0)$$

$$P(\tau < 0|\text{data}) \approx \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(\tau^{(s)} < 0)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

For a two-sided hypothesis test against the null $H_0 : \tau = 0$, we define a decision metric as the smaller of these two posterior probabilities:

$$p_{\text{Bayes}} = \min\left(P(\tau > 0|\text{data}), P(\tau < 0|\text{data})\right)$$

This value, p_{Bayes}, represents the posterior probability of the parameter having a sign opposite to the one suggested by the bulk of the posterior mass (i.e., the posterior median). We then reject the null hypothesis $H_0 : \tau = 0$ if this posterior probability is below a pre-specified significance level, such as $\alpha_{H_0} = 0.05$ used in this paper. This approach, while analogous to a frequentist p-value in its use as a decision threshold, provides a more direct probabilistic statement about the parameter itself, which is a distinct advantage of the Bayesian framework.

Incidentally, although such a treatment effect statistical testing practice (especially regarding the frequentist models) for applied research is not perfect and has received some critique in the past years (POOLE, 2001; GREENLAND *et al.*, 2016; GANNON; PEREIRA; POLPO, 2019; BLUME *et al.*, 2018; WASSERSTEIN; LAZAR, 2016; WASSERSTEIN; SCHIRM; LAZAR, 2019), it is still the standard practice in the applied causal inference research realm (WASSERSTEIN; LAZAR, 2016; WASSERSTEIN;

SCHIRM; LAZAR, 2019); hence, proposing models that can properly perform such a treatment effect statistical test is key for the model adoption in the applied research domain. Additionally, it is important to note that when the treatment effect is set to zero, DGP 4 and DGP 5 become structurally equivalent to DGP 1 and DGP 3, respectively. The complete error metric results for the simulation studies where the treatment effect is zero are presented in Appendix C. As the conclusions from these results converge with the findings from our main simulation studies, we have placed them in the appendix to ensure the sparsity of the main text.

The choice of these evaluation criteria provides a multifaceted view of estimator performance. RMSE highlights the impact of larger errors, MAE offers a robust measure of average error magnitude, MAPE provides a relative assessment of accuracy, and the analysis of power and size evaluates the reliability of the statistical inference. Together, they allow for a nuanced understanding of how well the proposed model and each benchmark model recover the true causal parameters under different simulation conditions, characterized by varying complexities such as staggered treatment adoption (DGP 2, 3, 5), selection on observables (DGP 3, 4, 5), and conditional treatment effect heterogeneity (DGP 4, 5). The tables present the error metrics as 'mean ± standard deviation' across the simulation runs.

4.5 Results and Discussion

DGP 1

The results of the Monte Carlo study for DGP 1 can be found in Table 7. It is worth remembering that the target estimand here is the ATT.

Table 7 – Overall Performance Comparison

Metric	RMSE	MAE	MAPE
Setting 1			
DiD-BCF	0.1522 ± 0.0548	0.1214 ± 0.0473	0.0405 ± 0.0158
TWFE	0.5596 ± 0.1355	0.5275 ± 0.1374	0.1758 ± 0.0458
DiD DR	0.5530 ± 0.1879	0.4378 ± 0.1598	0.1459 ± 0.0533
DiD2s	0.1589 ± 0.0585	0.1256 ± 0.0503	0.0419 ± 0.0168
SDiD	$\mathbf{0.0981 \pm 0.0628}$	$\mathbf{0.0756 \pm 0.0628}$	$\mathbf{0.0252 \pm 0.0209}$
DoubleML_did	0.7237 ± 0.2434	0.5734 ± 0.2204	0.1911 ± 0.0735
Setting 2			
DiD-BCF	0.5079 ± 0.3397	$\mathbf{0.2678 \pm 0.2155}$	$\mathbf{0.0893 \pm 0.0718}$
TWFE	1.0243 ± 0.4133	0.8269 ± 0.3979	0.2756 ± 0.1326
DiD DR	1.2927 ± 0.4652	1.0275 ± 0.4135	0.3425 ± 0.1378
DiD2s	0.7528 ± 0.2782	0.5901 ± 0.2388	0.1967 ± 0.0796
SDiD	$\mathbf{0.4667 \pm 0.2801}$	0.3744 ± 0.2801	0.1248 ± 0.0934
DoubleML_did	1.4818 ± 0.5402	1.1707 ± 0.4690	0.3902 ± 0.1563
Setting 3			
DiD-BCF	0.5310 ± 0.3430	$\mathbf{0.3112 \pm 0.2434}$	$\mathbf{0.0622 \pm 0.0487}$
TWFE	2.6140 ± 0.6926	2.4520 ± 0.7001	0.4904 ± 0.1400
DiD DR	1.2927 ± 0.4652	1.0275 ± 0.4135	0.2055 ± 0.0827
DiD2s	0.7528 ± 0.2782	0.5901 ± 0.2388	0.1180 ± 0.0478
SDiD	$\mathbf{0.4670 \pm 0.2785}$	0.3760 ± 0.2785	0.0752 ± 0.0557
DoubleML_did	1.3963 ± 0.4716	1.1123 ± 0.4341	0.2225 ± 0.0868

In Setting 1, where the data generating process is fully linear and the treatment effect $\tau = 3$, the SDiD model emerges as the top performer, achieving the lowest RMSE (0.0981), MAE (0.0756), and MAPE (2.52%). This is a significant finding, as the synthetic control-based reweighting of SDiD proves to be exceptionally effective in this pure linear setting, even more than correctly specified linear models. It is followed closely by DiD2s (RMSE 0.1589) and DiD-BCF (RMSE 0.1522), which deliver nearly identical, strong performance and are the next best models. In contrast, the traditional TWFE model, along with the more complex DiD DR and DoubleML_did estimators, exhibit substantially higher errors.

As we move to Setting 2, which introduces partial non-linearity, the performance landscape becomes more nuanced and competitive. SDiD still achieves the lowest RMSE (0.4667), demonstrating its robustness to nonlinearity thanks to its weighting strategy. However, DiD-BCF shows superior performance in terms of MAE (0.2678 vs. SDiD's 0.3744) and MAPE (8.93% vs. SDiD's 12.48%). This suggests that while SDiD produces estimates with a smaller variance, DiD-BCF's estimates are, on average, closer to the true value and have a smaller relative error. Both SDiD and DiD-BCF significantly outperform

the other estimators. Linear models like TWFE and DiD2s suffer from misspecification bias, as expected, while DiD DR and DoubleML_did continue to show the highest errors, struggling to adapt effectively in this simple ATT scenario.

In Setting 3, the scenario becomes significantly more challenging with full non-linearity and a quadratic time trend. The pattern from Setting 2 continues: SDiD again shows a slight edge in RMSE (0.4670), while DiD-BCF once more leads on MAE (0.3112 vs. 0.3760) and MAPE (6.22% vs. 7.52%). This consistency highlights the distinct advantages of the proposed model and SDID in complex environments. In stark contrast, all other benchmark models exhibit a dramatic deterioration in performance. The rigid assumptions of TWFE and DiD2s are heavily violated, leading to massive errors. Similarly, the DiD DR and DoubleML_did estimators are unable to cope with the high degree of non-linearity, cementing the superior adaptability of SDiD and DiD-BCF.

Figure 15 shows the frequency of rejecting $H_0 : \tau = 0$ with the p-value threshold $\alpha_{H_0} = 0.05$. While the power for all models besides DoubleML_did is virtually 1 for Setting 1, only the models DiD-BCF, DiD2s and SDiD have a great power for the other settings, with SDID having a slight superiority. Yet, when considering the cases where $\tau = 0$, we can see that SDID is too conservative, indicating that its power would presumably be smaller than DiD-BCF and especially DiD2s for values of τ closer to zero. Thus, it can be concluded that for DGP 1, the DiD2s is the best model concerning treatment effect detection.

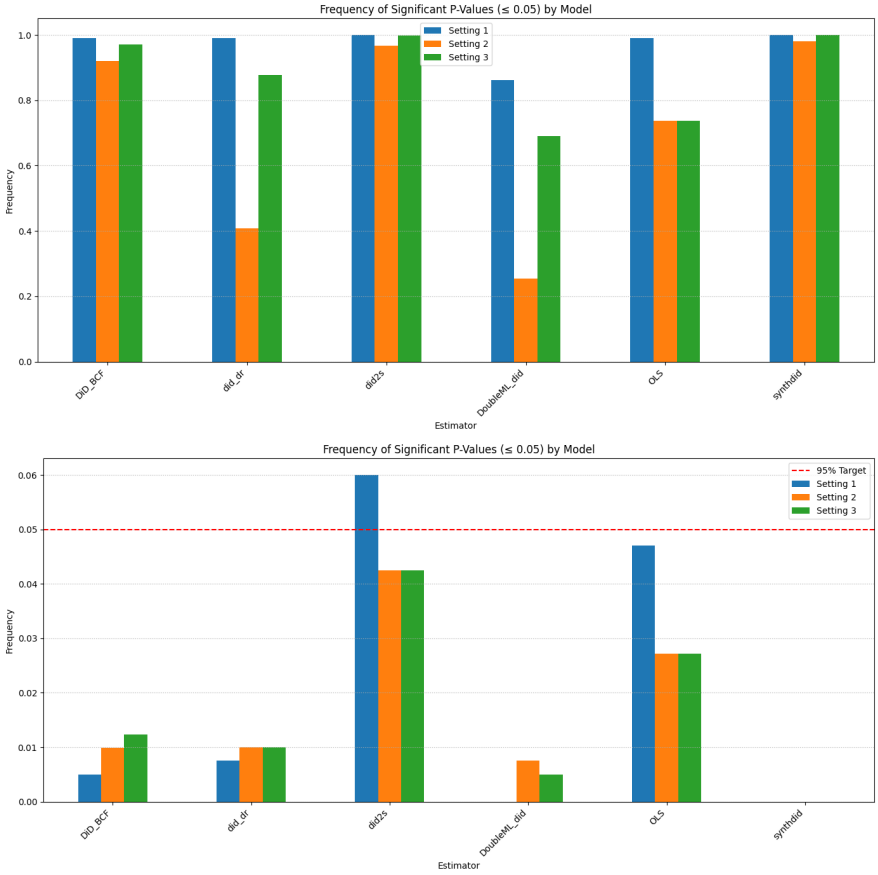

Figure 15 – Frequency of $H_0 : \tau = 0$ being reject for DGP 1 (left figure is for $\tau \neq 0$ and right figure is for $\tau = 0$, and OLS = TWFE model).

DGP 2

Table 8 show the results for DGP 2. It is worth remembering that the target estimand here is the GATT and we are mimicking here more an experimental setting than an observational one given the perfect randomness of treatment assignment.

Table 8 – Overall Performance Comparison

Metric	RMSE	MAE	MAPE (in %)
Setting 1			
DiD-BCF	0.2967 ± 0.1699	0.2012 ± 0.1483	0.0894 ± 0.0659
TWFE	1.0241 ± 0.0204	0.7576 ± 0.0519	0.1831 ± 0.0288
DiD DR	0.8469 ± 0.1979	0.6901 ± 0.1698	0.2300 ± 0.0630
DiD2s	$\mathbf{0.1171 \pm 0.0461}$	$\mathbf{0.0955 \pm 0.0373}$	$\mathbf{0.0368 \pm 0.0164}$
SDiD	1.3038 ± 0.0058	1.1191 ± 0.0558	0.2461 ± 0.0372
DoubleML_did	1.1104 ± 0.2509	0.8903 ± 0.2053	0.2941 ± 0.0768
Setting 2			
DiD-BCF	0.3427 ± 0.2077	0.2270 ± 0.1831	0.1009 ± 0.0814
TWFE	1.0476 ± 0.0390	0.7930 ± 0.0834	0.1958 ± 0.0403
DiD DR	0.9326 ± 0.2255	0.7623 ± 0.1918	0.2540 ± 0.0706
DiD2s	$\mathbf{0.1688 \pm 0.0693}$	$\mathbf{0.1366 \pm 0.0585}$	$\mathbf{0.0522 \pm 0.0261}$
SDiD	1.3073 ± 0.0117	1.1087 ± 0.0722	0.2392 ± 0.0481
DoubleML_did	1.2406 ± 0.2928	0.9946 ± 0.2399	0.3241 ± 0.0875
Setting 3			
DiD-BCF	0.5527 ± 0.4323	0.2874 ± 0.3773	0.0766 ± 0.1006
TWFE	1.7171 ± 0.0404	1.2733 ± 0.1102	0.1868 ± 0.0346
DiD DR	0.6604 ± 0.1917	0.5493 ± 0.1659	0.1091 ± 0.0338
DiD2s	$\mathbf{0.2230 \pm 0.0870}$	$\mathbf{0.1817 \pm 0.0726}$	$\mathbf{0.0421 \pm 0.0195}$
SDiD	2.1753 ± 0.0151	1.8549 ± 0.1045	0.2420 ± 0.0418
DoubleML_did	0.8530 ± 0.2555	0.6821 ± 0.1938	0.1328 ± 0.0404

In Setting 1, the DiD2s estimator demonstrates exceptionally strong performance. This superior performance in a correctly specified linear environment can be attributed to its two-stage estimation strategy. When the model is correctly specified and treatment timing is random (as in DGP2), the direct, sequential OLS approach of DiD2s is highly efficient. In comparison, DiD-BCF also performs well, significantly outperforming other benchmark models.

The DiD DR estimator, while also designed for staggered adoption and correctly specified in this setting, shows higher errors than DiD2s. This difference might arise because DiD DR, often a doubly robust method involving estimation of outcome regression and/or propensity score models, might introduce slightly more finite-sample variance from estimating these nuisance components, even if they are simple. DiD2s, by its construction in this specific DGP, relies on a more direct OLS approach for $Y(0)$ components from a clean sample, which can be more efficient than a more general DR framework when its assumptions are perfectly met.

The remarkable performance of DiD2s continues into Setting 2. The robustness of DiD2s to this form of misspecification in the first stage (which assumes linear covari-

ate effects) is noteworthy. While the linear model for $Y_{it}(0)$ estimated in its first stage is now misspecified, the unit and period fixed effects can still absorb the average group and time variations. Crucially, because the assignment to treatment timing groups in DGP2 is entirely random, the unmodelled non-linear components of $Y_{it}(0)$ (after linear adjustment) might remain largely orthogonal to the treatment indicators in the second stage. This orthogonality would prevent the first-stage misspecification from severely biasing the GATT estimate in the second stage. DiD-BCF also exhibits strong robustness, with its error metrics increasing only moderately. It consistently outperforms TWFE, DiD DR, SynthDiD, and DoubleML_did, whose performances degrade more substantially due to the non-linearities.

Similarly, in Setting 3, DiD2s continues to lead by a significant margin (RMSE 0.2230, MAE 0.1817, MAPE interpreted as 4.21%). DiD-BCF is anew the next best performer, demonstrating its capacity to handle significant non-linearities better than the other remaining benchmarks.

Figure 16 depicts the frequency of rejecting $H_0 : \tau = 0$ with the standard p-value threshold. Similarly to DGP 1, DiD2s is the best model concerning treatment effect detection, followed by DiD-BCF and TWFE. Anew, SDID has a great power yet is too conservative, indicating that its power would be smaller.

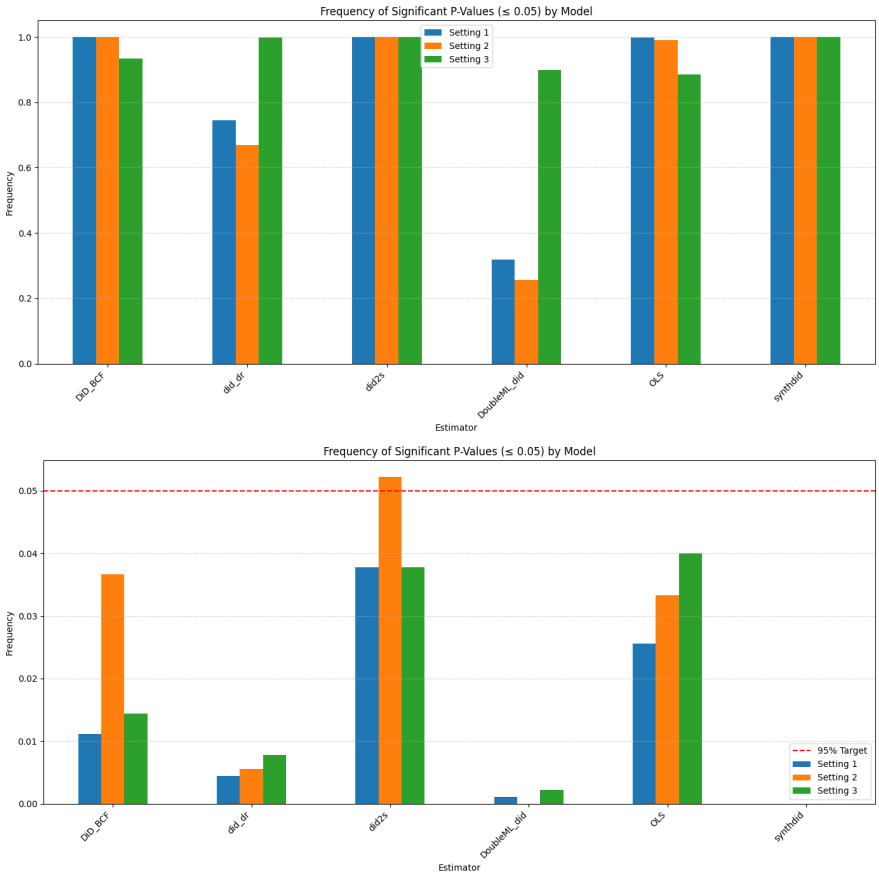

Figure 16 – Frequency of $H_0 : \tau = 0$ being reject for DGP 2 (left figure is for $\tau \neq 0$ and right figure is for $\tau = 0$, and OLS = TWFE model).

DGP 3

The results of the simulation study for DGP 3 can be found in Table 9. The target estimand here is anew the GATT but now we are mimicking an observational setting.

Table 9 – Overall Performance Comparison

Metric	RMSE	MAE	MAPE (in %)
Setting 1			
DiD-BCF	$\mathbf{0.3090 \pm 0.1767}$	$\mathbf{0.2063 \pm 0.1549}$	$\mathbf{0.0917 \pm 0.0688}$
TWFE	1.1844 ± 0.1055	0.8444 ± 0.0990	0.1411 ± 0.0527
DiD DR	0.7553 ± 0.3529	0.6314 ± 0.2991	0.2137 ± 0.1082
DiD2s	N/A	N/A	N/A
SDiD	0.9981 ± 0.0489	0.6768 ± 0.1502	0.1374 ± 0.0621
DoubleML_did	3.8608 ± 10.4084	2.5502 ± 5.4474	0.8571 ± 2.0345
Setting 2			
DiD-BCF	$\mathbf{0.3499 \pm 0.2159}$	$\mathbf{0.2263 \pm 0.1892}$	$\mathbf{0.1006 \pm 0.0841}$
TWFE	1.7580 ± 0.4468	1.4017 ± 0.4335	0.3733 ± 0.1734
DiD DR	3.1489 ± 1.3033	2.6181 ± 1.1663	0.8754 ± 0.4173
DiD2s	N/A	N/A	N/A
SDiD	1.2100 ± 0.2878	0.9377 ± 0.4007	0.2380 ± 0.1741
DoubleML_did	10.7815 ± 9.0965	7.5092 ± 4.9488	2.4504 ± 1.8014
Setting 3			
DiD-BCF	$\mathbf{0.4365 \pm 0.2480}$	$\mathbf{0.2906 \pm 0.2097}$	$\mathbf{0.0775 \pm 0.0559}$
TWFE	2.3976 ± 0.4381	1.8724 ± 0.4182	0.2638 ± 0.1236
DiD DR	3.5792 ± 1.6355	2.9983 ± 1.4708	0.6051 ± 0.3114
DiD2s	N/A	N/A	N/A
SDiD	1.8170 ± 0.2918	1.3020 ± 0.5248	0.1794 ± 0.1399
DoubleML_did	11.2100 ± 18.6878	7.6799 ± 8.3195	1.5219 ± 1.8701

N/A indicates that the model is not applicable for this DGP and Setting due to unbalanced panel data.

In Setting 1, DiD-BCF achieves the best performance among the available estimators with an RMSE of 0.3090, MAE of 0.2063, and MAPE of 9.17%. Its ability to flexibly model the outcome $\mu(\cdot)$ conditional on all covariates, while exploiting the its simpler formulation, likely helps in controlling for the selection bias. SynthDiD and DiD DR perform next best among the remaining benchmarks, suggesting they can handle selection on observables to some extent, though less effectively than DiD-BCF. TWFE struggles more, as it does not explicitly account for the selection mechanism in the same way. DoubleML_did, on the other hand, exhibits very high errors and variability (RMSE 3.8608 ± 10.4084), indicating potential instability or difficulty in correctly specifying/estimating its nuisance functions in this selection scenario with staggered adoption. Presumably, this is the case since for certain simulations, the number of units of a certain group is below 10% of the total number, increasing the limitation that DoubleML_did has, namely its need for a great amount of data to properly converge given its slow convergence due to its nonparametric and formulation nature.

Under Setting 2, DiD-BCF maintains its lead and barely suffers any downgrade

in performance, showcasing its robustness to the introduction of non-linearities in the outcome model even when treatment selection is present. The other estimators, on the other hand, see a more marked decline, especially DiD DR, which indicates that its good performance for Setting 1 was solely due to its regression estimation part and not the IPW part. DoubleML_did continues to struggle significantly.

In Setting 3, DiD-BCF continues to be the top performer (RMSE 0.4365, MAE 0.2906, MAPE 7.75%). The increased complexity of the DGP further highlights the limitations of the other methods, with DoubleML_did again showing very high errors.

The frequency of rejecting $H_0 : \tau = 0$ with $\alpha_{H_0} = 0.05$ is presented in Figure 17. For DGP 3, DiD-BCF is clearly the most performing model for treatment effect detection for Setting 2 and 3 (though a bit conservative). Yet for Setting 1, TWFE is as powerful as DiD-BCF, yet it does not suffer from being too conservative as DiD-BCF. Nonetheless, it is worth mentioning that while the TWFE detects *a general* treatment effect, DiD-BCF can effectively detect treatment effects per group (thus, giving more value for researchers and practitioners). As a result, for DGP 3, it can be concluded that DiD-BCF is the best model concerning (group) treatment effect detection.

