Hierarchical Bias-Driven Stratification for Interpretable Causal Effect Estimation

Lucile Ter-Minassian
University of Oxford, IBM Research Israel
lucile.ter-minassian@spc.ox.ac.uk

Liran Szlak
IBM Research Israel

Ehud Karavani
IBM Research Israel

Chris Holmes
University of Oxford

Yishai Shimoni
IBM Research Israel
yishais@il.ibm.com

Abstract

Modelling causal effects from observational data for deciding policy actions can benefit from being interpretable and transparent; both due to the high stakes involved and the inherent lack of ground truth labels to evaluate the accuracy of such models. To date, attempts at transparent causal effect estimation consist of applying post hoc explanation methods to black-box models, which are not interpretable. Here, we present BICauseTree: an interpretable balancing method that identifies clusters where natural experiments occur locally. Our approach builds on decision trees with a customized objective function to improve balancing and reduce treatment allocation bias. Consequently, it can additionally detect subgroups presenting positivity violations, exclude them, and provide a covariate-based definition of the target population we can infer from and generalize to. We evaluate the method's performance using synthetic and realistic datasets, explore its bias-interpretability tradeoff, and show that it is comparable with existing approaches.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

The goal of causal inference is to estimate the strength of a pre-specified intervention on some outcome of interest. It is therefore a valuable tool for data-driven decision-making in public health, economics, political science, and more. These are high-stakes, high-visibility domains in which data-based conclusions should ideally be understood by both the policy makers and the general public, which lack the expertise in the inner workings of statistical models. This highlights the importance of using *interpretable* models for estimating causal effects.

Following Rudin (2019), *interpretability* means that each decision in the algorithm is inherently explicit and traceable, contrasting with *explainability* where decisions are justified post-hoc using an external model. Therefore, when involving laypeople, interpretable models may often be preferred over explainable ones, as the latter might require understanding the assumptions and implications of an additional explainability model on top of the effect estimation model. Simplicity may incur better trustworthiness.

Interpretable or not, modeling causal effects requires estimating the *potential outcomes* - the outcome a person would experience under every possible intervention (Rubin, 1974). However, at any given time, a person can either be treated or not, but not both. This "fundamental problem of causal inference" (Holland, 1986) implies we never have ground truth labels and thus can never evaluate the accuracy of estimated causal effects. Furthermore, in nonexperimental settings where treatment is not assigned randomly, groups that do or do not receive treatment may not be comparable

in their attributes, and such attributes can influence the outcome too, introducing confounding bias. These two difficulties can harm the trustworthiness of data-driven conclusions perceived by both policy makers and the general public. However, by being explicitly transparent, interpretable models allow a new way to evaluate causal models by judging the sensibility of the conclusions they made and whether they adhere to one's common sense.

In this paper, we introduce BICauseTree: Bias-balancing Interpretable Causal Tree, an interpretable balancing method for observational data with binary treatment that can handle high-dimensional datasets. We use a binary decision tree to stratify on imbalanced covariates and identify subpopulations with similar propensities to be treated, when they exist in the data. The resulting clusters act as local naturally randomized experiments. This newly formed partition can be used for effect estimation, as well as propensity score or outcome estimation. Our primary objective is ATE estimation via transparent natural experiment identification. However, our method can further identify positivity-violating regions of the data space, i.e., subsets of the population where treatment allocation is highly unbalanced. By doing so, we generate a transparent, covariate-based definition of the target population, i.e. the population with sufficient overlap for inferring a causal effect.

Our contributions are as follows:

1. Our BICauseTree method can identify "natural experiments" i.e. subgroups with lower treatment imbalance, when they exist.
2. BICauseTree compares with existing methods for causal effect estimation in terms of bias while maintaining interpretability. BICauseTree is consistent and robust in the sub-populations it creates.
3. Our method provides users with a built-in prediction abstention mechanism for covariate spaces lacking common support. We show the value of defining the inferentiable population using a clinical example with matched twins data.
4. We show that under mild assumptions (weakly correlated variables, normally distributed data or Gaussian copula model), our framework is guaranteed to decrease the treatment imbalance.
5. The resulting tree can further be used for propensity or outcome estimation.
6. We release open-source code with detailed documentation for implementation of our method, and reproducibility of our results.

2 RELATED WORK

2.1 Effect estimation methods

Causal inference provides a wide range of methods for effect estimation from data with unbalanced treatment allocation. There are two modelling strategies: *balancing* methods, and *outcome* models.

In *balancing* methods, such as matching or weighting methods, the data is pre-processed to create subgroups with lower treatment imbalance or "natural experiments". *Matching* methods consist of clustering (based on a distance metric) similar units from the treatment and control groups to reduce imbalance. However, as the notion of distance becomes problematic in high dimensional spaces, covariate-based matching tends to become ineffective (Beyer et al., 1999). *Weighting* methods aim at balancing the covariate distribution across treatment groups, with Inverse Probability Weighting (IPW) (Horvitz and Thompson, 1952) being the most popular approach. Sample weights are the inverse of the estimated *propensity* scores, i.e. the probability of a unit to be assigned to its observed group. However, extreme IPW weights can increase the estimation variance.

Contrastingly, in *outcome* models the causal effect is estimated from regression outcome models where both treatment and covariates act as predictors of the outcome. These regressions can be fitted through various methods like linear regression (Imbens and Rubin, 2015), neural networks (Shi et al., 2019; Shalit et al., 2017a), or tree-based models (Athey and Imbens, 2016).

Under this taxonomy, BICauseTree is a *balancing* method, i.e., a data-driven mechanism for achieving conditional exchangeability.

2.2 Positivity violations

Causal inference relies on the *