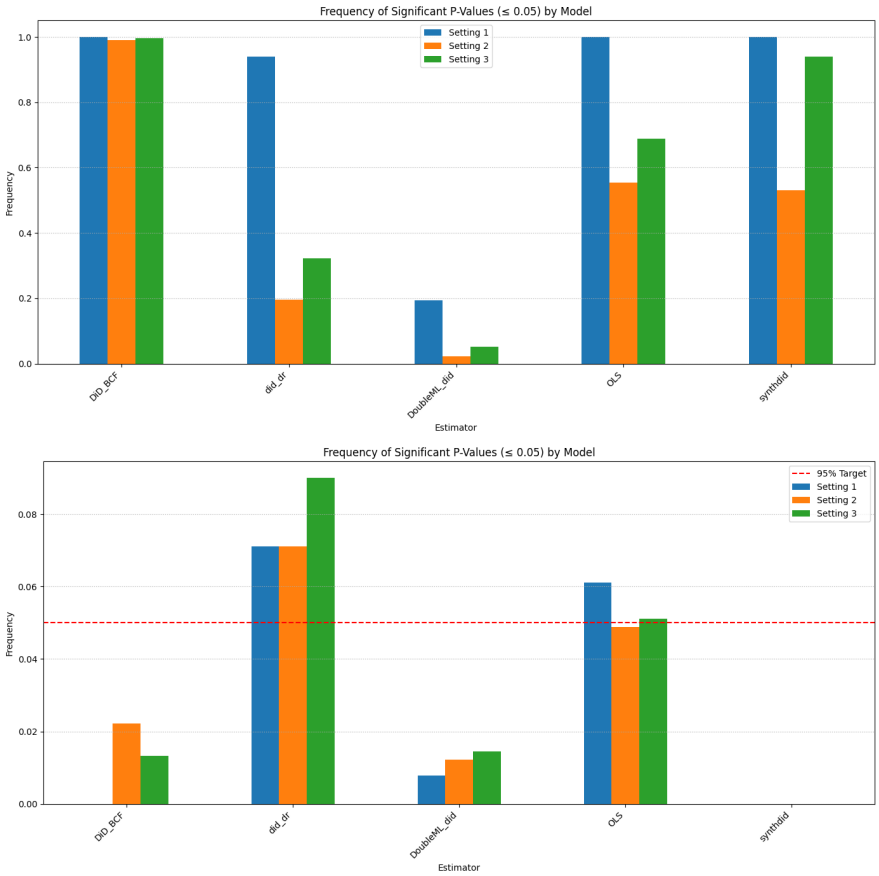

Figure 17 – Frequency of $H_0 : \tau = 0$ being reject for DGP 3 (left figure is for $\tau \neq 0$ and right figure is for $\tau = 0$), and OLS = TWFE model.

DGP 4

Table 10 present the results for DGP 4. It is worth remembering that the target estimand here is the CATT and we are mimicking an observational setting with only two groups and CHTE.

Table 10 – Overall Performance Comparison for DGP 4

Metric	RMSE	MAE	MAPE
Setting 1			
DiD-BCF	$\mathbf{0.3623 \pm 0.0446}$	$\mathbf{0.2849 \pm 0.0390}$	$\mathbf{0.0850 \pm 0.0126}$
TWFE	0.8984 ± 0.0324	0.7938 ± 0.0265	0.2398 ± 0.0156
DiD DR	1.0762 ± 0.1424	0.8989 ± 0.1099	0.2740 ± 0.0458
DiD2s	N/A	N/A	N/A
SDiD	0.8856 ± 0.0219	0.7908 ± 0.0212	0.2377 ± 0.0107
DoubleML_did	1.1536 ± 0.1615	0.9559 ± 0.1238	0.2944 ± 0.0462
Setting 2			
DiD-BCF	$\mathbf{0.3654 \pm 0.0440}$	$\mathbf{0.2876 \pm 0.0380}$	$\mathbf{0.0859 \pm 0.0125}$
TWFE	0.9055 ± 0.0376	0.7962 ± 0.0292	0.2404 ± 0.0187
DiD DR	1.0594 ± 0.1315	0.8856 ± 0.1015	0.2693 ± 0.0442
DiD2s	N/A	N/A	N/A
SDiD	0.8886 ± 0.0235	0.7922 ± 0.0217	0.2384 ± 0.0125
DoubleML_did	1.1268 ± 0.1411	0.9332 ± 0.1059	0.2862 ± 0.0438
Setting 3			
DiD-BCF	$\mathbf{0.3789 \pm 0.0543}$	$\mathbf{0.2969 \pm 0.0469}$	$\mathbf{0.0547 \pm 0.0092}$
TWFE	1.0191 ± 0.0993	0.8583 ± 0.0727	0.1597 ± 0.0199
DiD DR	1.1105 ± 0.1922	0.9188 ± 0.1426	0.1693 ± 0.0291
DiD2s	N/A	N/A	N/A
SDiD	0.9200 ± 0.0603	0.7989 ± 0.0333	0.1456 ± 0.0101
DoubleML_did	1.1331 ± 0.1877	0.9374 ± 0.1370	0.1739 ± 0.0275

N/A indicates that the model is not applicable for this DGP and Setting due to unbalanced panel data.

In Setting 1, DiD-BCF clearly provides the most accurate estimates of CATT, with an RMSE of 0.3623, MAE of 0.2849, and MAPE interpreted as 8.50%. This is expected, as BCF is specifically designed to model and estimate CHTE. The benchmark models are primarily designed to estimate average (group) effects. While they incorporate covariates, they do not inherently model treatment effect heterogeneity as a function of those covariates without specific modifications. Consequently, their error metrics are substantially higher Their reported error metrics results reflect their inability to capture the true CATT, likely estimating some average effect that poorly approximates the varying individual treatment effects.

This pattern of DiD-BCF dominance continues and becomes even more pronounced

in Setting 2 and Setting 3. The superior performance of DiD-BCF in DGP 4 is a strong testament to its core design: the BCF component explicitly models $\tau(\mathbf{X}_{it}, k_{it})$ as a flexible function of covariates, allowing it to capture CHTE. Its simple yet effective formulation, presence of a flexible the treatment effect function $\tau(\cdot)$, and nonparametric modeling are crucial when treatment effects are not constant and the underlying DGP is (high) nonlinear and consequently (high) complex.

Figure 18 presents the frequency of rejecting $H_0 : \tau = 0$ with the p-value threshold $\alpha_{H_0} = 0.05$. When considering the model's capacity of detecting *a general* treatment effect per post-treatment period, the TWFE model is anew the best one. Yet, for detecting treatment effect per post-treatment period conditional on covariates, then only the DiD-BCF model can do it. In fact, DiD-BCF does it quite effectively, albeit being a bit too conservative.

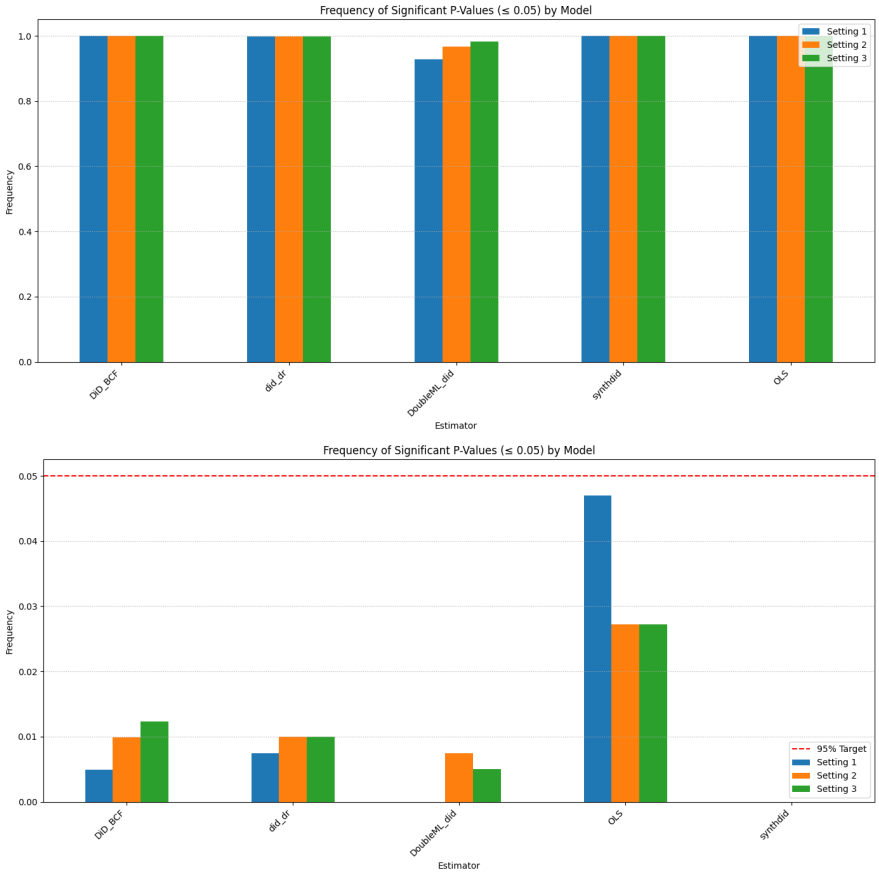

Figure 18 – Frequency of $H_0 : \tau = 0$ being reject for DGP 4 (left figure is for $\tau \neq 0$ and right figure is for $\tau = 0$, and OLS = TWFE model).

DGP 5

The results for DGP 5 can be found in Table 11. The target estimand here is anew the CATT but now with a the presence of a staggered treatment.

Table 11 – Overall Performance Comparison for DGP 5

Metric	RMSE	MAE	MAPE
Setting 1			
DiD-BCF	**0.3831 ± 0.0816**	**0.2866 ± 0.0721**	**0.0982 ± 0.0257**
TWFE	1.3491 ± 0.0615	1.0885 ± 0.0497	0.2570 ± 0.0209
DiD DR	1.1332 ± 0.2645	0.9241 ± 0.2159	0.2961 ± 0.0771
DiD2s	N/A	N/A	N/A
SDiD	1.4088 ± 0.0440	1.0677 ± 0.0408	0.2124 ± 0.0112
DoubleML_did	5.0916 ± 18.2942	3.0191 ± 7.2245	0.9572 ± 2.4663
Setting 2			
DiD-BCF	**1.2552 ± 0.9417**	**1.0046 ± 0.9012**	0.3402 ± 0.3065
TWFE	1.8212 ± 0.3651	1.4899 ± 0.3466	0.4052 ± 0.1318
DiD DR	3.2677 ± 1.2665	2.7074 ± 1.1293	0.8367 ± 0.3796
DiD2s	N/A	N/A	N/A
SDiD	1.5766 ± 0.2483	1.2429 ± 0.2752	**0.2820 ± 0.1075**
DoubleML_did	10.5661 ± 16.2515	7.5608 ± 7.4628	2.3921 ± 2.6577
Setting 3			
DiD-BCF	**1.0481 ± 1.2632**	**0.8246 ± 1.2098**	**0.1730 ± 0.2537**
TWFE	2.2960 ± 0.3991	1.8356 ± 0.4088	0.2886 ± 0.1018
DiD DR	3.6880 ± 1.5986	3.0709 ± 1.4375	0.5836 ± 0.2881
DiD2s	N/A	N/A	N/A
SDiD	2.1237 ± 0.2727	1.5764 ± 0.4026	0.2013 ± 0.0992
DoubleML_did	8.9943 ± 6.4537	6.8248 ± 3.9980	1.2782 ± 0.8295

N/A indicates that the model is not applicable for this DGP and Setting due to unbalanced panel data.

In Setting 1, DiD-BCF once again delivers the best performance with a performance similar to the previous DGPs; in other words, though the DGPs get increasingly more complex (and thus more realistic), the performance of our proposed model remains intact. Its ability to concurrently handle staggered adoption, selection on observables, and CHTE sets it apart. The benchmark models struggle considerably more, especially DoubleML_did. Anew, this is probably the case since for certain simulations, the number of units of a certain group is below 10% of the total number, increasing the limitation that DoubleML_did has, namely its need for a great amount of data to properly converge given its slow convergence due to its nonparametric and formulation nature.

As non-linearities are introduced in Setting 2 , DiD-BCF's performance, while degrading in absolute terms (RMSE 1.2552, MAE 1.0046, MAPE interpreted as 34.02%), still remains relatively better or competitive compared to the alternatives that are also struggling. The errors for DiD-BCF are notably higher in this setting of DGP5 compared to similar settings in other DGPs, suggesting that the combination of all complexities (staggered adoption, selection, CHTE, and non-linear $Y(0)$) poses a very significant chal-

lenge for all estimators, including DiD-BCF. In fact, when also considering the standard deviation of the Monte Carlo DGP simulations, one could arguably affirm that the SDiD would be a preferred model.

In Setting 3, the challenging nature of the DGP is evident for all models. DiD-BCF shows an RMSE of 1.0481, MAE of 0.8246, and MAPE interpreted as 17.30%. While these errors are higher than in simpler DGPs, DiD-BCF still offers the most reasonable performance compared to the benchmarks, which are severely affected. Here, even when considering the standard deviation, it would make more sense to use DiD-BCF over SDiD, especially when trying to discover the heterogeneity nature of the treatment effect. The ability of DiD-BCF to provide more stable and accurate estimates in such a demanding scenario underscores the value of its comprehensive approach, namely 1. leveraging BCF to model CHTE, 2. employing a flexible structure for the prognostic score $\mu(\cdot)$ to absorb complex main effects and selection, and 3. benefiting from the PTA-based reparameterization to simplify the treatment effect estimation task.

Figure 19 depicts the frequency of rejecting $H_0 : \tau = 0$ with $\alpha_{H_0} = 0.05$. For DGP 5, DiD-BCF and SDiD are the most powerful models for treatment effect detection, though the DiD-BCF can not only detect group treatment effects as SDiD, but also covariate-dependent treatment effects. Nonetheless, DiD-BCF is less conservative for Setting 2 and 3, indicating that for a treatment effect closer to 0, the power of DiD-BCF would be likely gretaer than the power of SDiD, making it a better model for treatment effect statistical testing for this complex DGP.

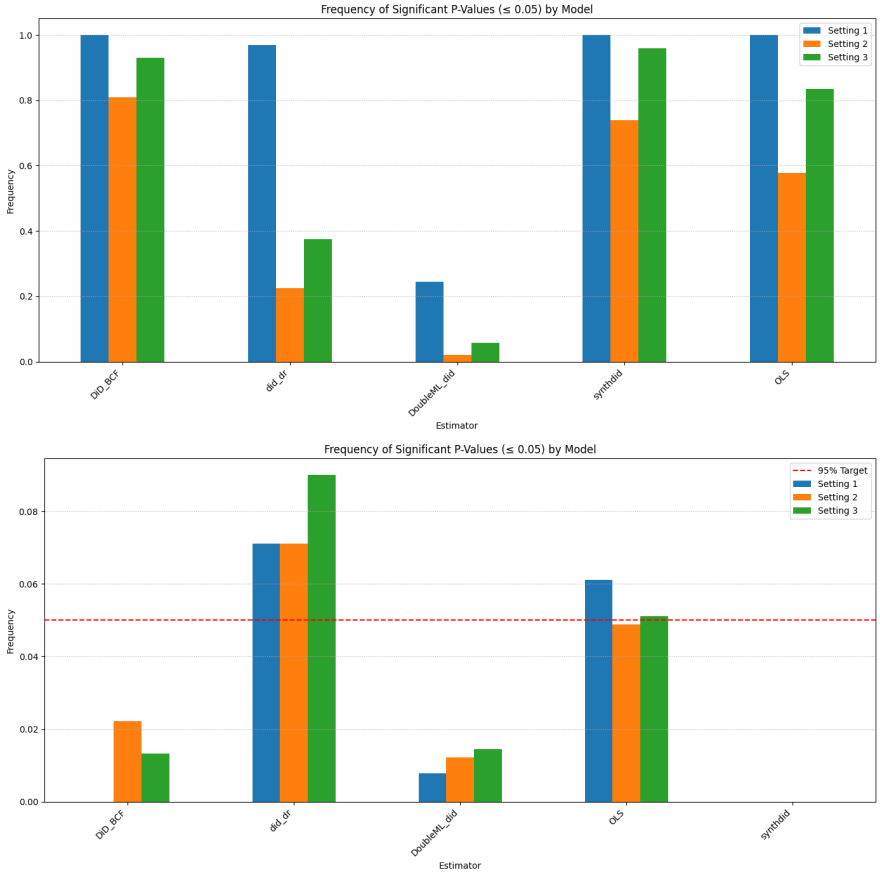

Figure 19 – Frequency of $H_0 : \tau = 0$ being reject for DGP 5 (left figure is for $\tau \neq 0$ and right figure is for $\tau = 0$, and OLS = TWFE model).

Overall Summary

Across all DGPs and settings, the DiD-BCF model consistently demonstrates strong and robust performance. It excels particularly when faced with non-linearities in the outcome model, selection on observables, and conditional heterogeneous treatment effects. Even in the simplest linear scenario, DGP1: Setting 1, DiD-BCF provided the second-most accurate estimates, while being the top model for Setting 2 and 3. Though there is the correct notion that flexible models necessarily underperform when simplicity is true, this is not the case (especially when comparing to the linear models) for DiD-BCF given its reparametrization exploring the PTA. Its advantage becomes more pronounced as the complexity of the DGP increases, highlighting the limitations of traditional linear estimators and even some more contemporary methods when their underlying assumptions are violated or when they are not designed for CHTE. The DiD2s estimator showed exceptional strength in staggered adoption settings with homogeneous effects (DGP2), aligning with its design focus. However, its reported inapplicability to DGPs involving more complex selection or panel structures (as interpreted from "unbalanced panel data" note) limited its comparison in those scenarios. The SynthDiD, on the other hand, offered a good performance throughtout all considered DGPs (especially in DGP 1), albeit

always being overperformed by our proposed model for Settings 2 and 3. The DoubleML_-did estimator, in its current configuration, often struggled, particularly in scenarios with selection and high complexity due to its slow convergence (and thus need for a great amount of data), suggesting potential challenges in its practical application or tuning for such DGPs. The consistent and superior performance of DiD-BCF, especially in capturing CATT and handling complex underlying DGPs simultaneously, underscores its potential as a powerful and versatile tool for causal inference in a wide range of DiD applications.

Beyond point estimate accuracy, the analysis of statistical inference revealed a nuanced landscape. In simpler settings (DGP1, DGP2), models like DiD2s demonstrated excellent power for detecting average treatment effects. However, as complexity increased, the inferential capabilities of DiD-BCF became paramount. In scenarios with selection bias and heterogeneity (DGP3, DGP4, DGP5), DiD-BCF consistently showed high statistical power while remaining well-calibrated, avoiding the overly conservative nature of SDiD in some settings. Critically, while methods like TWFE could detect a general effect, they could not pinpoint its source. DiD-BCF is unique in its ability to perform reliable statistical tests on group-specific (GATT) and conditional (CATT) effects, providing researchers with a far more granular and powerful inferential toolkit.

4.6 Real Life Application

4.6.1 Data and Context

To illustrate the practical utility and distinctive capabilities of our DiD-BCF model, we apply it to a salient policy question: the impact of minimum wage increases on teen employment (WILLIAMS, 2006; BROWN; GILROY; KOHEN, 1981; SEN; RYBCZYNSKI; WAAL, 2011; WELLINGTON, 1991). We utilize a publicly available dataset and compare our findings with established results, focusing particularly on treatment effect heterogeneity.

The data for this application are drawn from the `mpdta` dataset included in the `R` package `did` (CALLAWAY; SANT'ANNA, 2022). This dataset is a subset of the data employed in the study by Sant'Anna and Zhao (2020), focusing on county-level teen employment in the United States. While the original study by Sant'Anna and Zhao (2020) considered a period from 2001–2007 where the federal minimum wage was constant, the `mpdta` subset covers the years 2004 to 2007 and comprises 2000 observations across 500 counties. The key variables include the logarithm of teen employment in a county (`lemp`), the year of observation (`year`), a unique county identifier (`countyreal`), the year the state encompassing the county first raised its minimum wage (`first.treat`), the log of 1000s of population for the county (`lpop`), and an indicator for whether the county is treated in a given year (`treat`). The empirical strategy evaluates the effect of state-level mini-

mum wage increases on teen employment at the county level. States that increased their minimum wage above the federal level during the period are considered "treated" groups, with treatment timing varying by state. States that maintained the federal minimum wage serve as the "untreated" or control group.

For our analysis, and to align with one of the specifications in Sant'Anna and Zhao (2020) given the fact that we do not possess access to the full dataset, especially the other covariates besides population for the county, we focus our analysis under the Unconditional Parallel Trends Assumption (UPTA); that is, no covariates are needed to ensure parallel trends and the use of group and time terms already suffice.

4.6.2 Comparative Results and Heterogeneity Analysis

Sant'Anna and Zhao (2020) report several estimates for the effect of minimum wage increases on teen employment. For a single parameter estimate representing an overall average effect, their findings include:

- Two-Way Fixed Effects (TWFE): -0.037 (standard error 0.006)
- Simple Weighted Average: -0.052 (standard error 0.006)
- DiD DR: -0.090 (standard error 0.013)

These results suggest a significantly negative impact of minimum wage increases on teen employment, and as we increase the flexibility of the models (and presumably their suitability for the underlying dataset), the estimated negative effect increases in magnitude.

Applying our DiD-BCF model to the `mpdta` dataset, we obtain nuanced insights. The overall average treatment effect estimated by DiD-BCF is a single parameter of -0.143. As DiD-BCF estimates the treatment effect function non-parametrically, standard errors for a single coefficient are not directly analogous to parametric models; uncertainty quantification for DiD-BCF estimates is typically visualized, as presented in Figure 20 for our application.

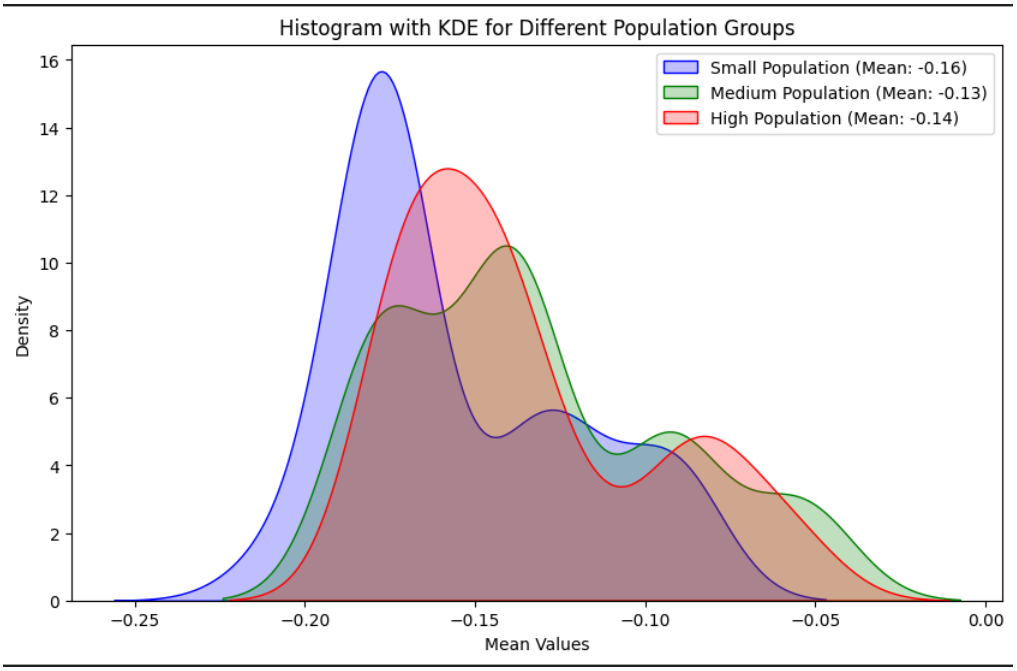

Figure 20 – Estimated Impact of Minimum Wage Increase on Teen Employment Grouped by County Population

A key advantage of DiD-BCF is its ability to explore treatment effect heterogeneity conditional on covariates. When examining the effect across different county population sizes (using `lpop`), we find notable variation:

- Counties with Small Population (up to the 33rd percentile): -0.164
- Counties with Medium Population (from the 33rd to the 66th percentile): -0.132
- Counties with High Population (above the 66th percentile): -0.141

These results indicate that the negative employment effect of minimum wage increases is estimated to be most pronounced in counties with smaller populations, while the effects are comparatively similar for medium and high-population counties, albeit still more negative than the average effects found by the benchmark linear models cited from Sant'Anna and Zhao (2020). The overall DiD-BCF estimate of -0.143 is also more negative than the benchmark averages.

4.6.3 Discussion and Supporting Literature

The heterogeneous findings from our DiD-BCF model, particularly the larger adverse effect in less populous counties, resonate with previous research in the minimum wage literature (THOMPSON, 2009; KALENKOSKI; LACOMBE, 2013).

Thompson (2009), using quarterly county and state data from 1996–2000 to examine federal minimum wage changes, demonstrated that state-level analyses could obscure

important cross-county differences. He found that while state-level analyses showed no significant effects, county-level analyses revealed that a 10% increase in the federal minimum wage led to a 2.6%–3.7% reduction in teen employment for all county sizes, and a more substantial 3.8–5.7% reduction for small counties. This highlights the importance of granular, local analysis and suggests that minimum wage effects are not uniform.

Several economic factors could explain why smaller, lower-population counties might experience more significant negative employment effects from minimum wage increases. Firstly, the prevalence of small businesses and tighter margins in these areas plays a crucial role (Chris Wheat and Stacey Chan, 2025). Lower-population counties often host a higher proportion of small businesses that typically operate with tighter profit margins and may possess less financial capacity to absorb mandated increases in labor costs compared to larger businesses often found in more populous, urbanized areas (Chris Wheat and Stacey Chan, 2025). Furthermore, businesses in smaller, potentially less competitive markets might find it more challenging to pass on increased labor costs to consumers via price hikes without substantially reducing demand, thereby compelling them to consider labor cost reductions more directly, including limiting teen employment. Secondly, the industry composition and economic diversification of less populous counties can exacerbate these effects (RATCLIFFE *et al.*, 2016). The industrial makeup might lean more heavily towards sectors that traditionally employ a larger share of minimum wage workers, such as agriculture, seasonal tourism, and small-scale retail or service establishments (RATCLIFFE *et al.*, 2016), which can be particularly sensitive to changes in wage floors (Economic Research Service, U.S. Department of Agriculture, 2025). In contrast, more populous areas frequently benefit from more diversified economies, featuring a broader mix of industries, including those where entry-level positions may already pay above the minimum wage (Economic Research Service, U.S. Department of Agriculture, 2025). Such economic diversity can cushion the aggregate impact of a minimum wage increase on overall teen employment.

The ability of DiD-BCF to uncover such covariate-driven heterogeneity, as seen with county population size, provides a richer and potentially more policy-relevant understanding than relying solely on average treatment effects. This application underscores the value of flexible, non-parametric approaches like DiD-BCF in applied econometric research.

4.7 Conclusion

This paper addressed these critical challenges of applying the DiD framework in real-world data by introducing the Difference-in-Differences Bayesian Causal Forest (DiD-BCF), a novel and robust framework for comprehensive DiD analysis. Our ap-

proach uniquely integrates the flexibility of Bayesian non-parametric modeling with a theoretically-grounded reparameterization strategy that leverages the Parallel Trends Assumption (PTA) to enhance estimation accuracy and stability. The DiD-BCF model provides a unified system for estimating Average Treatment Effects on the Treated (ATT), Group-Average Treatment Effects (GATT), and, critically, Conditional Average Treatment Effects on the Treated (CATT), across both classical and staggered DiD designs.

The empirical superiority and robustness of the DiD-BCF were rigorously demonstrated through extensive simulation studies. Across a diverse array of Data Generating Processes (DGPs)—encompassing varying degrees of non-linearity, selection mechanisms, and treatment effect heterogeneity—our proposed model consistently outperformed a suite of established benchmarks. Notably, DiD-BCF exhibited substantial gains in precision and lower treatment estimation error rates, particularly in scenarios mimicking complex observational data where other methods faltered. Even in simpler, correctly specified linear settings, DiD-BCF often matched or exceeded the performance of specialized linear estimators, underscoring the benefits of its principled regularization and PTA-based reparameterization which mitigates potential overfitting concerns. This consistent performance highlights the DiD-BCF's capacity to adapt to underlying data structures without sacrificing precision.

Furthermore, the inferential capabilities of DiD-BCF represent a critical advantage for applied research. Our analyses showed that the model not only provides high statistical power to detect treatment effects, especially in complex scenarios where other models struggle, but it does so while remaining well-calibrated. More importantly, it moves beyond the simple, binary question of whether an average effect exists. By enabling robust hypothesis testing for both group-level and conditional treatment effects, DiD-BCF empowers researchers to investigate the nuanced drivers of policy impacts. This dual capability—providing accurate estimates and facilitating detailed, reliable inference about effect heterogeneity—equips researchers with a more powerful and trustworthy lens to understand causal relationships in the real world.

The practical utility and distinctive capabilities of the DiD-BCF were further substantiated through an application to the salient policy question of minimum wage effects on teen employment. Beyond yielding an overall average effect estimate, our model uncovered significant and policy-relevant treatment effect heterogeneity conditional on county population size. Specifically, the findings indicated a more pronounced adverse employment impact in less populous counties. This granular insight, which aligns with existing economic literature suggesting differential impacts based on local economic conditions, would be obscured by traditional average effect estimators. This application underscores DiD-BCF's power not merely to estimate an average effect, but to reveal *how* and *for whom* an intervention's impact varies.

In conclusion, the DiD-BCF model represents a significant methodological advancement for applied causal inference. By adeptly handling staggered adoption, selection on observables, and heterogeneous treatment effects within a unified Bayesian framework, it offers researchers a more powerful, reliable, and nuanced tool than previously available. The ability to flexibly model complex outcome surfaces and treatment effect variations, validated through both simulation and real-world application, equips empirical researchers to draw more credible and fine-grained causal insights from DiD studies. As the demand for robust policy evaluation in complex settings grows, the DiD-BCF offers a promising path towards more precise and actionable evidence.

Supplementary Research Material and Code

The code used in this paper can be found in the following GitHub repository: Repository

CHAPTER

5

CONCLUSION

This Master's thesis embarked on a journey to address critical challenges in the estimation and evaluation of HTEs. The overarching ambition was to contribute to a more rigorous, reliable, and efficient landscape for causal inference, particularly in complex empirical settings where understanding nuanced treatment effect variation is paramount. Through a collection of three interconnected research papers, this work has presented methodological innovations, critical evaluations of existing practices, and novel tools designed to empower researchers in their pursuit of credible causal evidence. The research herein has sought not merely to introduce new techniques, but to fundamentally question and refine the processes by which causal claims are established and validated in the quantitative sciences.

The first core contribution, presented in Chapter 2, underscored the imperative of **critical model component evaluation** through the systematic use of ablation studies. By focusing on the BCF model and the role of the estimated propensity score $\hat{\boldsymbol{\pi}}(\mathbf{X})$, a component theoretically posited to mitigate RIC according to Hahn, Murray and Carvalho (2020), this research demonstrated that commonly accepted model components may not always enhance performance and can introduce unnecessary computational overhead. The finding that excluding $\hat{\boldsymbol{\pi}}(\mathbf{X})$ from the BCF model often leads to comparable, if not identical, treatment effect estimation accuracy and uncertainty quantification, while significantly reducing computational time by approximately 21%, carries important practical implications. It encourages a shift towards more parsimonious modeling, urging researchers to rigorously validate the necessity of each component in complex nonparametric estimators, particularly when such components add layers of estimation or computational burden. This work contributes to the causal inference literature by formally championing the use of ablation studies, a standard practice in other machine learning domains, within the specific context of causal effect estimation. This fosters a culture of critical model component evaluation and refinement, moving beyond reliance on theoretical justifica-

tions alone to demand empirical evidence of a component's utility in finite samples and realistic data scenarios. The implication is a potential re-evaluation of default settings in popular causal inference software and a call for greater transparency in reporting the contributions of individual model parts.

The second pillar of this thesis, detailed in Chapter 3, tackled prevalent methodological shortcomings in the **evaluation of novel treatment effect estimators via simulation studies**. Current practices often suffer from the arbitrary selection of simulation replications (J) and a lack of formal statistical comparisons, hindering reproducibility and potentially leading to underpowered or misleading conclusions regarding model superiority. To address this critical gap, TISCA was introduced. TISCA provides a statistically principled framework that integrates Welch's t-tests with power analysis, iteratively determining the optimal number of simulations required to achieve a desired statistical power for detecting a user-defined minimum detectable effect size. The bibliometric analysis presented within the chapter empirically confirmed the widespread heterogeneity and often ad-hoc nature of simulation counts in contemporary research. The case study, revisiting McJames *et al.* (2024), powerfully demonstrated TISCA's capacity to identify sufficient simulation counts (in that instance, half the number originally performed), thereby offering substantial savings in computational resources, namely 50 hours of simulations, without compromising statistical rigor. The broader implication of TISCA is the promotion of more transparent, efficient, and sustainable research practices. Its contribution lies in offering a concrete, algorithmic solution that enhances the credibility and comparability of simulation-based evaluations in causal inference and beyond, urging the field to move from convention-based to evidence-based simulation design. This has profound implications for the allocation of research resources and the trustworthiness of published comparative studies.

The third major contribution, elaborated in Chapter 4, introduced a **novel nonparametric estimator for Difference-in-Differences (DiD) settings**, a cornerstone methodology for policy evaluation and causal inference with panel data. The DiD-BCF model was developed to address contemporary challenges in DiD analysis, including staggered treatment adoption, selection on observables, and the nuanced estimation of CATE. A core innovation of DiD-BCF is its PTA-based reparameterization, a theoretically grounded modification designed to enhance estimation accuracy and stability in complex panel data by simplifying the functional form the treatment effect component must learn. Extensive simulation studies showcased DiD-BCF's superior performance over a suite of established benchmarks, particularly in scenarios characterized by non-linearity, selection biases, and treatment effect heterogeneity—conditions frequently encountered in real-world data. Furthermore, its application to U.S. minimum wage policy data successfully uncovered significant conditional treatment effect heterogeneity related to county population, an insight often masked by traditional, more restrictive DiD methods. The

DiD-BCF model offers applied researchers a robust and versatile tool for more nuanced causal inference, extending the power of Bayesian non-parametric methods to the increasingly complex landscape of DiD applications. This research contributes a significant methodological advancement to the DiD toolkit, enabling more credible and fine-grained causal insights from observational panel data. It specifically addresses the growing need for estimators that can flexibly handle the intricacies of modern DiD designs while providing interpretable HTEs.

Synthesizing these individual contributions, this thesis collectively champions a paradigm of **rigor, efficiency, and nuanced understanding in causal inference**. A common thread weaving through all chapters is the emphasis on moving beyond conventional or default approaches towards more critically evaluated, statistically justified, and context-adaptable methodologies. Whether it involves scrutinizing the components of existing models, refining the design of evaluation studies, or developing new estimators for challenging data structures, the underlying objective has been to enhance the reliability and practical utility of tools available for HTE estimation. This work argues for a reflective practice in quantitative causal analysis, where the assumptions underpinning our models and evaluation strategies are as rigorously interrogated as the substantive research questions themselves.

The implications of this body of work extend to various stakeholders in the research ecosystem. For methodologists, it highlights areas ripe for further refinement and innovation, such as the development of more adaptive regularization schemes or the exploration of alternative statistical foundations for simulation design. For applied researchers, it offers practical tools and frameworks that can lead to more credible, computationally feasible, and insightful findings from their data, enabling them to answer more complex questions with greater confidence. For the broader scientific community, it advocates for standards that can improve the overall quality, reproducibility, and resource stewardship of causal inference research, contributing to a more robust accumulation of knowledge.

Looking forward, several avenues for future research emerge from this thesis, promising to build upon its foundations. The principles of ablation studies championed in Chapter 2 can and should be extended to a wider array of complex causal estimators, including those based on deep learning (e.g., TNet (CURTH; SCHAAR, 2021), TARNet (SHALIT; JOHANSSON; SONTAG, 2017), DragonNet (SHI; BLEI; VEITCH, 2019), DR-CFR (HASSANPOUR; GREINER, 2020), and SNet (CURTH; SCHAAR, 2021)), and to other model components beyond propensity scores, such as specific network architectures or regularization terms. Further work could also focus on developing automated or semi-automated frameworks for conducting such studies, making them more accessible. For TISCA, introduced in Chapter {chapter:TISCA}, future extensions could include incorporating a broader range of statistical tests suitable for diverse performance metrics,

developing more sophisticated and interactive methods for MDE elicitation, and further developing of methods to deal with multiple statistical testing. The DiD-BCF model from Chapter 4 opens doors for further exploration into non-parametric DiD. This includes, but is not limited to, incorporating more complex dynamic treatment effect patterns, developing robust methods for addressing violations of the parallel trends assumption more directly within the Bayesian framework (perhaps through sensitivity analyses or incorporating auxiliary data), or extending the model to settings with multiple or continuous treatments, or even interference between units. Additionally, the computational aspects of all proposed methods, particularly their scalability to very large datasets (big data) and their implementation in user-friendly software packages, remain a critical area for ongoing optimization and development. The intersection of these methods with issues of fairness, accountability, and transparency in algorithmic decision-making also presents a vital frontier.

In conclusion, this Master's thesis has sought to make meaningful and actionable contributions to the evolving field of heterogeneous treatment effect estimation. By critically evaluating existing methods, proposing novel algorithmic solutions for study design, and developing advanced non-parametric estimators, this work aims to equip the research community with enhanced capabilities to navigate the complexities of causal inference. It is hoped that the methodologies and insights presented herein will foster more robust, reliable, and insightful research, ultimately leading to a deeper understanding of causal relationships and more effective, equitable, and evidence-based decision-making across a wide spectrum of disciplines. The pursuit of causal truth is an ongoing endeavor, and this thesis represents a dedicated step towards refining the tools and practices essential for that pursuit.

BIBLIOGRAPHY

ABADIE, A. Semiparametric difference-in-differences estimators. **The Review of Economic Studies**, Oxford University Press (OUP), v. 72, n. 1, p. 1–19, Jan. 2005. ISSN 0034-6527. Available: <http://dx.doi.org/10.1111/0034-6527.00321>. Citation on page 90.

ABADIE, A.; DIAMOND, A.; HAINMUELLER, J. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. **Journal of the American Statistical Association**, Informa UK Limited, v. 105, n. 490, p. 493–505, Jun. 2010. ISSN 1537-274X. Available: <http://dx.doi.org/10.1198/jasa.2009.ap08746>. Citations on pages 55 and 56.

ADHIKARI, A.; RAM, A.; TANG, R.; LIN, J. Rethinking complex neural network architectures for document classification. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4046–4051. Available: <https://aclanthology.org/N19-1408>. Citation on page 34.

ALCANTARA, R.; WANG, M.; HAHN, P. R.; LOPES, H. **Modified BART for Learning Heterogeneous Effects in Regression Discontinuity Designs**. arXiv, 2024. Available: <https://arxiv.org/abs/2407.14365>. Citations on pages 17, 93, 94, 166, and 167.

ANGRIST, J. D.; IMBENS, G. W.; RUBIN, D. B. Identification of causal effects using instrumental variables. **Journal of the American Statistical Association**, Informa UK Limited, v. 91, n. 434, p. 444–455, Jun. 1996. ISSN 1537-274X. Available: <http://dx.doi.org/10.1080/01621459.1996.10476902>. Citation on page 84.

ARKHANGELSKY, D.; ATHEY, S.; HIRSHBERG, D. A.; IMBENS, G. W.; WAGER, S. Synthetic difference-in-differences. **American Economic Review**, American Economic Association, v. 111, n. 12, p. 4088–4118, Dec. 2021. ISSN 0002-8282. Available: <http://dx.doi.org/10.1257/aer.20190159>. Citations on pages 86, 88, 96, 97, and 107.

ATHEY, S.; IMBENS, G. W. Design-based analysis in difference-in-differences settings with staggered adoption. **Journal of Econometrics**, Elsevier BV, v. 226, n. 1, p. 62–79, Jan. 2022. ISSN 0304-4076. Available: <http://dx.doi.org/10.1016/j.jeconom.2020.10.012>. Citation on page 84.

ATHEY, S.; WAGER, S. Estimating treatment effects with causal forests: An application. **Observational Studies**, Project MUSE, v. 5, n. 2, p. 37–51, 2019. ISSN 2767-3324. Available: <http://dx.doi.org/10.1353/obs.2019.0001>. Citations on pages 87 and 89.

BACH, P.; CHERNOZHUKOV, V.; KURZ, M. S.; SPINDLER, M. DoubleML – An object-oriented implementation of double machine learning in Python. **Journal of Ma-**

chine Learning Research, v. 23, n. 53, p. 1–6, 2022. Available: <http://jmlr.org/papers/v23/21-0862.html>. Citations on pages 97 and 107.

BACH, P.; CHERNOZHUKOV, V.; SPINDLER, M. **Closing the U.S. gender wage gap requires understanding its heterogeneity**. arXiv, 2018. Available: <https://arxiv.org/abs/1812.04345>. Citations on pages 55 and 56.

BAIL, C. A.; GUAY, B.; MALONEY, E.; COMBS, A.; HILLYGUS, D. S.; MERHOUT, F.; FREELON, D.; VOLFOVSKY, A. Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 117, n. 1, p. 243–250, Nov. 2019. ISSN 1091-6490. Available: <http://dx.doi.org/10.1073/pnas.1906420116>. Citations on pages 34 and 56.

BAIOCCHI, M.; CHENG, J.; SMALL, D. S. Instrumental variable methods for causal inference: Instrumental variable methods for causal inference. **Statistics in Medicine**, Wiley, v. 33, n. 13, p. 2297–2340, Mar. 2014. ISSN 0277-6715. Available: <http://dx.doi.org/10.1002/sim.6128>. Citation on page 84.

BALLINARI, D.; BEARTH, N. **Improving the Finite Sample Performance of Double/Debiased Machine Learning with Propensity Score Calibration**. arXiv, 2024. Available: <https://arxiv.org/abs/2409.04874>. Citations on pages 35 and 37.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press (OUP), v. 57, n. 1, p. 289–300, Jan. 1995. ISSN 1467-9868. Available: <http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x>. Citations on pages 72 and 80.

BIANCHINI, M.; SCARSELLI, F. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. **IEEE Transactions on Neural Networks and Learning Systems**, v. 25, n. 8, p. 1553–1565, 2014. Citation on page 34.

BLUME, J. D.; MCGOWAN, L. D.; DUPONT, W. D.; GREEVY, R. A. Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. **PLOS ONE**, Public Library of Science (PLoS), v. 13, n. 3, p. e0188299, Mar. 2018. ISSN 1932-6203. Available: <http://dx.doi.org/10.1371/journal.pone.0188299>. Citation on page 107.

BOR, J.; MOSCOE, E.; BäRNIGHAUSEN, T. Three approaches to causal inference in regression discontinuity designs. **Epidemiology**, Ovid Technologies (Wolters Kluwer Health), v. 26, n. 2, p. e28–e30, Mar. 2015. ISSN 1044-3983. Available: <http://dx.doi.org/10.1097/EDE.0000000000000256>. Citation on page 84.

BOR, J.; MOSCOE, E.; MUTEVEDZI, P.; NEWELL, M.-L.; BäRNIGHAUSEN, T. Regression discontinuity designs in epidemiology: Causal inference without randomized trials. **Epidemiology**, Ovid Technologies (Wolters Kluwer Health), v. 25, n. 5, p. 729–737, Sep. 2014. ISSN 1044-3983. Available: <http://dx.doi.org/10.1097/EDE.0000000000000138>. Citation on page 84.

BOWDEN, J.; BORNKAMP, B.; GLIMM, E.; BRETZ, F. Connecting instrumental variable methods for causal inference to the estimand framework. **Statistics in Medicine**, Wiley, v. 40, n. 25, p. 5605–5627, Jul. 2021. ISSN 1097-0258. Available: <http://dx.doi.org/10.1002/sim.9143>. Citation on page 84.

BOYER, C. B.; DAHABREH, I. J.; STEINGRIMSSON, J. A. **Estimating and evaluating counterfactual prediction models**. arXiv, 2023. Available: <https://arxiv.org/abs/2308.13026>. Citation on page 84.

BROSAMLER, G. A. An almost everywhere central limit theorem. **Mathematical Proceedings of the Cambridge Philosophical Society**, Cambridge University Press (CUP), v. 104, n. 3, p. 561–574, Nov. 1988. ISSN 1469-8064. Available: <http://dx.doi.org/10.1017/S0305004100065750>. Citation on page 67.

BROWN, C.; GILROY, C.; KOHEN, A. **Time-Series Evidence of the Effect of the Minimum Wage on Youth Employment and Unemployment**. [s.n.], 1981. Available: <http://dx.doi.org/10.3386/w0790>. Citation on page 122.

BUTTS, K.; GARDNER, J.; MCDERMOTT, G.; BERGE, L. **did2s: Two-Stage Difference-in-Differences Following Gardner (2021)**. [S.l.], 2023. R package version 1.0.2. Available: <https://kylebutts.github.io/did2s/>. Citations on pages 88 and 96.

CALLAWAY, B.; SANT'ANNA, P. H. C. **did: Treatment Effects with Multiple Periods and Groups**. [S.l.], 2022. R package version 2.1.2. Available: <https://bcallaway11.github.io/did/,https://github.com/bcallaway11/did/>. Citations on pages 96 and 122.

CALLAWAY, B.; SANT'ANNA, P. H. Difference-in-differences with multiple time periods. **Journal of Econometrics**, Elsevier BV, v. 225, n. 2, p. 200–230, Dec. 2021. ISSN 0304-4076. Available: <http://dx.doi.org/10.1016/j.jeconom.2020.12.001>. Citations on pages 29, 84, 86, 88, and 89.

CANIGLIA, E. C.; MURRAY, E. J. Difference-in-difference in the time of cholera: a gentle introduction for epidemiologists. **Current Epidemiology Reports**, Springer Science and Business Media LLC, v. 7, n. 4, p. 203–211, Sep. 2020. ISSN 2196-2995. Available: <http://dx.doi.org/10.1007/s40471-020-00245-2>. Citation on page 84.

CAO, D.; WANG, Y.; DUAN, J.; ZHANG, C.; ZHU, X.; HUANG, C.; TONG, Y.; XU, B.; BAI, J.; TONG, J.; ZHANG, Q. Spectral temporal graph neural network for multivariate time-series forecasting. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 17766–17778. Available: <https://proceedings.neurips.cc/paper_files/paper/2020/file/cdf6581cb7aca4b7e19ef136c6e601a5-Paper.pdf>. Citation on page 34.

CARD, D. The causal effect of education on earnings. In: _____. **Handbook of Labor Economics**. Elsevier, 1999. p. 1801–1863. Available: <http://dx.doi.org/10.1016/S1573-4463(99)03011-4>. Citations on pages 55 and 56.

CARD, D.; KRUEGER, A. **Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania**. [s.n.], 1993. Available: <http://dx.doi.org/10.3386/w4509>. Citation on page 101.

CATTANEO, M. D.; TITIUNIK, R. Regression discontinuity designs. **Annual Review of Economics**, Annual Reviews, v. 14, n. 1, p. 821–851, Aug. 2022. ISSN 1941-1391. Available: <http://dx.doi.org/10.1146/annurev-economics-051520-021409>. Citation on page 84.

CAUSAL inference and data fusion in econometrics. **The Econometrics Journal**, Oxford University Press (OUP), v. 28, n. 1, p. 128–128, Apr. 2024. ISSN 1368-423X. Available: <http://dx.doi.org/10.1093/ectj/utae008>. Citations on pages 29 and 83.

CHAISEMARTIN, C. de; D'HAULTFœUILLE, X. Two-way fixed effects estimators with heterogeneous treatment effects. **American Economic Review**, American Economic Association, v. 110, n. 9, p. 2964–2996, Sep. 2020. ISSN 0002-8282. Available: <http://dx.doi.org/10.1257/aer.20181169>. Citations on pages 86 and 88.

CHANG, N.-C. Double/debiased machine learning for difference-in-differences models. **The Econometrics Journal**, Oxford University Press (OUP), v. 23, n. 2, p. 177–191, Feb. 2020. ISSN 1368-423X. Available: <http://dx.doi.org/10.1093/ectj/utaa001>. Citation on page 97.

CHERNOZHUKOV, V.; HANSEN, C.; KALLUS, N.; SPINDLER, M.; SYRGKANIS, V. **Applied Causal Inference Powered by ML and AI**. [S.l.]: Online, 2024. Citation on page 29.

CHIPMAN, H. A.; GEORGE, E. I.; MCCULLOCH, R. E. Bart: Bayesian additive regression trees. **The Annals of Applied Statistics**, Institute of Mathematical Statistics, v. 4, n. 1, Mar. 2010. ISSN 1932-6157. Available: <http://dx.doi.org/10.1214/09-AOAS285>. Citations on pages 36, 56, 74, 92, 93, 151, 153, 154, 156, and 157.

CHIRINKO, R. S.; WILSON, D. J. State investment tax incentives: A zero-sum game? **Journal of Public Economics**, Elsevier BV, v. 92, n. 12, p. 2362–2384, Dec. 2008. ISSN 0047-2727. Available: <http://dx.doi.org/10.1016/j.jpubeco.2008.07.005>. Citation on page 103.

Chris Wheat and Stacey Chan. **The rural divide: small business revenue milestones in the U.S.** [S.l.], 2025. Available: <https://www.jpmorganchase.com/institute/all-topics/business-growth-and-entrepreneurship/the-rural-divide-small-business-revenue-milestones-in-the-us>. Citation on page 125.

CINTRON, D. W.; ADLER, N. E.; GOTTLIEB, L. M.; HAGAN, E.; TAN, M. L.; VLAHOV, D.; GLYMOUR, M. M.; MATTHAY, E. C. Heterogeneous treatment effects in social policy studies: An assessment of contemporary articles in the health and social sciences. **Annals of Epidemiology**, Elsevier BV, v. 70, p. 79–88, Jun. 2022. ISSN 1047-2797. Available: <http://dx.doi.org/10.1016/j.annepidem.2022.04.009>. Citations on pages 29 and 87.

CORNEJO, F.; FRANCHINI, N.; CORTéS, B. I.; ELGUETA, D.; CANCINO, G. I. Neural conditional ablation of the protein tyrosine phosphatase receptor delta ptprd impairs gliogenesis in the developing mouse brain cortex. **Frontiers in Cell and Developmental Biology**, Frontiers Media SA, v. 12, Feb. 2024. ISSN 2296-634X. Available: <http://dx.doi.org/10.3389/fcell.2024.1357862>. Citation on page 39.

CURTH, A.; SCHAAR, M. van der. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In: BANERJEE, A.; FUKUMIZU, K. (Ed.). **Proceedings of The 24th International Conference on Artificial Intelligence and Statistics**. PMLR, 2021. (Proceedings of Machine Learning Research, v. 130), p. 1810–1818. Available: <https://proceedings.mlr.press/v130/curth21a.html>. Citations on pages 29, 56, 57, 92, 93, and 131.

CURTH, A.; SVENSSON, D.; WEATHERALL, J.; SCHAAR, M. van der. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In: VANSCHOREN, J.; YEUNG, S. (Ed.). **Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks**. Curran, 2021. v. 1. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/2a79ea27c279e471f4d180b08d62b00a-Paper-round2.pdf>. Citations on pages 29, 56, and 57.

DAI, X. Nonparametric estimation via partial derivatives. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press (OUP), v. 87, n. 2, p. 319–336, Sep. 2024. ISSN 1467-9868. Available: <http://dx.doi.org/10.1093/jrsssb/qkae093>. Citation on page 90.

DEAN, A.; VOSS, D. Fractional factorial experiments. In: _____. **Design and Analysis of Experiments**. Springer-Verlag, 1999. p. 483–545. ISBN 0387985611. Available: <http://dx.doi.org/10.1007/0-387-22634-6_15>. Citation on page 67.

DEB, N.; MUKHERJEE, D. **Trade-off Between Dependence and Complexity for Nonparametric Learning – an Empirical Process Approach**. arXiv, 2024. Available: <https://arxiv.org/abs/2401.08978>. Citations on pages 90 and 91.

DOUTRELIGNE, M.; VAROQUAUX, G. How to select predictive models for decision-making or causal inference. **GigaScience**, Oxford University Press (OUP), v. 14, 2025. ISSN 2047-217X. Available: <http://dx.doi.org/10.1093/gigascience/giaf016>. Citation on page 84.

DU, L. How much deep learning does neural style transfer really need? an ablation study. In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. [S.l.: s.n.], 2020. Citation on page 34.

DUDLEY, R. M. Central limit theorems for empirical measures. **The Annals of Probability**, JSTOR, p. 899–929, 1978. Citation on page 67.

DUNN, O. J. Multiple comparisons among means. **Journal of the American Statistical Association**, Informa UK Limited, v. 56, n. 293, p. 52–64, Mar. 1961. ISSN 1537-274X. Available: <http://dx.doi.org/10.1080/01621459.1961.10482090>. Citations on pages 72 and 80.

Economic Research Service, U.S. Department of Agriculture. **County Typology Codes - Descriptions and Maps**. Economic Research Service, U.S. DEPARTMENT OF AGRICULTURE, 2025. Updated: 4/11/2025; Contact: Austin Sanders. Available: <https://www.ers.usda.gov/data-products/county-typology-codes/descriptions-and-maps>. Citation on page 125.

ELLMAN, D. G.; LUND, M. C.; NISSEN, M.; NIELSEN, P. S.; SøRENSEN, C.; LESTER, E. B.; THOUGAARD, E.; JøRGENSEN, L. H.; NEDOSPASOV, S. A.; ANDERSEN, D. C.; STUBBE, J.; BRAMBILLA, R.; DEGN, M.; LAMBERTSEN, K. L. Conditional ablation of myeloid tnf improves functional outcome and decreases lesion size after spinal cord injury in mice. **Cells**, MDPI AG, v. 9, n. 11, p. 2407, Nov. 2020. ISSN 2073-4409. Available: <http://dx.doi.org/10.3390/cells9112407>. Citation on page 39.

EWALD, F. K.; BOTHMANN, L.; WRIGHT, M. N.; BISCHL, B.; CASALICCHIO, G.; KöNIG, G. A guide to feature importance methods for scientific inference. In: _____. **Explainable Artificial Intelligence**. Springer Nature Switzerland, 2024. p. 440–464. ISBN 9783031637971. Available: <http://dx.doi.org/10.1007/978-3-031-63797-1_22>. Citation on page 39.

FAGERLAND, M. W. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? **BMC Medical Research Methodology**, Springer Science and Business Media LLC, v. 12, n. 1, Jun. 2012. ISSN 1471-2288. Available: <http://dx.doi.org/10.1186/1471-2288-12-78>. Citation on page 41.

FENG, Z.; PROSPERI, M.; GUO, Y.; BIAN, J. Variational temporal deconfounder for individualized treatment effect estimation with longitudinal observational data. Research Square Platform LLC, Feb. 2023. Available: <http://dx.doi.org/10.21203/rs.3.rs-2536079/v1>. Citation on page 84.

FIGLIO, D. N. Short-term effects of a 1990s-era property tax limit: Panel evidence on oregon's measure 5. **National Tax Journal**, University of Chicago Press, v. 51, n. 1, p. 55–70, Mar. 1998. ISSN 1944-7477. Available: <http://dx.doi.org/10.1086/NTJ41789311>. Citation on page 104.

FREYALDENHOVEN, S.; HANSEN, C.; PéREZ, J. P.; SHAPIRO, J. M. **Visualization, Identification, and Estimation in the Linear Panel Event-Study Design**. [S.l.], 2021. (Working Paper Series, 29170). Available: <http://www.nber.org/papers/w29170>. Citation on page 84.

FRIEDRICH, S.; GROLL, A.; ICKSTADT, K.; KNEIB, T.; PAULY, M.; RAHNENFüHRER, J.; FRIEDE, T. Regularization approaches in clinical biostatistics: A review of methods and their applications. **Statistical Methods in Medical Research**, SAGE Publications, v. 32, n. 2, p. 425–440, Nov. 2022. ISSN 1477-0334. Available: <http://dx.doi.org/10.1177/09622802221133557>. Citation on page 91.

GANGL, M. Causal inference in sociological research. **Annual Review of Sociology**, Annual Reviews, v. 36, n. 1, p. 21–47, Jun. 2010. ISSN 1545-2115. Available: <http://dx.doi.org/10.1146/annurev.soc.012809.102702>. Citations on pages 29 and 83.

GANNON, M. A.; PEREIRA, C. A. de B.; POLPO, A. Blending bayesian and classical tools to define optimal sample-size-dependent significance levels. **The American Statistician**, Informa UK Limited, v. 73, n. sup1, p. 213–222, Mar. 2019. ISSN 1537-2731. Available: <http://dx.doi.org/10.1080/00031305.2018.1518268>. Citation on page 107.

GARDNER, J. **Two-stage differences in differences**. arXiv, 2022. Available: <https://arxiv.org/abs/2207.05943>. Citations on pages 84, 88, 96, and 107.

GOLDKUHLE, M.; HIRSCH, C.; IANNIZZI, C.; ZORGER, A.-M.; BENDER, R.; DALEN, E. C. van; HEMKENS, L. G.; MONSEF, I.; KREUZBERGER, N.; SKOETZ, N. Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: a meta-epidemiological study. **BMC Medical Research Methodology**, Springer Science and Business Media LLC, v. 24, n. 1, Nov. 2024. ISSN 1471-2288. Available: <http://dx.doi.org/10.1186/s12874-024-02401-4>. Citations on pages 59 and 80.

GOMEZ, A. N.; REN, M.; URTASUN, R.; GROSSE, R. B. The reversible residual network: Backpropagation without storing activations. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Available: <https://proceedings.neurips.cc/paper_files/paper/2017/file/f9be311e65d81a9ad8150a60844bb94c-Paper.pdf>. Citation on page 34.

GOODMAN-BACON, A. Difference-in-differences with variation in treatment timing. **Journal of Econometrics**, Elsevier BV, v. 225, n. 2, p. 254–277, Dec. 2021. ISSN 0304-4076. Available: <http://dx.doi.org/10.1016/j.jeconom.2021.03.014>. Citation on page 86.

GREENLAND, S.; SENN, S. J.; ROTHMAN, K. J.; CARLIN, J. B.; POOLE, C.; GOODMAN, S. N.; ALTMAN, D. G. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. **European Journal of Epidemiology**, Springer Science and Business Media LLC, v. 31, n. 4, p. 337–350, Apr. 2016. ISSN 1573-7284. Available: <http://dx.doi.org/10.1007/s10654-016-0149-3>. Citations on pages 59 and 107.

GREENSTONE, M. The impacts of environmental regulations on industrial activity: Evidence from the 1970 and 1977 clean air act amendments and the census of manufactures. **Journal of Political Economy**, University of Chicago Press, v. 110, n. 6, p. 1175–1219, Dec. 2002. ISSN 1537-534X. Available: <http://dx.doi.org/10.1086/342808>. Citation on page 105.

GRIMMER, J. We are all social scientists now: How big data, machine learning, and causal inference work together. **PS: Political Science & Politics**, Cambridge University Press (CUP), v. 48, n. 01, p. 80–83, Dec. 2014. ISSN 1537-5935. Available: <http://dx.doi.org/10.1017/S1049096514001784>. Citations on pages 29 and 83.

GROENIGER, J. O.; NOORDZIJ, K.; WAAL, J. van der; KOSTER, W. de. Dutch covid-19 lockdown measures increased trust in government and trust in science: A difference-in-differences analysis. **Social Science & Medicine**, Elsevier BV, v. 275, p. 113819, Apr. 2021. ISSN 0277-9536. Available: <http://dx.doi.org/10.1016/j.socscimed.2021.113819>. Citation on page 84.

HAHN, P. R.; CARVALHO, C. M.; PUELZ, D.; HE, J. Regularization and confounding in linear regression for treatment effect estimation. **Bayesian Analysis**, Institute of Mathematical Statistics, v. 13, n. 1, Mar. 2018. ISSN 1936-0975. Available: <http://dx.doi.org/10.1214/16-BA1044>. Citations on pages 35 and 93.

HAHN, P. R.; MURRAY, J. S.; CARVALHO, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). **Bayesian Analysis**, Institute of Mathematical Statistics, v. 15, n. 3, Sep. 2020. ISSN

1936-0975. Available: <http://dx.doi.org/10.1214/19-BA1195>. Citations on pages 29, 34, 35, 36, 37, 38, 39, 42, 49, 52, 53, 56, 57, 67, 74, 87, 93, 129, 161, 162, and 163.

HAMMERTON, G.; MUNAFò, M. R. Causal inference with observational data: the need for triangulation of evidence. **Psychological Medicine**, Cambridge University Press (CUP), v. 51, n. 4, p. 563–578, Mar. 2021. ISSN 1469-8978. Available: <http://dx.doi.org/10.1017/S0033291720005127>. Citation on page 84.

HANSEN, P. R.; LUNDE ASGER AMD NASON, J. M. The model confidence set. **Econometrica**, The Econometric Society, v. 79, n. 2, p. 453–497, 2011. ISSN 0012-9682. Available: <http://dx.doi.org/10.3982/ECTA5771>. Citations on pages 58 and 59.

HASANPOUR, S. H.; ROUHANI, M.; FAYYAZ, M.; SABOKROU, M. **Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures**. arXiv, 2016. Available: <https://arxiv.org/abs/1608.06037>. Citation on page 34.

HASSANPOUR, N.; GREINER, R. Learning disentangled representations for counterfactual regression. In: **International Conference on Learning Representations**. [s.n.], 2020. Available: <https://openreview.net/forum?id=HkxBJT4YvB>. Citations on pages 34, 56, and 131.

HATAMYAR, J.; KREIF, N.; ROCHA, R.; HUBER, M. **Machine Learning for Staggered Difference-in-Differences and Dynamic Treatment Effect Heterogeneity**. arXiv, 2023. Available: <https://arxiv.org/abs/2310.11962>. Citations on pages 15, 87, 89, 97, and 99.

HE, J.; HAHN, P. R. Stochastic tree ensembles for regularized nonlinear regression. **Journal of the American Statistical Association**, Informa UK Limited, v. 118, n. 541, p. 551–570, Aug. 2021. ISSN 1537-274X. Available: <http://dx.doi.org/10.1080/01621459.2021.1942012>. Citations on pages 93, 95, 166, and 167.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. Citation on page 34.

HECKMAN, J.; ICHIMURA, H.; SMITH, J.; TODD, P. Characterizing selection bias using experimental data. **Econometrica**, JSTOR, v. 66, n. 5, p. 1017, Sep. 1998. ISSN 0012-9682. Available: <http://dx.doi.org/10.2307/2999630>. Citation on page 90.

HECKMAN, J. J.; ICHIMURA, H.; TODD, P. E. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. **The Review of Economic Studies**, Oxford University Press (OUP), v. 64, n. 4, p. 605–654, Oct. 1997. ISSN 1467-937X. Available: <http://dx.doi.org/10.2307/2971733>. Citation on page 90.

HEDGES, L. V.; PIGOTT, T. D. The power of statistical tests in meta-analysis. **Psychological Methods**, American Psychological Association (APA), v. 6, n. 3, p. 203–217, 2001. ISSN 1082-989X. Available: <http://dx.doi.org/10.1037/1082-989X.6.3.203>. Citation on page 59.

HERNáN, M. A.; ROBINS, J. M. Instruments for causal inference: An epidemiologist's dream? **Epidemiology**, Ovid Technologies (Wolters Kluwer Health), v. 17, n. 4,

p. 360–372, Jul. 2006. ISSN 1044-3983. Available: <http://dx.doi.org/10.1097/01.ede.0000222409.00878.37>. Citation on page 84.

HILL, J. L. Bayesian nonparametric modeling for causal inference. **Journal of Computational and Graphical Statistics**, Informa UK Limited, v. 20, n. 1, p. 217–240, Jan. 2011. ISSN 1537-2715. Available: <http://dx.doi.org/10.1198/jcgs.2010.08162>. Citations on pages 34 and 57.

HITSCH, G. J.; MISRA, S.; ZHANG, W. W. Heterogeneous treatment effects and optimal targeting policy evaluation. **Quantitative Marketing and Economics**, Springer Science and Business Media LLC, v. 22, n. 2, p. 115–168, Apr. 2024. ISSN 1573-711X. Available: <http://dx.doi.org/10.1007/s11129-023-09278-5>. Citations on pages 29 and 87.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric Statistical Methods**. Wiley, 2015. ISSN 1940-6347. ISBN 9781119196037. Available: <http://dx.doi.org/10.1002/9781119196037>. Citations on pages 58 and 59.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian journal of statistics**, JSTOR, p. 65–70, 1979. Citations on pages 72 and 80.

HOOVER, K. D. The logic of causal inference: Econometrics and the conditional analysis of causation. **Economics and Philosophy**, Cambridge University Press (CUP), v. 6, n. 2, p. 207–234, Oct. 1990. ISSN 1474-0028. Available: <http://dx.doi.org/10.1017/S026626710000122X>. Citations on pages 29 and 83.

HOROWITZ, J. L. **Semiparametric and Nonparametric Methods in Econometrics**. Springer US, 2009. ISSN 0172-7397. ISBN 9780387928708. Available: <http://dx.doi.org/10.1007/978-0-387-92870-8>. Citation on page 90.

IMAI, K.; KIM, I. S.; WANG, E. H. Matching methods for causal inference with time-series cross-sectional data. **American Journal of Political Science**, Wiley, v. 67, n. 3, p. 587–605, Dec. 2021. ISSN 1540-5907. Available: <http://dx.doi.org/10.1111/ajps.12685>. Citation on page 84.

IMBENS, G. W. Causal inference in the social sciences. **Annual Review of Statistics and Its Application**, Annual Reviews, v. 11, n. 1, p. 123–152, Apr. 2024. ISSN 2326-831X. Available: <http://dx.doi.org/10.1146/annurev-statistics-033121-114601>. Citations on pages 29 and 83.

JENNISON, C.; TURNBULL, B. W. **Group sequential methods with applications to clinical trials**. [S.l.]: CRC Press, 2000. Citation on page 73.

KALENKOSKI, C. M.; LACOMBE, D. J. Minimum wages and teen employment: A spatial panel approach. **Papers in Regional Science**, Elsevier BV, v. 92, n. 2, p. 407–418, Jun. 2013. ISSN 1056-8190. Available: <http://dx.doi.org/10.1111/j.1435-5957.2012.00453.x>. Citation on page 124.

KALLUS, N. Generalized optimal matching methods for causal inference. **Journal of Machine Learning Research**, v. 21, n. 62, p. 1–54, 2020. Available: <http://jmlr.org/papers/v21/19-120.html>. Citation on page 84.

KATTENBERG, M.; SCHEER, B.; THIEL, J. **Causal forests with fixed effects for treatment effect heterogeneity in difference-in-differences**. CPB Netherlands Bureau for Economic Policy Analysis, 2023. Available: <http://www.cpb.nl/en/causal-forests-fixed-effects-treatment-effect-heterogeneity-difference-differences>. Citations on pages 15, 87, 89, 97, and 98.

KEOGH, R. H.; GELOVEN, N. V. Prediction under interventions: Evaluation of counterfactual performance using longitudinal observational data. **Epidemiology**, Ovid Technologies (Wolters Kluwer Health), v. 35, n. 3, p. 329–339, Apr. 2024. ISSN 1044-3983. Available: <http://dx.doi.org/10.1097/EDE.0000000000001713>. Citation on page 84.

KING, G.; NIELSEN, R.; COBERLEY, C.; POPE, J. E.; WELLS, A. Comparative effectiveness of matching methods for causal inference. **Unpublished manuscript, Institute for Quantitative Social Science, Harvard University, Cambridge, MA**, 2011. Citation on page 84.

KLAASSEN, C. A. J.; PUTTER, H. Efficient estimation of banach parameters in semiparametric models. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 33, n. 1, Feb. 2005. ISSN 0090-5364. Available: <http://dx.doi.org/10.1214/009053604000000913>. Citations on pages 90 and 91.

KRANTSEVICH, N.; HE, J.; HAHN, P. R. Stochastic tree ensembles for estimating heterogeneous effects. In: RUIZ, F.; DY, J.; MEENT, J.-W. van de (Ed.). **Proceedings of The 26th International Conference on Artificial Intelligence and Statistics**. PMLR, 2023. (Proceedings of Machine Learning Research, v. 206), p. 6120–6131. Available: <https://proceedings.mlr.press/v206/krantsevich23a.html>. Citations on pages 94, 95, 166, and 167.

KüNZEL, S. R.; SEKHON, J. S.; BICKEL, P. J.; YU, B. Metalearners for estimating heterogeneous treatment effects using machine learning. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 116, n. 10, p. 4156–4165, Feb. 2019. ISSN 1091-6490. Available: <http://dx.doi.org/10.1073/pnas.1804597116>. Citations on pages 92 and 93.

KWIATKOWSKI, D.; PHILLIPS, P. C.; SCHMIDT, P.; SHIN, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. **Journal of Econometrics**, Elsevier BV, v. 54, n. 1–3, p. 159–178, Oct. 1992. ISSN 0304-4076. Available: <http://dx.doi.org/10.1016/0304-4076(92)90104-Y>. Citation on page 67.

KöNIG, G.; MOLNAR, C.; BISCHL, B.; GROSSE-WENTRUP, M. Relative feature importance. In: **2020 25th International Conference on Pattern Recognition (ICPR)**. [S.l.: s.n.], 2021. p. 9318–9325. Citation on page 39.

LANNELONGUE, L.; GREALEY, J.; INOUYE, M. Green algorithms: Quantifying the carbon footprint of computation. **Advanced Science**, Wiley, v. 8, n. 12, May 2021. ISSN 2198-3844. Available: <http://dx.doi.org/10.1002/advs.202100707>. Citations on pages 79 and 80.

LEER, J. After the big bang: Estimating the effects of decentralization on educational outcomes in indonesia through a difference-in-differences analysis. **International Journal of Educational Development**, Elsevier BV, v. 49, p. 80–90, Jul. 2016. ISSN 0738-0593. Available: <http://dx.doi.org/10.1016/j.ijedudev.2016.02.005>. Citation on page 84.

LINDEN, A.; ADAMS, J. L. Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. **Journal of Evaluation in Clinical Practice**, Wiley, v. 18, n. 2, p. 317–325, Feb. 2012. ISSN 1365-2753. Available: <http://dx.doi.org/10.1111/j.1365-2753.2011.01768.x>. Citation on page 84.

LINDROOTH, R. C.; PERRAILLON, M. C.; HARDY, R. Y.; TUNG, G. J. Understanding the relationship between medicaid expansions and hospital closures. **Health Affairs**, Health Affairs (Project Hope), v. 37, n. 1, p. 111–120, Jan. 2018. ISSN 1544-5208. Available: <http://dx.doi.org/10.1377/hlthaff.2017.0976>. Citation on page 101.

LOWENSTEIN, M.; HOSSAIN, E.; YANG, W.; GRANDE, D.; PERRONE, J.; NEUMAN, M. D.; ASHBURN, M.; DELGADO, M. K. Impact of a state opioid prescribing limit and electronic medical record alert on opioid prescriptions: a difference-in-differences analysis. **Journal of General Internal Medicine**, Springer Science and Business Media LLC, v. 35, n. 3, p. 662–671, Oct. 2019. ISSN 1525-1497. Available: <http://dx.doi.org/10.1007/s11606-019-05302-1>. Citation on page 84.

MCJAMES, N.; O'SHEA, A.; GOH, Y. C.; PARNELL, A. Bayesian causal forests for multivariate outcomes: application to irish data from an international large scale education assessment. **Journal of the Royal Statistical Society Series A: Statistics in Society**, Oxford University Press (OUP), v. 188, n. 2, p. 428–450, May 2024. ISSN 1467-985X. Available: <http://dx.doi.org/10.1093/jrsssa/qnae049>. Citations on pages 15, 19, 25, 30, 55, 60, 74, 75, 76, 77, 78, 79, 80, and 130.

MENON, V. Multiple testing and protection against type i error using p value correction: Application in cross-sectional study designs. **Indian Journal of Psychological Medicine**, SAGE Publications, v. 41, n. 2, p. 197–197, Mar. 2019. ISSN 0975-1564. Available: <http://dx.doi.org/10.4103/IJPSYM.IJPSYM_12_19>. Citation on page 72.

MEYES, R.; LU, M.; PUISEAU, C. W. de; MEISEN, T. **Ablation Studies in Artificial Neural Networks**. arXiv, 2019. Available: <https://arxiv.org/abs/1901.08644>. Citation on page 34.

NEWEY, W. K. Semiparametric efficiency bounds. **Journal of Applied Econometrics**, Wiley, v. 5, n. 2, p. 99–135, Apr. 1990. ISSN 1099-1255. Available: <http://dx.doi.org/10.1002/jae.3950050202>. Citation on page 90.

NICHOLS, A. Causal inference with observational data. **The Stata Journal: Promoting communications on statistics and Stata**, SAGE Publications, v. 7, n. 4, p. 507–541, Dec. 2007. ISSN 1536-8734. Available: <http://dx.doi.org/10.1177/1536867X0800700403>. Citation on page 84.

OHLSSON, H.; KENDLER, K. S. Applying causal inference methods in psychiatric epidemiology: A review. **JAMA Psychiatry**, American Medical Association (AMA), v. 77, n. 6, p. 637, Jun. 2020. ISSN 2168-622X. Available: <http://dx.doi.org/10.1001/jamapsychiatry.2019.3758>. Citations on pages 29, 56, and 83.

OLDENBURG, C. E.; MOSCOE, E.; BäRNIGHAUSEN, T. Regression discontinuity for causal effect estimation in epidemiology. **Current Epidemiology Reports**, Springer Science and Business Media LLC, v. 3, n. 3, p. 233–241, Aug. 2016. ISSN 2196-2995. Available: <http://dx.doi.org/10.1007/s40471-016-0080-x>. Citation on page 84.

OLIVARES, K. G.; CHALLU, C.; MARCJASZ, G.; WERON, R.; DUBRAWSKI, A. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. **International Journal of Forecasting**, Elsevier BV, v. 39, n. 2, p. 884–900, Apr. 2023. ISSN 0169-2070. Available: <http://dx.doi.org/10.1016/j.ijforecast.2022.03.001>. Citation on page 34.

PENG, A.; FORDE, J. Z.; SHAVIT, Y.; FRANKLE, J. **Strengthening Subcommunities: Towards Sustainable Growth in AI Research**. arXiv, 2022. Available: <https://arxiv.org/abs/2204.08377>. Citation on page 59.

POOLE, C. Low p-values or narrow confidence intervals: which are more durable? **Epidemiology**, LWW, v. 12, n. 3, p. 291–294, 2001. Citation on page 107.

PRADO, E. a. B.; O'NEILL, E.; HERNÁNDEZ, B.; PARNELL, A. C.; MORAL, R. A. Contributed discussion. In: HAHN, P. R.; MURRAY, J. S.; CARVALHO, C. M. (Ed.). **Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)**. [s.n.], 2020. v. 15, n. 3, p. 1029–1031. A discussion contribution to the main article by Hahn, Murray, and Carvalho. Available: <http://dx.doi.org/10.1214/19-BA1195>. Citations on pages 15, 49, and 52.

PREL, J.-B. du; RöHRIG, B.; HOMMEL, G.; BLETTNER, M. Choosing statistical tests. **Deutsches Ärzteblatt international**, Deutscher Arzte-Verlag GmbH, May 2010. ISSN 1866-0452. Available: <http://dx.doi.org/10.3238/arztebl.2010.0343>. Citation on page 59.

RAMBACHAN, A.; ROTH, J. A more credible approach to parallel trends. **Review of Economic Studies**, Oxford University Press (OUP), v. 90, n. 5, p. 2555–2591, Feb. 2023. ISSN 1467-937X. Available: <http://dx.doi.org/10.1093/restud/rdad018>. Citation on page 91.

RASCH, D.; KUBINGER, K. D.; MODER, K. The two-sample t test: pre-testing its assumptions does not pay off. **Statistical Papers**, Springer Science and Business Media LLC, v. 52, n. 1, p. 219–231, Apr. 2009. ISSN 1613-9798. Available: <http://dx.doi.org/10.1007/s00362-009-0224-x>. Citation on page 67.

RATCLIFFE, M.; BURD, C.; HOLDER, K.; FIELDS, A. Defining rural at the us census bureau. **American community survey and geography brief**, US Department of Commerce Economics and Statistics Administration, US Census …, v. 1, n. 8, p. 1–8, 2016. Citation on page 125.

REHILL, P.; BIDDLE, N. **Fairness Implications of Heterogeneous Treatment Effect Estimation with Machine Learning Methods in Policy-making**. arXiv, 2023. Available: <https://arxiv.org/abs/2309.00805>. Citations on pages 29 and 87.

ROCHON, J.; GONDAN, M.; KIESER, M. To test or not to test: Preliminary assessment of normality when comparing two independent samples. **BMC Medical Research Methodology**, Springer Science and Business Media LLC, v. 12, n. 1, Jun. 2012. ISSN 1471-2288. Available: <http://dx.doi.org/10.1186/1471-2288-12-81>. Citations on pages 41 and 67.

ROTH, J. Pretest with caution: Event-study estimates after testing for parallel trends. **American Economic Review: Insights**, American Economic Association, v. 4, n. 3,

p. 305–322, Sep. 2022. ISSN 2640-2068. Available: <http://dx.doi.org/10.1257/aeri.20210236>. Citations on pages 84 and 86.

_____. **Interpreting Event-Studies from Recent Difference-in-Differences Methods**. arXiv, 2024. Available: <https://arxiv.org/abs/2401.12309>. Citation on page 84.

ROTHMAN, K. J.; GREENLAND, S. Causation and causal inference in epidemiology. **American Journal of Public Health**, American Public Health Association, v. 95, n. S1, p. S144–S150, Jul. 2005. ISSN 1541-0048. Available: <http://dx.doi.org/10.2105/AJPH.2004.059204>. Citations on pages 29, 56, and 83.

RUXTON, G. D. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. **Behavioral Ecology**, Oxford University Press (OUP), v. 17, n. 4, p. 688–690, May 2006. ISSN 1045-2249. Available: <http://dx.doi.org/10.1093/beheco/ark016>. Citation on page 41.

SALINAS, P.; SOLé-OLLé, A. Partial fiscal decentralization reforms and educational outcomes: A difference-in-differences analysis for spain. **Journal of Urban Economics**, Elsevier BV, v. 107, p. 31–46, Sep. 2018. ISSN 0094-1190. Available: <http://dx.doi.org/10.1016/j.jue.2018.08.003>. Citation on page 84.

SANT'ANNA, P. H.; ZHAO, J. Doubly robust difference-in-differences estimators. **Journal of Econometrics**, Elsevier BV, v. 219, n. 1, p. 101–122, Nov. 2020. ISSN 0304-4076. Available: <http://dx.doi.org/10.1016/j.jeconom.2020.06.003>. Citations on pages 84, 86, 90, 96, 97, 107, 122, 123, and 124.

SCHUCANY, W. R.; NG, H. K. T. Preliminary goodness-of-fit tests for normality do not validate the one-sample studentt. **Communications in Statistics - Theory and Methods**, Informa UK Limited, v. 35, n. 12, p. 2275–2286, Dec. 2006. ISSN 1532-415X. Available: <http://dx.doi.org/10.1080/03610920600853308>. Citations on pages 41 and 67.

SEN, A.; RYBCZYNSKI, K.; WAAL, C. V. D. Teen employment, poverty, and the minimum wage: Evidence from canada. **Labour Economics**, Elsevier BV, v. 18, n. 1, p. 36–47, Jan. 2011. ISSN 0927-5371. Available: <http://dx.doi.org/10.1016/j.labeco.2010.06.003>. Citation on page 122.

SHALIT, U.; JOHANSSON, F. D.; SONTAG, D. Estimating individual treatment effect: generalization bounds and algorithms. In: PRECUP, D.; TEH, Y. W. (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 3076–3085. Available: <https://proceedings.mlr.press/v70/shalit17a.html>. Citations on pages 34, 56, and 131.

SHEIKHOLESLAMI, S.; MEISTER, M.; WANG, T.; PAYBERAH, A. H.; VLASSOV, V.; DOWLING, J. Autoablation: Automated parallel ablation studies for deep learning. In: **Proceedings of the 1st Workshop on Machine Learning and Systems**. ACM, 2021. (EuroSys '21). Available: <http://dx.doi.org/10.1145/3437984.3458834>. Citation on page 34.

SHI, C.; BLEI, D.; VEITCH, V. Adapting neural networks for the estimation of treatment effects. In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2019. v. 32. Available: <https://proceedings.neurips.cc/

paper_files/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf>. Citations on pages 56 and 131.

SOUTO, H. G.; LOUZADA, F. **Ablation Studies for Novel Treatment Effect Estimation Models**. arXiv, 2024. Available: <https://arxiv.org/abs/2410.15560>. Citations on pages 93 and 95.

SOUTO, H. G.; MORADI, A. Can transformers transform financial forecasting? **China Finance Review International**, Emerald, Jun. 2024. ISSN 2044-1398. Available: <http://dx.doi.org/10.1108/CFRI-01-2024-0032>. Citation on page 34.

_____. Introducing nbeatsx to realized volatility forecasting. **Expert Systems with Applications**, Elsevier BV, v. 242, p. 122802, May 2024. ISSN 0957-4174. Available: <http://dx.doi.org/10.1016/j.eswa.2023.122802>. Citations on pages 58 and 59.

SOUTO, H. G.; NETO, F. L. **Advancing Causal Inference: A Nonparametric Approach to ATE and CATE Estimation with Continuous Treatments**. arXiv, 2024. Available: <https://arxiv.org/abs/2409.06593>. Citation on page 40.

_____. **Beyond Arbitrary Replications: A Principled Approach to Simulation Design in Causal Inference**. arXiv, 2024. Available: <https://arxiv.org/abs/2409.05161>. Citations on pages 29 and 41.

STREINER, D. L.; NORMAN, G. R. Correction for multiple testing. **Chest**, Elsevier BV, v. 140, n. 1, p. 16–18, Jul. 2011. ISSN 0012-3692. Available: <http://dx.doi.org/10.1378/chest.11-0523>. Citation on page 72.

STUART, E. A. Matching methods for causal inference: A review and a look forward. **Statistical Science**, Institute of Mathematical Statistics, v. 25, n. 1, Feb. 2010. ISSN 0883-4237. Available: <http://dx.doi.org/10.1214/09-STS313>. Citation on page 84.

SUN, L.; ABRAHAM, S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. **Journal of Econometrics**, Elsevier BV, v. 225, n. 2, p. 175–199, Dec. 2021. ISSN 0304-4076. Available: <http://dx.doi.org/10.1016/j.jeconom.2020.09.006>. Citation on page 86.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. Citation on page 34.

TAN, Z. A distributional approach for causal inference using propensity scores. **Journal of the American Statistical Association**, Informa UK Limited, v. 101, n. 476, p. 1619–1637, Dec. 2006. ISSN 1537-274X. Available: <http://dx.doi.org/10.1198/016214506000000023>. Citation on page 84.

TCHETGEN, E. J. T.; PARK, C.; RICHARDSON, D. B. Universal difference-in-differences for causal inference in epidemiology. **Epidemiology**, Ovid Technologies (Wolters Kluwer Health), v. 35, n. 1, p. 16–22, Nov. 2023. ISSN 1044-3983. Available: <http://dx.doi.org/10.1097/EDE.0000000000001676>. Citation on page 84.

THAL, D. R. C.; FORROW, L. V.; LIPMAN, E. R.; STARLING, J. E.; FINUCANE, M. M. **Aggregate Bayesian Causal Forests: The ABCs of Flexible Causal Inference for Hierarchically Structured Data**. arXiv, 2024. Available: <https://arxiv.org/abs/2407.07067>. Citation on page 34.

THOMPSON, J. P. Using local labor market data to re-examine the employment effects of the minimum wage. **ILR Review**, SAGE Publications, v. 62, n. 3, p. 343–366, Apr. 2009. ISSN 2162-271X. Available: <http://dx.doi.org/10.1177/001979390906200305>. Citation on page 124.

UM, S.; LINERO, A. R.; SINHA, D.; BANDYOPADHYAY, D. Bayesian additive regression trees for multivariate skewed responses. **Statistics in Medicine**, Wiley, v. 42, n. 3, p. 246–263, Nov. 2022. ISSN 1097-0258. Available: <http://dx.doi.org/10.1002/sim.9613>. Citation on page 74.

VALE, H. M. do; ESTOFOLETE, C. F.; PEREIRA, A. E. F. F. e. D.; VALE, E. P. B. M. do. Researchers and journals – what are their responsibilities? **International Journal of Innovative Research in Multidisciplinary Education**, IJSSHMR Publication, v. 03, n. 12, Dec. 2024. ISSN 2833-4531. Available: <http://dx.doi.org/10.58806/ijirme.2024.v3i12n01>. Citations on pages 59 and 80.

VANDENBROUCKE, J. P.; BROADBENT, A.; PEARCE, N. Causality and causal inference in epidemiology: the need for a pluralistic approach. **International Journal of Epidemiology**, Oxford University Press (OUP), v. 45, n. 6, p. 1776–1786, Jan. 2016. ISSN 1464-3685. Available: <http://dx.doi.org/10.1093/ije/dyv341>. Citations on pages 29, 56, and 83.

VARIAN, H. R. Causal inference in economics and marketing. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 113, n. 27, p. 7310–7315, Jul. 2016. ISSN 1091-6490. Available: <http://dx.doi.org/10.1073/pnas.1510479113>. Citations on pages 29, 55, 56, and 83.

WAGER, S.; ATHEY, S. Estimation and inference of heterogeneous treatment effects using random forests. **Journal of the American Statistical Association**, Informa UK Limited, v. 113, n. 523, p. 1228–1242, Jun. 2018. ISSN 1537-274X. Available: <http://dx.doi.org/10.1080/01621459.2017.1319839>. Citations on pages 34 and 56.

WASSERSTEIN, R. L.; LAZAR, N. A. The asa statement on p-values: Context, process, and purpose. **The American Statistician**, Informa UK Limited, v. 70, n. 2, p. 129–133, Apr. 2016. ISSN 1537-2731. Available: <http://dx.doi.org/10.1080/00031305.2016.1154108>. Citations on pages 107 and 108.

WASSERSTEIN, R. L.; SCHIRM, A. L.; LAZAR, N. A. Moving to a world beyond "p<0.05". **The American Statistician**, Informa UK Limited, v. 73, n. sup1, p. 1–19, Mar. 2019. ISSN 1537-2731. Available: <http://dx.doi.org/10.1080/00031305.2019.1583913>. Citations on pages 107 and 108.

WELCH, B. L. The generalization of 'student's' problem when several different population varlances are involved. **Biometrika**, Oxford University Press (OUP), v. 34, n. 1–2, p. 28–35, 1947. ISSN 1464-3510. Available: <http://dx.doi.org/10.1093/biomet/34.1-2.28>. Citations on pages 41, 66, and 67.

WELLINGTON, A. J. Effects of the minimum wage on the employment status of youths: An update. **The Journal of Human Resources**, JSTOR, v. 26, n. 1, p. 27, 1991. ISSN 0022-166X. Available: <http://dx.doi.org/10.2307/145715>. Citation on page 122.

WILLIAMS, N. Regional effects of the minimum wage on teenage employment. **Applied Economics**, Informa UK Limited, v. 25, n. 12, p. 1517–1528, Jul. 2006. ISSN 1466-4283. Available: <http://dx.doi.org/10.1080/00036849300000156>. Citation on page 122.

WITT, S. F.; SONG, H.; LOUVIERIS, P. Statistical testing in forecasting model selection. **Journal of Travel Research**, SAGE Publications, v. 42, n. 2, p. 151–158, Nov. 2003. ISSN 1552-6763. Available: <http://dx.doi.org/10.1177/0047287503253941>. Citations on pages 58 and 59.

WU, X.; MEALLI, F.; KIOUMOURTZOGLOU, M.-A.; DOMINICI, F.; BRAUN, D. Matching on generalized propensity scores with continuous exposures. **Journal of the American Statistical Association**, Informa UK Limited, v. 119, n. 545, p. 757–772, Dec. 2022. ISSN 1537-274X. Available: <http://dx.doi.org/10.1080/01621459.2022.2144737>. Citation on page 34.

WU, X.; XIA, Y.; ZHU, J.; WU, L.; XIE, S.; QIN, T. A study of bert for context-aware neural machine translation. **Machine Learning**, Springer Science and Business Media LLC, v. 111, n. 3, p. 917–935, Jan. 2022. ISSN 1573-0565. Available: <http://dx.doi.org/10.1007/s10994-021-06070-y>. Citation on page 34.

WYNTER, A. de. **Awes, Laws, and Flaws From Today's LLM Research**. arXiv, 2024. Available: <https://arxiv.org/abs/2408.15409>. Citation on page 59.

YAO, L.; CHU, Z.; LI, S.; LI, Y.; GAO, J.; ZHANG, A. A survey on causal inference. **ACM Transactions on Knowledge Discovery from Data**, Association for Computing Machinery (ACM), v. 15, n. 5, p. 1–46, May 2021. ISSN 1556-472X. Available: <http://dx.doi.org/10.1145/3444944>. Citation on page 29.

YEAGER, D. S.; BRYAN, C. J.; GROSS, J. J.; MURRAY, J. S.; COBB, D. K.; SANTOS, P. H. F.; GRAVELDING, H.; JOHNSON, M.; JAMIESON, J. P. A synergistic mindsets intervention protects adolescents from stress. **Nature**, Springer Science and Business Media LLC, v. 607, n. 7919, p. 512–520, Jul. 2022. ISSN 1476-4687. Available: <http://dx.doi.org/10.1038/s41586-022-04907-7>. Citations on pages 34 and 56.

YEAGER, D. S.; HANSELMAN, P.; WALTON, G. M.; MURRAY, J. S.; CROSNOE, R.; MULLER, C.; TIPTON, E.; SCHNEIDER, B.; HULLEMAN, C. S.; HINOJOSA, C. P.; PAUNESKU, D.; ROMERO, C.; FLINT, K.; ROBERTS, A.; TROTT, J.; IACHAN, R.; BUONTEMPO, J.; YANG, S. M.; CARVALHO, C. M.; HAHN, P. R.; GOPALAN, M.; MHATRE, P.; FERGUSON, R.; DUCKWORTH, A. L.; DWECK, C. S. A national experiment reveals where a growth mindset improves achievement. **Nature**, Springer Science and Business Media LLC, v. 573, n. 7774, p. 364–369, Aug. 2019. ISSN 1476-4687. Available: <http://dx.doi.org/10.1038/s41586-019-1466-y>. Citations on pages 34 and 56.

YEON, J.; KIM, S. .; SONG, K.; KIM, J. Examining the impact of short-term rental regulation on peer-to-peer accommodation performance: a difference-in-differences approach. **Current Issues in Tourism**, Informa UK Limited, v. 25, n. 19, p. 3212–3224, Sep. 2020. ISSN 1747-7603. Available: <http://dx.doi.org/10.1080/13683500.2020.1814704>. Citation on page 84.

YU, S.; LUO, M.; MADASU, A.; LAL, V.; HOWARD, P. **Is Your Paper Being Reviewed by an LLM? Investigating AI Text Detectability in Peer Review**. arXiv, 2024. Available: <https://arxiv.org/abs/2410.03019>. Citation on page 59.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: _____. **Computer Vision – ECCV 2014**. Springer International Publishing, 2014. p. 818–833. ISBN 9783319105901. Available: <http://dx.doi.org/10.1007/978-3-319-10590-1_53>. Citation on page 34.

ZHANG, J.; ZHAO, J.; JIANG, W.-j.; SHAN, X.-w.; YANG, X.-m.; GAO, J.-g. Conditional gene manipulation: Cre-ating a new biological era. **Journal of Zhejiang University SCIENCE B**, Zhejiang University Press, v. 13, n. 7, p. 511–524, Jul. 2012. ISSN 1862-1783. Available: <http://dx.doi.org/10.1631/jzus.B1200042>. Citation on page 39.

ZHOU, Z.; ZHAO, Y.; SHEN, C.; LAI, S.; NAWAZ, R.; GAO, J. Evaluating the effect of hierarchical medical system on health seeking behavior: A difference-in-differences analysis in china. **Social Science & Medicine**, Elsevier BV, v. 268, p. 113372, Jan. 2021. ISSN 0277-9536. Available: <http://dx.doi.org/10.1016/j.socscimed.2020.113372>. Citation on page 84.

ZIMMERMAN, D. W. A note on preliminary tests of equality of variances. **British Journal of Mathematical and Statistical Psychology**, Wiley, v. 57, n. 1, p. 173–181, May 2004. ISSN 2044-8317. Available: <http://dx.doi.org/10.1348/000711004849222>. Citations on pages 41 and 67.

APPENDIX

A

ADVANCED BAYESIAN REGRESSION TREE MODELS FOR PREDICTION AND CAUSAL INFERENCE (APPENDIX)

A.1 Introduction

Bayesian Additive Regression Trees (BART), introduced by Chipman, George and McCulloch (2010), represent a powerful non-parametric Bayesian approach for regression and classification tasks. BART models the relationship between a response variable Y and a set of predictors $\boldsymbol{X} = (X_1, ..., X_p)$ using a sum of many regression trees. Unlike single-tree models or frequentist ensemble methods like Random Forests or Gradient Boosting, BART employs a full Bayesian framework. This involves specifying a prior distribution over the entire function space represented by the sum-of-trees model and performing posterior inference. BART utilizes carefully constructed regularization priors to prevent overfitting, allowing the ensemble of shallow trees to capture complex relationships, including interactions, without letting any single tree dominate the fit. The resulting posterior distribution provides not only point predictions but also principled uncertainty quantification through posterior credible intervals.

This document provides a detailed exposition of the BART model, its estimation via Bayesian backfitting Markov chain Monte Carlo (MCMC), and several important extensions relevant to modern statistical practice. We will delve into Bayesian Causal Forests (BCF) for estimating heterogeneous treatment effects, accelerated algorithms like XBART and XBCF designed to improve computational efficiency for large datasets, and the warm-start BCF (wsBCF) technique aimed at improving stability and convergence. The discussion assumes familiarity with fundamental concepts in Bayesian statistics (prior specification, likelihood, posterior inference, MCMC) and the Gibbs sampler.

A.2 The Bayesian Additive Regression Trees (BART) Model

The BART model comprises two fundamental components: a sum-of-trees structure defining the functional form for the mean, and a carefully specified regularization prior placed on the parameters of this structure, including the tree structures themselves and the values assigned at their terminal nodes.

A.2.1 The Sum-of-Trees Model

At its core, BART models the conditional expectation of a response variable Y given predictors $\boldsymbol{X}$ as a sum of functions, where each function is represented by a regression tree:

$$E(Y|\boldsymbol{X}) = f(\boldsymbol{X}) = \sum_{j=1}^{m} g(\boldsymbol{X}; T_j, M_j) \tag{A.1}$$

Assuming normally distributed errors, which is common in regression settings, the full model is:

$$Y_i = f(\boldsymbol{X_i}) + \varepsilon_i = \sum_{j=1}^{m} g(\boldsymbol{X_i}; T_j, M_j) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{A.2}$$

for observations $i = 1, ..., n$. Here, m is the number of trees in the ensemble, and each $g(\boldsymbol{X}; T_j, M_j)$ represents the contribution of the j-th tree. Let's dissect the structure of a single tree, T_j, and its associated parameters, M_j.

A single regression tree, T, is a binary tree structure. It consists of:

- A set of interior (non-terminal) nodes. Each interior node k contains a decision rule involving a specific predictor X_{d_k} and a split point c_k. For a continuous predictor X_{d_k}, the rule partitions the predictor space based on whether $x_{d_k} \leq c_k$ or $x_{d_k} > c_k$ (where x_{d_k} is the realization of the random variable X_{d_k}). For a categorical predictor, the rule partitions based on subsets of its categories.
- A set of terminal nodes (leaves). Let b be the number of terminal nodes in tree T. Each terminal node represents a distinct partition (region) of the predictor space defined by the sequence of splits from the root to that node.

Associated with the tree T is a set of parameters $M = \{\mu_1, \mu_2, ..., \mu_b\}$, where each μ_l is the parameter value assigned to the l-th terminal node.

The function $g(\boldsymbol{X}; T, M)$ maps a given predictor vector $\boldsymbol{X}$ to one of the terminal node parameters μ_l. This mapping is determined by dropping $\boldsymbol{X}$ down the tree T. Starting at the root node, the decision rule is evaluated. Based on the outcome, $\boldsymbol{X}$ proceeds to the

left or right child node. This process continues until a terminal node l is reached. The value assigned to $\boldsymbol{X}$ by this tree is the parameter μ_l associated with that terminal node: $g(\boldsymbol{X};T,M) = \mu_{l(\boldsymbol{X})}$.

In a single-tree model ($m = 1$), $Y = g(\boldsymbol{X};T,M) + \varepsilon$, the parameter μ_l directly represents the conditional mean $E(Y|\boldsymbol{X})$ for all $\boldsymbol{X}$ falling into the region defined by the l-th terminal node.

However, in the BART sum-of-trees model (Equation A.2) with $m > 1$, each individual terminal node parameter μ_{lj} (the parameter for the l-th terminal node of the j-th tree, T_j) represents only a small fraction of the overall conditional mean $E(Y|\boldsymbol{X})$. The sum $\sum_{j=1}^{m} \mu_{l_j(\boldsymbol{X}),j}$, where $\mu_{l_j(\boldsymbol{X}),j}$ denotes the parameter from the terminal node of tree T_j that $\boldsymbol{X}$ falls into, constitutes the estimate of $f(\boldsymbol{X})$. This structure allows each tree to be simple (e.g., a stump or very shallow), focusing on capturing a small part of the overall pattern, such as a main effect or a low-order interaction.

This sum-of-trees structure provides immense flexibility. Each tree T_j can capture different aspects of the relationship between $\boldsymbol{X}$ and Y. If $g(\boldsymbol{X};T_j,M_j)$ depends only on a single predictor x_k, it contributes to modeling the main effect of that variable. If it depends on multiple predictors (i.e., the path to the terminal node involves splits on multiple variables), it captures interaction effects between those variables. By summing many trees, potentially of varying depths and complexities (though typically kept shallow by the prior), BART can approximate complex, non-linear functions involving high-order interactions.

The use of many trees introduces redundancy; the same function $f(\boldsymbol{X})$ can often be represented by different combinations of $(T_1,M_1),...,(T_m,M_m)$. This "overcomplete basis" property, combined with regularization priors, is key to BART's success. It allows the model to adapt flexibly to the data without overfitting, as the regularization ensures that individual tree contributions remain small.

A.2.2 The Regularization Prior

To complete the BART specification and control the complexity inherent in the sum-of-trees model, a prior distribution is placed over all unknown parameters: the tree structures $(T_1,...,T_m)$, their associated terminal node parameters $(M_1,...,M_m)$, and the error variance σ^2. The key idea, central to BART's philosophy, is to use priors that regularize the fit by keeping the influence of each individual tree small and favoring simpler structures.

Chipman, George and McCulloch (2010) propose a prior structure that simplifies specification and computation, assuming prior independence between the tree components and the error variance, and independence among the tree components themselves:

$$p((T_1,M_1),...,(T_m,M_m),\sigma^2) = \left[\prod_{j=1}^{m} p(M_j|T_j)p(T_j)\right] p(\sigma^2) \tag{A.3}$$

Furthermore, within each tree T_j, the terminal node parameters in $M_j = \{\mu_{1j},...,\mu_{b_j j}\}$ are assumed to be conditionally independent given the tree structure T_j:

$$p(M_j|T_j) = \prod_{l=1}^{b_j} p(\mu_{lj}|T_j) \tag{A.4}$$

This factorization reduces the complex task of specifying a joint prior over all parameters to specifying three distinct components, typically using identical functional forms across all trees $j = 1,...,m$ for simplicity and exchangeability:

1. $p(T_j)$: The prior on the structure (topology and splitting rules) of tree j.
2. $p(\mu_{lj}|T_j)$: The prior on the parameter μ_{lj} associated with the l-th terminal node of tree j.
3. $p(\sigma^2)$: The prior on the error variance.

These priors are controlled by a small number of hyperparameters. While these can be set based on subjective expert knowledge, BART implementations often employ data-informed default strategies to calibrate them, making the method relatively automatic and robust in practice.

A.2.2.1 Prior on Tree Structure: $p(T_j)$

The prior on the tree structure T_j is crucial for regularization, primarily by controlling the expected size and depth of the individual trees. It is defined by three elements:

1. **Node Splitting Probability:** The probability that a node at depth d ($d = 0$ for the root) is an internal (non-terminal) node is specified as:

$$P_{\text{split}}(d) = \eta(1+d_i)^{-\beta}, \quad \eta \in (0,1), \beta \geq 0 \tag{A.5}$$

This prior implies that nodes deeper in the tree are less likely to be split. The hyperparameter η controls the base probability of splitting at the root, while β controls the rate at which the splitting probability decreases with depth. Smaller values of η or larger values of β induce stronger regularization by favoring shallower trees. The default values suggested by Chipman, George and McCulloch (2010), $\eta = 0.95$ and $\beta = 2$, place high prior probability on very small trees (e.g., with 2 or 3 terminal nodes), enforcing the idea that each tree should contribute only a

small, simple component to the overall sum. However, the prior does not preclude the growth of larger trees if strongly supported by the data.

2. **Splitting Variable Distribution:** At each internal node, a predictor variable X_k must be chosen to define the split. A discrete uniform prior over the set of available predictor variables is typically used. This reflects an initial indifference regarding which variables are important.

3. **Splitting Rule Distribution:** Conditional on selecting a predictor variable X_k for a split at node d, a splitting value c (or subset for categorical variables) must be selected from the available values of X_k within the data points reaching node d. A common choice is a discrete uniform prior over the unique values of X_k present in the node. This choice, while computationally convenient, is invariant to monotone transformations of continuous predictors.

This overall prior structure $p(T_j)$ favors small individual trees, ensuring that each $g(\boldsymbol{x};T_j,M_j)$ contributes only a modest, regularized component to the overall sum $f(\boldsymbol{x})$, preventing any single tree from dominating the fit.

A.2.2.2 Prior on Terminal Node Parameters: $p(\mu_{lj}|T_j)$

Given a tree structure T_j, a prior is needed for the parameters μ_{lj} associated with its b_j terminal nodes. A conjugate Normal prior is typically used, primarily for computational convenience as it facilitates the marginalization required in the MCMC algorithm:

$$\mu_{lj} \overset{iid}{\sim} N(\mu_\mu, \sigma_\mu^2) \tag{A.6}$$

Note that the prior is usually assumed to be identical for all terminal nodes across all trees. Since $E(Y|\boldsymbol{X}) = \sum_{j=1}^{m} \mu_{l_j(\boldsymbol{X}),j}$, the prior on the individual μ_{lj} induces a prior on the overall function mean $f(\boldsymbol{X})$. Assuming μ_{lj} are i.i.d. $N(0,\sigma_\mu^2)$ (i.e., setting the prior mean $\mu_\mu = 0$), the implied prior for $f(\boldsymbol{X})$ at any given $\boldsymbol{X}$ is $N(0, m\sigma_\mu^2)$.

A common data-informed strategy is employed to set σ_μ^2. First, the response variable Y is linearly transformed (shifted and scaled) so that its observed values y_i range from a minimum y_{min} to a maximum y_{max} (e.g., often scaled to $[-0.5, 0.5]$). Let this transformed variable still be denoted by Y. Then, the prior mean μ_μ is typically set to 0 (or the mean of the transformed Y). The prior standard deviation σ_μ is chosen such that the bulk of the prior probability mass for the sum $f(\boldsymbol{X})$ falls within the observed range $[y_{min}, y_{max}]$. Specifically, we set:

$$\mu_\mu + k\sqrt{m}\sigma_\mu = y_{max} \quad \text{and} \quad \mu_\mu - k\sqrt{m}\sigma_\mu = y_{min} \tag{A.7}$$

If Y is scaled to $[-0.5, 0.5]$ and $\mu_\mu = 0$, this simplifies to $k\sqrt{m}\sigma_\mu = 0.5$, yielding $\sigma_\mu = \frac{0.5}{k\sqrt{m}}$. Here, k is a hyperparameter controlling the degree of shrinkage towards the prior mean μ_μ. A typical default value is $k = 2$, implying that $f(\boldsymbol{x})$ is a priori expected to be within $[y_{min}, y_{max}]$ with approximately 95% probability. Larger values of k or a larger number of trees m lead to smaller σ_μ, resulting in stronger shrinkage of the μ_{lj} values towards the center. This shrinkage is crucial for regularization, preventing any single tree from having an overly large effect and acting analogously to the learning rate or shrinkage parameter in boosting algorithms.

A.2.2.3 Prior on Error Variance: $p(\sigma^2)$

For the error variance σ^2 in Equation A.2, a conjugate Inverse Gamma prior is typically used, often parameterized via the equivalent Inverse Chi-Squared distribution:

$$\sigma^2 \sim \text{Inv-Gamma}(\nu/2, \nu\lambda/2) \quad \Leftrightarrow \quad \sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu} \tag{A.8}$$

Here, ν is the degrees of freedom parameter, and λ relates to the scale parameter (it is the prior sum of squares). The Inverse Gamma prior is conjugate to the Normal likelihood for the variance parameter, simplifying the Gibbs sampling step for σ^2. Similar to the prior for μ_{lj}, a data-informed approach is often used to set the hyperparameters ν and λ. First, a rough estimate $\hat{\sigma}$ of the residual standard deviation is obtained (e.g., the sample standard deviation of Y, or the residual standard deviation from an initial linear regression of Y on $\boldsymbol{x}$). Then, ν and λ are chosen such that the prior assigns substantial probability to values around $\hat{\sigma}$, while remaining relatively uninformative. This is often done by selecting a degrees of freedom parameter ν (e.g., $\nu = 3$, which implies a relatively flat prior) and then choosing λ such that a specific upper quantile (e.g., the 90th percentile, $q = 0.90$) of the induced prior distribution for σ (not σ^2) matches $\hat{\sigma}$. That is, $P(\sigma < \hat{\sigma}) = q$. This strategy aims to center the prior in a reasonable range suggested by the data, avoiding priors that are overly concentrated or excessively diffuse relative to the likely scale of the noise.

A.2.2.4 Choice of the Number of Trees: m

The number of trees, m, is a crucial tuning parameter. While it could, in principle, be treated as an unknown parameter with its own prior and estimated within the MCMC, this is rarely done due to computational complexity and potential identifiability issues. Instead, m is typically treated as a fixed hyperparameter selected by the user. Common practice is to select a relatively large value for m, such as $m = 50$, $m = 100$, or often $m = 200$ as a default. Empirical studies (CHIPMAN; GEORGE; MCCULLOCH, 2010) suggest that BART's predictive performance improves rapidly as m increases from small values, then

plateaus, and may slowly degrade if m becomes excessively large (due to the interaction with the prior on σ_μ, which shrinks more heavily as m increases). Therefore, choosing a sufficiently large m (Chipman, George and McCulloch (2010) suggest the default value of $m = 200$) is often adequate for prediction tasks and robust across various problems. Cross-validation could be used to select m from a grid of values, but this significantly increases computational cost. The choice of m interacts strongly with the prior settings, particularly σ_μ (Equation A.7), which scales inversely with $\sqrt{m}$. Yet, most commonly, researchers simple use the suggested default value of $m = 200$.

A.3 Posterior Inference: Bayesian Backfitting MCMC for BART

Given the BART model specification (Equation A.2) and the priors described above, the goal is to perform inference based on the joint posterior distribution of all unknown parameters given the observed data $(\boldsymbol{y}, \boldsymbol{x})$:

$$p((T_1,M_1),...,(T_m,M_m),\sigma^2|\boldsymbol{y},\boldsymbol{x}) \propto p(\boldsymbol{y}|(T_1,M_1),...,(T_m,M_m),\sigma^2,\boldsymbol{x})p((T_1,M_1),...,(T_m,M_m),\sigma^2) \tag{A.9}$$

Due to the high dimensionality and complexity of the parameter space, particularly the variable structure of the trees $(T_1,...,T_m)$, this posterior distribution is intractable to compute or sample from directly. BART utilizes a Markov Chain Monte Carlo (MCMC) algorithm, specifically a Gibbs sampler incorporating Metropolis-Hastings steps for the tree structures. This approach is often referred to as "Bayesian backfitting" because it iteratively updates each tree conditional on the others, similar in spirit to the backfitting algorithm used for fitting generalized additive models.

The core idea of the backfitting MCMC is to iteratively sample each component (or block of components) of the model conditional on the current values of all other components. The algorithm cycles through sampling:

1. **Each tree component** (T_j,M_j): For $j = 1,...,m$, sample (T_j,M_j) from its full conditional distribution $p((T_j,M_j)|\{(T_k,M_k)\}_{k\neq j},\sigma^2,\boldsymbol{y},\boldsymbol{x})$.

2. **The error variance** σ^2: Sample σ^2 from its full conditional distribution $p(\sigma^2|\{(T_k,M_k)\}_{k=1}^m,\boldsymbol{y},\boldsymbol{x}$

Under regularity conditions, after a sufficient number of iterations (burn-in), the samples drawn from this iterative process converge in distribution to the target joint posterior distribution.

A.3.1 Sampling the Tree Components (T_j, M_j)

Sampling a specific tree component (T_j, M_j) involves updating both its structure T_j and its terminal node parameters M_j. This is done conditional on the current state of all other parameters and the data. The key insight is to define the partial residuals for the j-th tree:

$$\boldsymbol{R}_j = \boldsymbol{y} - \sum_{k \neq j} g(\boldsymbol{x}; T_k, M_k) \tag{A.10}$$

where $g(\boldsymbol{x}; T_k, M_k)$ denotes the vector of predictions from tree T_k for all n observations. The vector $\boldsymbol{R}_j$ represents the part of the response variable $\boldsymbol{y}$ that is currently unexplained by the other $m-1$ trees. Conditional on the other trees and σ^2, the model for $\boldsymbol{R}_j$ simplifies to :

$$\boldsymbol{R}_j = g(\boldsymbol{x}; T_j, M_j) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{A.11}$$

This effectively reduces the problem to performing Bayesian inference for a single-tree model applied to the partial residuals $\boldsymbol{R}_j$. The sampling for (T_j, M_j) typically proceeds in two stages within the Gibbs iteration:

1. **Integrate out M_j to sample T_j**: Given the conjugate Normal prior $p(\mu_{lj}|T_j) = N(\mu_\mu, \sigma_\mu^2)$, the terminal node parameters $M_j = \{\mu_{lj}\}$ can be analytically integrated out (marginalized) from the joint distribution $p(\boldsymbol{R}_j, M_j | T_j, \sigma^2)$. This yields the marginal likelihood of the tree structure T_j given the partial residuals $\boldsymbol{R}_j$ and σ^2, denoted as $p(\boldsymbol{R}_j | T_j, \sigma^2)$.

   Let I_{lj} be the set of indices of observations $(i = 1, ..., n)$ that fall into the l-th terminal node of tree T_j. Let $n_{lj} = |I_{lj}|$ be the number of observations in that node, and let $R^*_{lj} = \sum_{i \in I_{lj}} (R_{ji} - \mu_\mu)$ be the sum of centered partial residuals in that node. The marginal likelihood $p(\boldsymbol{R}_j | T_j, \sigma^2)$ can be computed by multiplying the marginal likelihood contributions from each terminal node, which involve terms related to n_{lj}, R^*_{lj}, σ^2, and σ_μ^2. (The exact formula involves Normal likelihood calculations integrated over the Normal prior for μ_{lj}).

   Since the space of possible tree structures T_j is vast and complex, sampling T_j directly from its conditional posterior $p(T_j | \boldsymbol{R}_j, \sigma^2) \propto p(\boldsymbol{R}_j | T_j, \sigma^2) p(T_j)$ is infeasible. Instead, a Metropolis-Hastings (MH) algorithm is employed. Starting from the current tree structure T_j^{current}, a new candidate structure T_j^{proposal} is generated using a proposal mechanism $q(T_j^{\text{proposal}} | T_j^{\text{current}})$. This typically involves one of several local moves with their respective invariant proposal probabilities:

- **GROW (0.25):** Randomly select a terminal node and split it into two new terminal nodes by introducing a new decision rule (randomly choosing a splitting variable and splitting value).
- **PRUNE (0.25):** Randomly select an internal node whose children are both terminal nodes (a "pair of leaves") and collapse it back into a single terminal node, removing the split.
- **CHANGE (0.40):** Randomly select an internal node and modify its decision rule (e.g., by picking a new splitting variable or a new splitting value for the existing variable).
- **SWAP (0.10):** Randomly select a parent-child pair of internal nodes and swap their decision rules. (This move is less common in basic implementations).

The proposal T_j^{proposal} is accepted with probability:

$$\alpha(T_j^{\text{current}}, T_j^{\text{proposal}}) = \min\left\{1, \frac{p(T_j^{\text{proposal}}|\boldsymbol{R}_j, \sigma^2)}{p(T_j^{\text{current}}|\boldsymbol{R}_j, \sigma^2)} \times \frac{q(T_j^{\text{current}}|T_j^{\text{proposal}})}{q(T_j^{\text{proposal}}|T_j^{\text{current}})}\right\} \quad \text{(A.12)}$$

The first ratio inside the minimum is the ratio of posterior probabilities, which decomposes into the ratio of marginal likelihoods $\frac{p(\boldsymbol{R}_j|T_j^{\text{proposal}}, \sigma^2)}{p(\boldsymbol{R}_j|T_j^{\text{current}}, \sigma^2)}$ and the ratio of tree structure priors $\frac{p(T_j^{\text{proposal}})}{p(T_j^{\text{current}})}$. The second ratio is the Hastings correction, accounting for the proposal probabilities (e.g., the probability of proposing a PRUNE move from T_j^{proposal} that results in T_j^{current} divided by the probability of the original GROW move from T_j^{current} to T_j^{proposal}).

2. **Sample M_j given T_j**: Once the (potentially updated) tree structure T_j is determined for the current MCMC iteration, the terminal node parameters $M_j = \{\mu_{lj}\}$ are sampled from their conditional posterior distribution $p(M_j|T_j, \boldsymbol{R}_j, \sigma^2)$. Due to conjugacy, this involves drawing each μ_{lj} independently from its updated Normal posterior distribution:

$$\mu_{lj}|T_j, \boldsymbol{R}_j, \sigma^2 \sim N\left(\frac{\frac{n_{lj}}{\sigma^2}\bar{R}_{lj} + \frac{1}{\sigma_\mu^2}\mu_\mu}{\frac{n_{lj}}{\sigma^2} + \frac{1}{\sigma_\mu^2}}, \left(\frac{n_{lj}}{\sigma^2} + \frac{1}{\sigma_\mu^2}\right)^{-1}\right) \quad \text{(A.13)}$$

where $\bar{R}_{lj} = (1/n_{lj})\sum_{i \in I_{lj}} R_{ji}$ is the mean partial residual in node l. This posterior mean represents a weighted average of the data mean in the node ($barR_{lj}$) and the prior mean (
mu_{mu}), shrunk towards the prior mean based on the relative precisions.

This entire process (calculating partial residuals, MH step for T_j, Gibbs step for M_j) is repeated for each tree $j = 1, ..., m$ within a single iteration of the main Gibbs sampler.

A.3.2 Sampling the Error Variance σ^2

Conditional on all the tree structures and terminal node parameters $(T_1, M_1), ..., (T_m, M_m)$, the full model residuals can be computed for all observations:

$$\boldsymbol{e} = \boldsymbol{y} - \sum_{j=1}^{m} g(\boldsymbol{x}; T_j, M_j) \tag{A.14}$$

Given the Inverse Gamma prior $\sigma^2 \sim \text{Inv-Gamma}(\nu/2, \nu\lambda/2)$, the full conditional posterior distribution for σ^2 is also Inverse Gamma (due to conjugacy with the Normal likelihood):

$$\sigma^2 | \{(T_k, M_k)\}_k, \boldsymbol{y}, \boldsymbol{x} \sim \text{Inv-Gamma}\left(\frac{\nu + n}{2}, \frac{\nu\lambda + \sum_{i=1}^{n} e_i^2}{2}\right) \tag{A.15}$$

Equivalently, in the Inverse Chi-Squared parameterization:

$$\sigma^2 | \{(T_k, M_k)\}_k, \boldsymbol{y}, \boldsymbol{x} \sim \frac{\nu\lambda + ||\boldsymbol{e}||^2}{\chi^2_{\nu+n}} \tag{A.16}$$

where n is the number of observations and $||\boldsymbol{e}||^2 = \sum_{i=1}^{n} e_i^2$ is the sum of squared residuals. Sampling from this distribution is straightforward.

A.3.3 Algorithm Summary and Output

The Bayesian backfitting MCMC algorithm for BART proceeds as follows:

1. Initialize $(T_1^{(0)}, M_1^{(0)}), ..., (T_m^{(0)}, M_m^{(0)})$ and $\sigma^{2(0)}$. Often, initial trees are simple stumps (root node only with parameter $\mu_{1j} = y_i/m$ or 0 if Y is centered) and $\sigma^{2(0)}$ is set to the sample variance of Y or residuals from a linear model.

2. For MCMC iteration $t = 1, ..., N_{\text{iter}}$:

   a) For $j = 1, ..., m$ (iterate through trees):

      i. Compute partial residuals $\boldsymbol{R}_j^{(t-1)} = \boldsymbol{y} - \sum_{k<j} g(\boldsymbol{x}; T_k^{(t)}, M_k^{(t)}) - \sum_{k>j} g(\boldsymbol{x}; T_k^{(t-1)}, M_k^{(t-1)})$.

      ii. Sample $T_j^{(t)}$ from $p(T_j | \boldsymbol{R}_j^{(t-1)}, \sigma^{2(t-1)})$ using a Metropolis-Hastings step (propose T_j^{proposal} from $T_j^{(t-1)}$ via GROW/PRUNE/CHANGE/SWAP, accept/reject based on marginal likelihood and prior ratios).

      iii. Sample terminal node parameters $M_j^{(t)} = \{\mu_{lj}^{(t)}\}$ from $p(M_j | T_j^{(t)}, \boldsymbol{R}_j^{(t-1)}, \sigma^{2(t-1)})$ (Gibbs step using the Normal posterior).

   b) Compute full residuals $\boldsymbol{e}^{(t)} = \boldsymbol{y} - \sum_{j=1}^{m} g(\boldsymbol{x}; T_j^{(t)}, M_j^{(t)})$.

   c) Sample $\sigma^{2(t)}$ from $p(\sigma^2 | \boldsymbol{e}^{(t)}, \nu, \lambda) = \text{Inv-Gamma}((\nu + n)/2, (\nu\lambda + ||\boldsymbol{e}^{(t)}||^2)/2)$.

3. Discard initial burn-in samples (e.g., first N_{burn} iterations) and collect the remaining $N_{\text{post}} = N_{\text{iter}} - N_{\text{burn}}$ samples.

The output of the algorithm is a set of samples from the posterior distribution: $\{(\{(T_j^{(t)}, M_j^{(t)})\}_{j=1}^m, \sigma^{2(t)})\}_{t=N_{\text{burn}}+1}^{N_{\text{iter}}}$. Posterior inference is based on these samples. For prediction at a new point $\boldsymbol{x}_{new}$, we compute $f^{(t)}(\boldsymbol{x}_{new}) = \sum_{j=1}^m g(\boldsymbol{x}_{new}; T_j^{(t)}, M_j^{(t)})$ for each post-burn-in sample t. The average of these $f^{(t)}(\boldsymbol{x}_{new})$ values provides the posterior mean prediction $E[f(\boldsymbol{x}_{new})|\boldsymbol{y}, \boldsymbol{X}]$, and the quantiles of the samples provide posterior credible intervals for $f(\boldsymbol{x}_{new})$, offering a natural measure of uncertainty.

A.4 Bayesian Causal Forests (BCF)

While BART provides a flexible framework for prediction ($E[Y|\boldsymbol{X}]$), estimating causal effects, particularly heterogeneous treatment effects (HTEs), requires modifications to the standard BART structure to properly account for treatment assignment and potential confounding. BCF, introduced by Hahn, Murray and Carvalho (2020) (Hahn, Murray, and Carvalho, 2020), adapt the BART framework specifically for causal inference in settings with a binary treatment $Z \in \{0, 1\}$.

Consider a setting with n units, where for each unit i, we observe a vector of pre-treatment covariates $\boldsymbol{X_i}$, a binary treatment assignment D_i, and an outcome Y_i. Under the potential outcomes framework, let $Y_i(1)$ and $Y_i(0)$ be the potential outcomes for unit i if they receive the treatment ($D_i = 1$) or control ($D_i = 0$), respectively. The fundamental problem of causal inference is that we only observe one potential outcome for each unit: $Y_i = Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0)$.

The goal is often to estimate the Conditional Average Treatment Effect (CATE) function, $\tau(\boldsymbol{X})$, defined as the expected difference in potential outcomes conditional on covariates:

$$\tau(\boldsymbol{x}_i) = E[Y_i(1) - Y_i(0)|\boldsymbol{X_i} = \boldsymbol{x_i}] \tag{A.17}$$

BCF aims to estimate this function $\tau(\boldsymbol{X})$ flexibly, allowing the treatment effect to vary across different covariate profiles $\boldsymbol{X}$.

BCF models the observed outcome Y_i by decomposing the conditional expectation $E[Y_i|\boldsymbol{X_i}, D_i]$:

$$E[Y_i|\boldsymbol{X_i}, D_i] = E[Y_i(0)|\boldsymbol{X_i}, D_i] + D_i(E[Y_i(1)|\boldsymbol{X_i}, D_i] - E[Y_i(0)|\boldsymbol{X_i}, D_i]) \tag{A.18}$$

$$= \mu(\boldsymbol{X_i}) + \tau(\boldsymbol{X_i}) D_i \tag{A.19}$$

This decomposition holds under the assumption of unconfoundedness (also known as ignorability or selection on observables), which states that treatment assignment D_i is independent of the potential outcomes $(Y_i(0), Y_i(1))$ conditional on the covariates $\boldsymbol{X_i}$: $(Y_i(0), Y_i(1)) \perp D_i | \boldsymbol{X_i}$. Under this assumption, $E[Y_i(d)|\boldsymbol{X_i}, D_i = d] = E[Y_i(d)|\boldsymbol{X_i}]$. The full BCF model is then:

$$Y_i = \mu(\boldsymbol{X_i}) + \tau()\boldsymbol{X_i})D_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{A.20}$$

Here:

- $\mu(\boldsymbol{x_i}) = E[Y_i(0)|\boldsymbol{X_i} = \boldsymbol{x_i}]$ is the conditional mean outcome under the control condition (the prognostic effect function).
- $\tau(\boldsymbol{x_i}) = E[Y_i(1)|\boldsymbol{X_i} = \boldsymbol{x_i}] - E[Y_i(0)|\boldsymbol{X_i} = \boldsymbol{x_i}]$ is the CATE function.

The key innovation of BCF is to model both the prognostic effect $\mu(\boldsymbol{X})$ and the treatment effect modification $\tau(\boldsymbol{X})$ using BART-like sum-of-trees priors, but with modifications designed to regularize the estimation of the treatment effect $\tau(\boldsymbol{X})$ specifically and robustly, often more strongly than the prognostic effect $\mu(\boldsymbol{X})$.

A.4.1 BCF Model Structure and Priors

BCF uses separate sum-of-trees models for $\mu(\boldsymbol{X})$ and $\tau(\boldsymbol{X})$:

$$\mu(\boldsymbol{X}) = \sum_{j=1}^{m_\mu} g_\mu(\boldsymbol{X}; T_{\mu j}, M_{\mu j}) \tag{A.21}$$

$$\tau(\boldsymbol{X}) = \sum_{k=1}^{m_\tau} g_\tau(\boldsymbol{X}; T_{\tau k}, M_{\tau k}) \tag{A.22}$$

However, a direct application of standard BART priors to both $\mu(\boldsymbol{X})$ and $\tau(\boldsymbol{X})$ within the model $Y_i = \mu(\boldsymbol{X_i}) + \tau(\boldsymbol{X_i})D_i + \varepsilon_i$ can be problematic, especially in observational studies where the treatment assignment D_i itself depends on $\mu(\boldsymbol{X})$ (HAHN; MURRAY; CARVALHO, 2020). This leads to bias in the estimation of $\tau(\boldsymbol{X})$, especially with with strong confounding and weak signal-to-noise ratios (HAHN; MURRAY; CARVALHO, 2020).

To address this and improve the estimation of $\tau(\boldsymbol{X})$, BCF incorporates an estimate of the propensity score, $\pi(\boldsymbol{x_i}) = P(D_i = 1|\boldsymbol{X_i} = \boldsymbol{x_i})$, directly into the model for the prognostic effect μ. The modified BCF model structure becomes:

$$Y_i = \mu(\boldsymbol{X_i}, \hat{\pi}(\boldsymbol{X_i})) + \tau(\boldsymbol{X_i})D_i + \varepsilon_i \tag{A.23}$$

where $\hat{\pi}(\boldsymbol{X_i})$ is an estimate of the propensity score (e.g., obtained from logistic regression or most commonly a flexible model like BART fitted to the treatment assignment D_i). This estimate $\hat{\pi}(\boldsymbol{X_i})$ is included as an additional predictor when modeling μ. The rationale is that its incorporation as another covariate for μ diminishes the regularization-induced confounding (RIC) bias (HAHN; MURRAY; CARVALHO, 2020). For more information about RIC, please see Hahn, Murray and Carvalho (2020).

Furthermore, BCF employs specific prior structures for the components of μ and τ aimed at regularizing $\tau(\boldsymbol{X})$ more strongly than $\mu(\boldsymbol{X})$. The justification is that treatment effects (τ) are often expected to be smaller and smoother (i.e., less complex) than baseline prognostic effects (μ), and thus require more shrinkage to prevent the estimation from being dominated by noise.

Specifically, the priors are set up as follows:

- **Prior for $\mu(\boldsymbol{X}, \pi)$**: A standard BART prior (as described in Section A.2.2) is used for the sum-of-trees model representing μ. This involves priors on the tree structures $T_{\mu j}$, terminal node parameters $M_{\mu j}$, and potentially the number of trees m_μ. The predictors used by these trees include both the original covariates $\boldsymbol{X}$ and the estimated propensity score $\hat{\pi}(\boldsymbol{X})$.
- **Prior for $\tau(\boldsymbol{x})$**: A BART prior is also used for $\tau(\boldsymbol{x})$, involving priors on $T_{\tau k}$, $M_{\tau k}$, and m_τ. However, the hyperparameters of this prior, particularly the prior variance for the terminal node parameters $\sigma^2_{\mu,\tau}$ (analogous to σ^2_μ in standard BART), are typically chosen to induce stronger shrinkage on $\tau(\boldsymbol{x})$ compared to $\mu(\boldsymbol{x})$. This might involve using a smaller value for $\sigma_{\mu,\tau}$ (e.g., by using a larger k_τ factor in Equation A.7). Additionally, the prior on the tree structures $T_{\tau k}$ might favor even simpler trees (e.g., using a smaller α_τ or larger β_τ in Equation A.5) than those used for μ.
- **Prior for σ^2**: An Inverse Gamma prior is used for the error variance, identical to standard BART.

This careful construction, combining propensity score adjustment with differential regularization, is designed to yield robust and reliable estimates of the CATE function $\tau(\boldsymbol{X})$.

A.5 BCF Estimation Algorithm

Estimation for the BCF model also relies on a Bayesian backfitting MCMC algorithm, similar in spirit to the one used for standard BART, but adapted to handle the two-component structure (μ and τ) and the specific roles of the treatment variable $\boldsymbol{D}$ and the propensity score $\hat{\pi}$.

The Gibbs sampler iteratively draws samples from the full conditional posterior distributions of the parameters:

- The components of the μ function: $(T_{\mu j}, M_{\mu j})$ for $j = 1, ..., m_\mu$.
- The components of the τ function: $(T_{\tau k}, M_{\tau k})$ for $k = 1, ..., m_\tau$.
- The error variance σ^2.

A.5.1 Sampling the μ Components

To sample the j-th tree component $(T_{\mu j}, M_{\mu j})$ of the prognostic function μ, we define partial residuals specific to this component:

$$\boldsymbol{R}_{\mu j} = \boldsymbol{y} - \left(\sum_{l \neq j} g_\mu(\boldsymbol{x}, \hat{\boldsymbol{\pi}}; T_{\mu l}, M_{\mu l}) \right) - \boldsymbol{D} \odot \tau(\boldsymbol{x}) \tag{A.24}$$

where $\odot$ denotes element-wise multiplication, $\tau(\boldsymbol{x})$ is the current estimate of the vector of treatment effects (sum of all g_τ terms), and $\hat{\boldsymbol{\pi}}$ is the vector of estimated propensity scores. The problem reduces to fitting a single tree $g_\mu(\cdot; T_{\mu j}, M_{\mu j})$ to these residuals $\boldsymbol{R}_{\mu j}$, using predictors $(\boldsymbol{x}, \hat{\boldsymbol{\pi}})$. The sampling proceeds exactly as in the standard BART MCMC step described in Section A.3.1: integrate out the terminal node parameters $M_{\mu j}$, use Metropolis-Hastings (GROW, PRUNE, CHANGE, SWAP) to sample the tree structure $T_{\mu j}$ based on the marginal likelihood $p(\boldsymbol{R}_{\mu j} | T_{\mu j}, \sigma^2)$ and the tree prior $p(T_{\mu j})$, and then sample the terminal node parameters $M_{\mu j}$ from their Normal conditional posterior given the updated $T_{\mu j}$.

A.5.2 Sampling the τ Components

Sampling the k-th tree component $(T_{\tau k}, M_{\tau k})$ of the treatment effect function τ requires careful consideration of the model structure $Y_i = \mu(\boldsymbol{X_I}, \hat{\pi}_i) + \tau(\boldsymbol{X_i}) D_i + \varepsilon_i$. Let $\tau_{-k}(\boldsymbol{x}) = \sum_{l \neq k} g_\tau(\boldsymbol{x}; T_{\tau l}, M_{\tau l})$ be the contribution of all other treatment effect trees.

Define the adjusted response by removing the current estimate of the prognostic effect:

$$\boldsymbol{Y}^* = \boldsymbol{y} - \mu(\boldsymbol{x}, \hat{\boldsymbol{\pi}}) \tag{A.25}$$

where $\mu(\boldsymbol{x}, \hat{\boldsymbol{\pi}})$ is the current estimate of the prognostic effect vector (sum of all g_μ terms). The model for $\boldsymbol{Y}^*$ is approximately:

$$\boldsymbol{Y}^* \approx \tau(\boldsymbol{x}) \odot \boldsymbol{D} + \boldsymbol{\varepsilon} = (\tau_{-k}(\boldsymbol{x}) + g_\tau(\boldsymbol{x}; T_{\tau k}, M_{\tau k})) \odot \boldsymbol{D} + \boldsymbol{\varepsilon} \tag{A.26}$$

Now, define the partial residuals for the k-th treatment effect tree:

$$\boldsymbol{R}_{\tau k} = \boldsymbol{Y}^* - \tau_{-k}(\boldsymbol{x}) \odot \boldsymbol{D} \tag{A.27}$$

Substituting the model, we get:

$$\boldsymbol{R}_{\tau k} \approx g_\tau(\boldsymbol{x}; T_{\tau k}, M_{\tau k}) \odot \boldsymbol{D} + \boldsymbol{\varepsilon} \tag{A.28}$$

This equation highlights that the k-th tree $g_\tau(\cdot; T_{\tau k}, M_{\tau k})$ only influences the residual $\boldsymbol{R}_{\tau k}$ for the treated units (where $D_i = 1$). For control units ($D_i = 0$), $R_{\tau k,i} \approx \varepsilon_i$. Therefore, the likelihood information for updating $(T_{\tau k}, M_{\tau k})$ comes only from the treated units.

The sampling procedure is thus modified:

1. Consider only the subset of observations where $D_i = 1$. Let $\boldsymbol{R}_{\tau k}^{(1)}$ and $\boldsymbol{X}^{(1)}$ be the residuals and covariates corresponding to these treated units.
2. Calculate the marginal likelihood $p(\boldsymbol{R}_{\tau k}^{(1)} | T_{\tau k}, \sigma^2)$ based only on these treated units, integrating out the parameters $M_{\tau k}$ using their prior (which is typically more regularized than the prior for μ).
3. Use Metropolis-Hastings (GROW, PRUNE, CHANGE, SWAP on predictors $\boldsymbol{X}$) to sample the tree structure $T_{\tau k}$ using this marginal likelihood and the specific tree prior $p(T_{\tau k})$ chosen for the treatment effect component.
4. Sample the terminal node parameters $M_{\tau k} = \{\mu_{lk\tau}\}$ from their conditional posterior distribution, which is again derived using only the information from the treated units ($\boldsymbol{R}_{\tau k}^{(1)}$, $\boldsymbol{x}^{(1)}$) falling into each terminal node, combined with the (stronger) prior $N(\mu_{\mu,\tau}, \sigma^2_{\mu,\tau})$.

This focus on the treated units for updating τ is crucial for correctly identifying the treatment effect modification.

A.5.3 Sampling σ^2

Given the current estimates of the full $\mu(\boldsymbol{x}, \hat{\boldsymbol{\pi}})$ and $\tau(\boldsymbol{x})$ functions, the overall model residuals are calculated:

$$\boldsymbol{e} = \boldsymbol{y} - \mu(\boldsymbol{x}, \hat{\boldsymbol{\pi}}) - \tau(\boldsymbol{x}) \odot \boldsymbol{D} \tag{A.29}$$

The error variance σ^2 is then sampled from its full conditional posterior, which remains an Inverse Gamma (or Inverse Chi-Squared) distribution, identical in form to the standard BART update:

$$\sigma^2 | \cdots \sim \text{Inv-Gamma}\left(\frac{\nu + n}{2}, \frac{\nu\lambda + ||\boldsymbol{e}||^2}{2}\right) \tag{A.30}$$

A.5.4 Algorithm Summary

The BCF MCMC algorithm involves initializing all parameters (often μ is initialized based on Y for controls, τ is initialized to 0) and then iterating the following steps for $t = 1, ..., N_{\text{iter}}$:

1. For $j = 1, ..., m_\mu$:
   - Compute partial residuals $\boldsymbol{R}_{\mu j}^{(t-1)}$.
   - Sample $(T_{\mu j}^{(t)}, M_{\mu j}^{(t)})$ conditional on $\boldsymbol{R}_{\mu j}^{(t-1)}$, $\sigma^{2(t-1)}$, and predictors $(\boldsymbol{x}, \hat{\boldsymbol{\pi}})$ using the standard BART MCMC step (MH for $T_{\mu j}$, Gibbs for $M_{\mu j}$).
2. For $k = 1, ..., m_\tau$:
   - Compute adjusted response $\boldsymbol{Y}^{*(t)}$ and partial residuals $\boldsymbol{R}_{\tau k}^{(t-1)}$.
   - Sample $(T_{\tau k}^{(t)}, M_{\tau k}^{(t)})$ conditional on $\boldsymbol{R}_{\tau k}^{(1),(t-1)}$ (treated units only), $\sigma^{2(t-1)}$, and predictors $\boldsymbol{X}$ using the modified BART MCMC step (MH for $T_{\tau k}$, Gibbs for $M_{\tau k}$, using τ-specific priors).
3. Compute full residuals $\boldsymbol{e}^{(t)}$.
4. Sample $\sigma^{2(t)}$ from its Inverse Gamma conditional posterior.

After discarding burn-in samples, the collected samples for $\tau^{(t)}(\boldsymbol{x}) = \sum_{k=1}^{m_\tau} g_\tau(\boldsymbol{x}; T_{\tau k}^{(t)}, M_{\tau k}^{(t)})$ provide draws from the posterior distribution of the CATE function. Posterior means, medians, and credible intervals for $\tau(\boldsymbol{x})$ at specific covariate values $\boldsymbol{x}$ can be computed from these draws, providing estimates of heterogeneous treatment effects and associated uncertainty.

A.6 Warm-Start Bayesian Causal Forests (ws-BCF)

Estimation in BART and BCF traditionally relies on backfitting MCMC algorithms, which can be computationally intensive and slow to converge (HE; HAHN, 2021), especially with large datasets, due to highly correlated tree samples (ALCANTARA *et al.*, 2024). He and Hahn (2021) proposed XBART (Accelerated BART), which uses a more efficient "Grow-From-Root" stochastic tree-fitting algorithm, described in Algorithm 2. This algorithm explores the tree space more rapidly.

Krantsevich, He and Hahn (2023) extended this to BCF, creating the XBCF algorithm. XBCF essentially applies the XBART fitting approach to the BCF model, often with a slight modification allowing for heteroskedastic errors by treatment status:

$$Y_i = a\mu(\mathbf{X_i}, \hat{\pi}(\mathbf{X_i})) + b_{D_i} D_i \tilde{\tau}(\mathbf{X_i}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$a \sim N(0,1), \quad b_0, b_1 \sim N(0, 1/2),$$

where $\mu(x)$ and $\tilde{\tau}(x)$ are XBART forests, the actual treatment effect is $\tau(x) = (b_1 - b_0)\tilde{\tau}(x)$ from the full parameterization, and a is an additional scaling factor, which enhances the learning of the prognostic term (KRANTSEVICH; HE; HAHN, 2023). The Grow-From-Root algorithm stochastically grows trees, offering faster posterior exploration.

Algorithm 2 – GrowFromRoot (as described in Alcantara *et al.* (2024))

procedure GrowFromRoot($y, \mathbf{X}, \Phi, \Psi, d, T,$ node) ▷ Modifies T by adding nodes and sampling associated leaf parameters μ.
  if the stopping conditions are met for node **then**
    $\mu_{\text{node}} \leftarrow$ SampleParameters($\emptyset$) ▷ Node becomes a leaf, update parameter
    return
  end if
  $s^{\emptyset} \leftarrow s(y, \mathbf{X}, \Psi, \mathscr{C}, \text{all})$
  for $c_{jk} \in \mathscr{C}$ **do**
    $s_{jk}^{(1)} \leftarrow s(y, \mathbf{X}, \Psi, \mathscr{C}, j, k, \text{left})$
    $s_{jk}^{(2)} \leftarrow s(y, \mathbf{X}, \Psi, \mathscr{C}, j, k, \text{right})$
    Calculate $L(c_{jk}) = m\left(s_{jk}^{(1)}; \Phi, \Psi\right) \times m\left(s_{jk}^{(2)}; \Phi, \Psi\right)$
  end for
  Calculate $L(\emptyset) = |\mathscr{C}|\left(\eta(1+d)^{-\beta} - 1\right) m\left(s^{\emptyset}; \Phi, \Psi\right)$
  Sample a cutpoint c_{jk}^* (from $\mathscr{C}$) or the null cutpoint $\emptyset$ with probabilities:
    $P(c_{jk}) = \frac{L(c_{jk})}{\sum_{c'_{j'k'} \in \mathscr{C}} L(c'_{j'k'}) + L(\emptyset)}$ for $c_{jk} \in \mathscr{C}$
    $P(\emptyset) = \frac{L(\emptyset)}{\sum_{c'_{j'k'} \in \mathscr{C}} L(c'_{j'k'}) + L(\emptyset)}$
  if the null cutpoint $\emptyset$ is selected (i.e., $c_{jk}^* = \emptyset$) **then**
    $\mu_{\text{node}} \leftarrow$ SampleParameters($\emptyset$)
    return
  else ▷ A non-null cutpoint c_{jk}^* was selected
    Create two new child nodes, **left_node** and **right_node**, for **node** in T
    Partition data $(y, \mathbf{X})$ at **node** into $(y_{\text{left}}, \mathbf{X}_{\text{left}})$ and $(y_{\text{right}}, \mathbf{X}_{\text{right}})$:
      $(y_{\text{left}}, \mathbf{X}_{\text{left}})$: data where $x_{ij'} \leq x_{jk}^{**}$
      $(y_{\text{right}}, \mathbf{X}_{\text{right}})$: data where $x_{ij'} > x_{jk}^{**}$
      (where x_{jk}^{**} is the value corresponding to the sampled cutpoint c_{jk}^*)
    Call GrowFromRoot($y_{\text{left}}, \mathbf{X}_{\text{left}}, \Phi, \Psi, d+1, T,$ **left_node**)
    Call GrowFromRoot($y_{\text{right}}, \mathbf{X}_{\text{right}}, \Phi, \Psi, d+1, T,$ **right_node**)
  end if
end procedure

One can use XBCF to obtain efficient initial estimates (a "warm start") for the more standard BCF MCMC, potentially speeding up convergence to the target posterior, which has been proposed by Krantsevich, He and Hahn (2023) and is named Warm-Start BCF (ws-BCF).

For more information about the Grow-From-Root algorithm and warm-start BCF, please see He and Hahn (2021) and Krantsevich, He and Hahn (2023) respectively.

APPENDIX

B

THE IMPORTANCE OF ABLATION STUDIES FOR COMPLEX NONPARAMETRIC CAUSAL MODELS (APPENDIX)

B.1 $\alpha = 1$

Table 12 – Mean and Standard Deviation for Different Metrics for $\alpha = 1$

DGP	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
DGP1	RMSE_{CATE}	$\mathbf{0.316 \pm 0.113}$	0.322 ± 0.114	0.336 ± 0.114
	MAE_{CATE}	$\mathbf{0.277 \pm 0.113}$	0.281 ± 0.114	0.297 ± 0.115
	MAPE_{CATE}	$\mathbf{0.709 \pm 0.296}$	0.716 ± 0.294	0.718 ± 0.249
	Cover_{CATE}	0.951 ± 0.128	$\mathbf{0.938 \pm 0.138}$	0.944 ± 0.123
	Len_{CATE}	1.31 ± 0.195	$\mathbf{1.26 \pm 0.185}$	1.33 ± 0.199
	RMSE_{ATE}	$\mathbf{0.244 \pm 0.145}$	0.248 ± 0.148	0.268 ± 0.147
	MAE_{ATE}	$\mathbf{0.244 \pm 0.145}$	0.248 ± 0.148	0.268 ± 0.147
	MAPE_{ATE}	$\mathbf{0.488 \pm 0.289}$	0.496 ± 0.294	0.537 ± 0.292
	Cover_{ATE}	$\mathbf{0.93 \pm 0.256}$	$\mathbf{0.93 \pm 0.256}$	$\mathbf{0.93 \pm 0.256}$
	Len_{ATE}	0.963 ± 0.147	$\mathbf{0.961 \pm 0.147}$	1.00 ± 0.152
	RMSE_{π}	0.438 ± 0.00107	$\mathbf{0 \pm 0}$	0.0463 ± 0.00840
	MAE_{π}	0.438 ± 0.00112	$\mathbf{0 \pm 0}$	0.0377 ± 0.00550
DGP2	RMSE_{CATE}	$\mathbf{0.262 \pm 0.0819}$	0.273 ± 0.0830	0.282 ± 0.0876
	MAE_{CATE}	$\mathbf{0.222 \pm 0.0783}$	0.230 ± 0.0796	0.241 ± 0.0847
	MAPE_{CATE}	0.611 ± 0.231	0.612 ± 0.214	$\mathbf{0.611 \pm 0.185}$
	Cover_{CATE}	0.963 ± 0.0820	0.946 ± 0.112	$\mathbf{0.951 \pm 0.100}$
	Len_{CATE}	1.15 ± 0.147	$\mathbf{1.11 \pm 0.144}$	1.16 ± 0.160
	RMSE_{ATE}	$\mathbf{0.167 \pm 0.118}$	0.176 ± 0.118	0.194 ± 0.124
	MAE_{ATE}	$\mathbf{0.167 \pm 0.118}$	0.176 ± 0.118	0.194 ± 0.124
	MAPE_{ATE}	$\mathbf{0.335 \pm 0.236}$	0.353 ± 0.237	0.390 ± 0.248
	Cover_{ATE}	0.96 ± 0.197	$\mathbf{0.95 \pm 0.219}$	0.93 ± 0.256
	Len_{ATE}	0.755 ± 0.0874	$\mathbf{0.746 \pm 0.0859}$	0.800 ± 0.0975
	RMSE_{π}	0.366 ± 0.00342	$\mathbf{0 \pm 0}$	0.0809 ± 0.0116
	MAE_{π}	0.361 ± 0.00348	$\mathbf{0 \pm 0}$	0.0634 ± 0.00940
DGP3	RMSE_{CATE}	$\mathbf{0.262 \pm 0.0819}$	0.273 ± 0.0830	0.282 ± 0.0876
	MAE_{CATE}	$\mathbf{0.222 \pm 0.0783}$	0.230 ± 0.0796	0.241 ± 0.0847
	MAPE_{CATE}	0.611 ± 0.231	0.612 ± 0.214	$\mathbf{0.611 \pm 0.185}$
	Cover_{CATE}	0.963 ± 0.0820	0.946 ± 0.112	$\mathbf{0.951 \pm 0.100}$
	Len_{CATE}	1.15 ± 0.147	$\mathbf{1.11 \pm 0.144}$	1.16 ± 0.160
	RMSE_{ATE}	$\mathbf{0.167 \pm 0.118}$	0.176 ± 0.118	0.194 ± 0.124
	MAE_{ATE}	$\mathbf{0.167 \pm 0.118}$	0.176 ± 0.118	0.194 ± 0.124
	MAPE_{ATE}	$\mathbf{0.335 \pm 0.236}$	0.353 ± 0.237	0.390 ± 0.248
	Cover_{ATE}	0.96 ± 0.197	$\mathbf{0.95 \pm 0.219}$	0.93 ± 0.256
	Len_{ATE}	0.755 ± 0.0874	$\mathbf{0.746 \pm 0.0859}$	0.800 ± 0.0975
	RMSE_{π}	0.366 ± 0.00342	$\mathbf{0 \pm 0}$	0.0809 ± 0.0116
	MAE_{π}	0.361 ± 0.00348	$\mathbf{0 \pm 0}$	0.0634 ± 0.00940

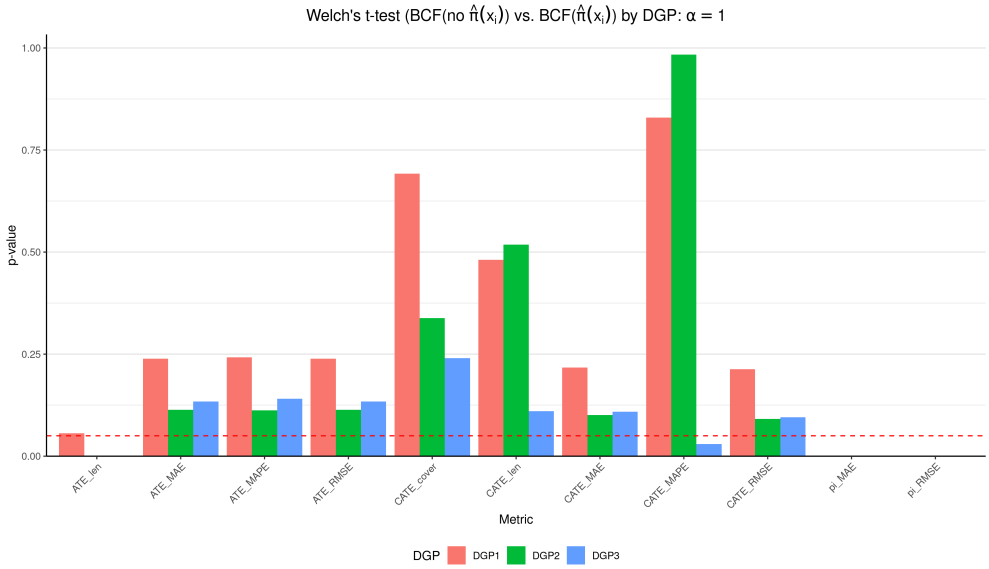

Figure 21 – Welch t-test p-values for BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) vs BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) for Main Ablation Study and $\alpha = 1$

B.2 $\alpha = 2$

Table 13 – Mean and Standard Deviation for Different Metrics for $\alpha = 2$

DGP	Variable	BCF (no $\hat{\pi}(\mathbf{X})$)	BCF ($\pi(\mathbf{X})$)	BCF ($\hat{\pi}(\mathbf{X})$)
DGP1	RMSE_{CATE}	$\mathbf{0.189 \pm 0.0795}$	0.193 ± 0.0831	0.192 ± 0.0930
	MAE_{CATE}	$\mathbf{0.169 \pm 0.0813}$	0.173 ± 0.0855	0.172 ± 0.0955
	MAPE_{CATE}	0.863 ± 0.456	0.870 ± 0.447	$\mathbf{0.862 \pm 0.490}$
	Cover_{CATE}	$\mathbf{0.997 \pm 0.0148}$	0.993 ± 0.0361	0.993 ± 0.0569
	Len_{CATE}	1.18 ± 0.164	$\mathbf{1.13 \pm 0.157}$	1.22 ± 0.164
	RMSE_{ATE}	$\mathbf{0.152 \pm 0.0981}$	0.157 ± 0.102	0.155 ± 0.111
	MAE_{ATE}	$\mathbf{0.152 \pm 0.0981}$	0.157 ± 0.102	0.155 ± 0.111
	MAPE_{ATE}	$\mathbf{0.605 \pm 0.388}$	0.625 ± 0.402	0.617 ± 0.438
	Cover_{ATE}	1.00 ± 0.000	$\mathbf{0.98 \pm 0.141}$	0.99 ± 0.1
	Len_{ATE}	0.885 ± 0.127	$\mathbf{0.871 \pm 0.127}$	0.933 ± 0.133
	RMSE_{π}	0.438 ± 0.00108	$\mathbf{0 \pm 0}$	0.0441 ± 0.00648
	MAE_{π}	0.438 ± 0.00113	$\mathbf{0 \pm 0}$	0.0361 ± 0.00524
DGP2	RMSE_{CATE}	$\mathbf{0.182 \pm 0.0750}$	0.187 ± 0.0750	0.192 ± 0.0799
	MAE_{CATE}	$\mathbf{0.161 \pm 0.0747}$	0.165 ± 0.0755	0.171 ± 0.0799
	MAPE_{CATE}	0.830 ± 0.415	0.829 ± 0.395	$\mathbf{0.829 \pm 0.387}$
	Cover_{CATE}	0.982 ± 0.0772	$\mathbf{0.973 \pm 0.0890}$	0.974 ± 0.0702
	Len_{CATE}	0.984 ± 0.159	$\mathbf{0.952 \pm 0.141}$	0.988 ± 0.155
	RMSE_{ATE}	$\mathbf{0.142 \pm 0.0930}$	0.146 ± 0.0943	0.154 ± 0.0967
	MAE_{ATE}	$\mathbf{0.142 \pm 0.0930}$	0.146 ± 0.0943	0.154 ± 0.0967
	MAPE_{ATE}	$\mathbf{0.569 \pm 0.374}$	0.587 ± 0.379	0.620 ± 0.388
	Cover_{ATE}	$\mathbf{0.94 \pm 0.239}$	$\mathbf{0.94 \pm 0.239}$	0.92 ± 0.273
	Len_{ATE}	0.665 ± 0.0916	$\mathbf{0.662 \pm 0.0792}$	0.696 ± 0.0838
	RMSE_{π}	0.366 ± 0.00326	$\mathbf{0 \pm 0}$	0.0792 ± 0.0106
	MAE_{π}	0.361 ± 0.00328	$\mathbf{0 \pm 0}$	0.0624 ± 0.00847
DGP3	RMSE_{CATE}	0.175 ± 0.0850	$\mathbf{0.168 \pm 0.0730}$	0.174 ± 0.0738
	MAE_{CATE}	0.150 ± 0.0832	$\mathbf{0.144 \pm 0.0716}$	0.150 ± 0.0734
	MAPE_{CATE}	0.982 ± 0.591	0.830 ± 0.466	$\mathbf{0.812 \pm 0.458}$
	Cover_{CATE}	0.973 ± 0.0993	$\mathbf{0.968 \pm 0.107}$	0.969 ± 0.113
	Len_{CATE}	0.917 ± 0.139	$\mathbf{0.878 \pm 0.134}$	0.883 ± 0.133
	RMSE_{ATE}	0.120 ± 0.0999	$\mathbf{0.113 \pm 0.0886}$	0.119 ± 0.0930
	MAE_{ATE}	0.120 ± 0.0999	$\mathbf{0.113 \pm 0.0886}$	0.119 ± 0.0930
	MAPE_{ATE}	0.482 ± 0.403	$\mathbf{0.456 \pm 0.358}$	0.480 ± 0.376
	Cover_{ATE}	0.91 ± 0.288	$\mathbf{0.93 \pm 0.256}$	$\mathbf{0.93 \pm 0.256}$
	Len_{ATE}	$\mathbf{0.584 \pm 0.0467}$	0.584 ± 0.0542	0.596 ± 0.0539
	RMSE_{π}	0.311 ± 0.00772	$\mathbf{0 \pm 0}$	0.121 ± 0.0125
	MAE_{π}	0.279 ± 0.00949	$\mathbf{0 \pm 0}$	0.0942 ± 0.00941

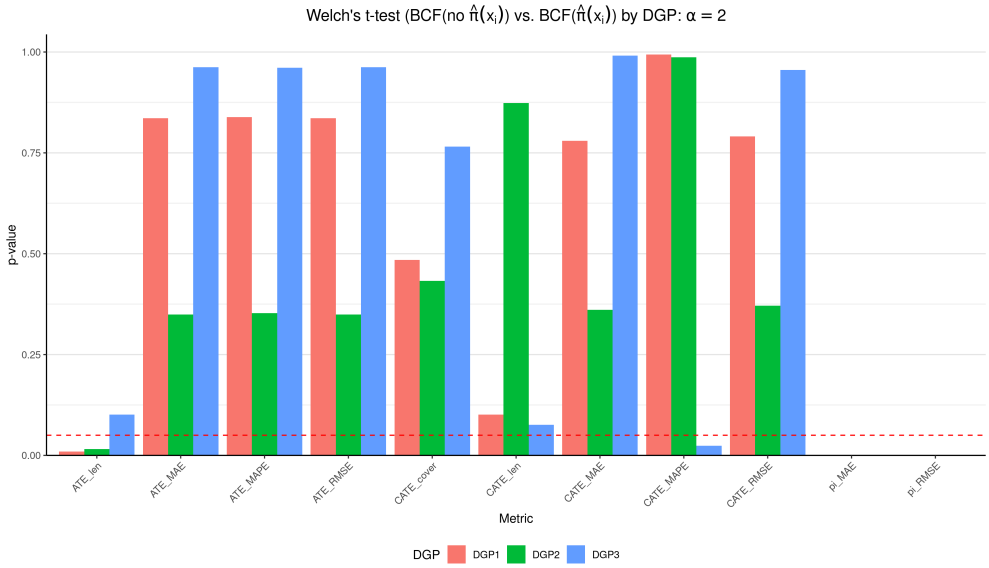

Figure 22 – Welch t-test p-values for BCF (no $\hat{\boldsymbol{\pi}}(\mathbf{X})$) vs BCF ($\hat{\boldsymbol{\pi}}(\mathbf{X})$) for Main Ablation Study and $\alpha = 2$

APPENDIX

C

FORESTS FOR DIFFERENCES: ROBUST CAUSAL INFERENCE BEYOND PARAMETRIC DID (APPENDIX)

C.1 Results and Discussion (TE=0)

To further assess the robustness of DiD-BCF and benchmark estimators, we conduct a series of placebo tests by setting the true treatment effect to zero ($\tau = 0$) across all Data Generating Processes (DGPs). This analysis is crucial for evaluating each model's ability to avoid finding spurious effects, a measure of its reliability and control over Type I error. In this context, lower error metrics values indicate less bias and a lower propensity to falsely detect a treatment effect. Since the true effect is zero, MAPE is not a meaningful metric and is excluded from these tables.

C.1.1 DGP 1

The results for DGP 1, where the estimand is the Average Treatment Effect on the Treated (ATT) and the true effect is zero, are presented in Table 14.

Table 14 – Overall Performance Comparison DGP 1 (True Treatment Effect = 0)

Metric	RMSE	MAE
Setting 1		
DiD-BCF	**0.0852 ± 0.0427**	**0.0607 ± 0.0340**
TWFE	0.1895 ± 0.0769	0.1490 ± 0.0733
DiD DR	0.5530 ± 0.1879	0.4378 ± 0.1598
DiD2s	0.1589 ± 0.0585	0.1256 ± 0.0503
SDiD	0.0954 ± 0.0623	0.0725 ± 0.0623
DoubleML_did	0.6598 ± 0.2233	0.5227 ± 0.1935
Setting 2		
DiD-BCF	**0.1189 ± 0.0639**	**0.0777 ± 0.0410**
TWFE	0.9237 ± 0.3548	0.7462 ± 0.3304
DiD DR	1.2927 ± 0.4652	1.0275 ± 0.4135
DiD2s	0.7528 ± 0.2782	0.5901 ± 0.2388
SDiD	0.4657 ± 0.2800	0.3731 ± 0.2800
DoubleML_did	1.4174 ± 0.5224	1.1308 ± 0.4940
Setting 3		
DiD-BCF	**0.1552 ± 0.0913**	**0.0996 ± 0.0655**
TWFE	0.9237 ± 0.3548	0.7462 ± 0.3304
DiD DR	1.2927 ± 0.4652	1.0275 ± 0.4135
DiD2s	0.7528 ± 0.2782	0.5901 ± 0.2388
SDiD	0.4648 ± 0.2790	0.3728 ± 0.2790
DoubleML_did	1.4787 ± 0.5510	1.1585 ± 0.4497

In Setting 1, the linear case, DiD-BCF emerges as the top-performing model, achieving the lowest RMSE (0.0852) and MAE (0.0607). This is a notable result, as it demonstrates that even in a simple, correctly specified linear environment, DiD-BCF is exceptionally well-calibrated to estimate a null effect. It slightly outperforms SDiD (RMSE 0.0954), which is the next best model. The other estimators, particularly DiD DR and DoubleML_did, exhibit substantially higher errors, suggesting they are more prone to finding spurious signals even under linearity settings.

The advantage of DiD-BCF becomes starkly evident in Setting 2, which introduces partial non-linearity. DiD-BCF's performance remains outstanding (RMSE 0.1189), with only a marginal increase in error. In stark contrast, all other benchmark models show a dramatic degradation in performance. SDiD's RMSE increases more than four-fold to 0.4657, and the errors for TWFE and DiD2s become very large. This indicates that when the parallel trends assumption holds but the functional form is misspecified, these other methods incorrectly attribute the unmodeled non-linearity to a treatment effect. DiD-BCF, by flexibly modeling the outcome, successfully disentangles the baseline trend from the (zero) treatment effect.

This pattern is cemented in Setting 3, the most challenging scenario with full non-linearity and a quadratic time trend. DiD-BCF is again the only model to produce reliable estimates, with a low RMSE of 0.1552. The other estimators are severely biased, producing estimates far from the true zero effect. This powerful result highlights DiD-BCF's robustness: it correctly identifies the absence of a treatment effect even in the presence of complex, non-linear confounding patterns that mislead all other methods.

C.1.2 DGP 2

Table 15 displays the results for DGP 2, which mimics an experimental setting with random treatment timing to estimate the Group-Average Treatment Effect (GATT).

Table 15 – Overall Performance Comparison DGP 2 (True Treatment Effect = 0)

Metric	RMSE	MAE
Setting 1		
DiD-BCF	0.1488 ± 0.0937	0.0967 ± 0.0812
TWFE	0.2207 ± 0.0842	0.1771 ± 0.0818
DiD DR	0.8469 ± 0.1979	0.6901 ± 0.1698
DiD2s	0.1171 ± 0.0461	0.0955 ± 0.0373
SDiD	$\mathbf{0.0913 \pm 0.0664}$	$\mathbf{0.0913 \pm 0.0664}$
DoubleML_did	1.1202 ± 0.2927	0.9062 ± 0.2366
Setting 2		
DiD-BCF	0.1928 ± 0.1328	0.1175 ± 0.1156
TWFE	0.3116 ± 0.1219	0.2453 ± 0.1122
DiD DR	0.9326 ± 0.2255	0.7623 ± 0.1918
DiD2s	0.1688 ± 0.0693	0.1366 ± 0.0585
SDiD	$\mathbf{0.1224 \pm 0.0916}$	$\mathbf{0.1224 \pm 0.0916}$
DoubleML_did	1.2489 ± 0.2829	1.0005 ± 0.2293
Setting 3		
DiD-BCF	0.2645 ± 0.2002	0.1424 ± 0.1726
TWFE	0.4123 ± 0.1579	0.3234 ± 0.1453
DiD DR	0.6604 ± 0.1917	0.5493 ± 0.1659
DiD2s	0.2230 ± 0.0870	0.1817 ± 0.0726
SDiD	$\mathbf{0.1687 \pm 0.1336}$	$\mathbf{0.1687 \pm 0.1336}$
DoubleML_did	0.7586 ± 0.2213	0.6148 ± 0.1777

In the linear Setting 1, SDiD shows the best performance (RMSE 0.0913), followed very closely by DiD2s (RMSE 0.1171). Their strong performance is consistent with the results when TE was non-zero and can be attributed to their efficiency in clean, randomized, and staggered treatment adoption settings where their underlying assumptions are met. DiD-BCF performs well (RMSE 0.1488), demonstrating that while it may have slightly

more variance than the specialized linear models in this simple case, it is not prone to significant bias and reliably estimates an effect close to zero.

As we introduce partial non-linearity in Setting 2, the top three performers remain the same: SDiD (RMSE 0.1224), DiD2s (RMSE 0.1688), and DiD-BCF (RMSE 0.1928). The random assignment of treatment timing appears to mitigate the biasing effects of functional form misspecification for these top models, as the unmodeled non-linearities are less likely to be correlated with treatment. All three prove adept at avoiding spurious findings.

In Setting 3, with full non-linearity, SDiD continues to lead (RMSE 0.1687), followed by DiD2s (RMSE 0.2230) and DiD-BCF (RMSE 0.2645). While SDiD's reweighting approach proves remarkably effective in this specific scenario, it is crucial to note that DiD-BCF remains a highly competitive and robust alternative. It consistently ranks among the top methods and, most importantly, its error remains low in absolute terms, confirming its ability to handle complex non-linearities without fabricating a treatment effect.

C.1.3 DGP 3

The results for DGP 3, representing a challenging observational setting with selection-on-observables, are presented in Table 16. This is a critical test of a model's ability to handle confounding.

Table 16 – Overall Performance Comparison DGP 3 (True Treatment Effect = 0)

Metric	RMSE	MAE
Setting 1		
DiD-BCF	$\mathbf{0.1273 \pm 0.0734}$	$\mathbf{0.0849 \pm 0.0633}$
TWFE	0.3547 ± 0.1576	0.2800 ± 0.1477
DiD DR	0.7553 ± 0.3529	0.6314 ± 0.2991
DiD2s	N/A	N/A
SDiD	0.1450 ± 0.1124	0.1450 ± 0.1124
DoubleML_did	3.0413 ± 4.3234	2.1047 ± 1.9597
Setting 2		
DiD-BCF	$\mathbf{0.1984 \pm 0.1189}$	$\mathbf{0.1320 \pm 0.1028}$
TWFE	1.3227 ± 0.5176	1.0680 ± 0.4924
DiD DR	3.1489 ± 1.3033	2.6181 ± 1.1663
DiD2s	N/A	N/A
SDiD	0.6061 ± 0.4679	0.6061 ± 0.4679
DoubleML_did	9.2586 ± 6.6156	6.8768 ± 3.8882
Setting 3		
DiD-BCF	$\mathbf{0.2195 \pm 0.1548}$	$\mathbf{0.1227 \pm 0.1281}$
TWFE	1.4859 ± 0.6109	1.1747 ± 0.5756
DiD DR	3.5792 ± 1.6355	2.9983 ± 1.4708
DiD2s	N/A	N/A
SDiD	0.6523 ± 0.5334	0.6523 ± 0.5334
DoubleML_did	8.1409 ± 4.2141	6.1865 ± 2.9206

N/A indicates that the model is not applicable for this DGP and Setting due to unbalanced panel data.

In Setting 1, which combines linearity with selection bias, DiD-BCF is the clear winner with the lowest RMSE (0.1273) and MAE (0.0849). It outperforms the next-best method, SDiD (RMSE 0.1450), while all other estimators show significantly higher errors. This demonstrates DiD-BCF's superior ability to control for selection bias by flexibly conditioning on the covariates that drive treatment assignment, thereby preventing confounding from creating a spurious treatment effect. The extremely high and variable errors of DoubleML_did again point to its instability and slow convergence.

The superiority of DiD-BCF becomes even more pronounced in Setting 2, where non-linearity is added to the selection bias. DiD-BCF is in a class of its own, maintaining a low RMSE of 0.1984. In contrast, the performance of all other methods deteriorates severely. SDiD's error triples (RMSE 0.6061), and the errors for TWFE and DiD DR become extremely large. This result underscores a core strength of our proposed model: it can simultaneously address confounding from both selection bias and complex functional forms, a scenario where other methods fail.

Finally, in the most difficult Setting 3, with full non-linearity and selection bias, DiD-BCF continues its exceptional performance (RMSE 0.2195), proving to be the only reliable estimator. The other methods are overwhelmed by the combined challenges, producing estimates that are heavily biased and far from the true null effect. These findings provide compelling evidence that DiD-BCF is a uniquely robust tool for modern DiD applications, capable of delivering accurate and reliable null estimates in complex observational settings where conventional and other machine learning-based methods are prone to significant error.


ICMC
USP
SÃO CARLOS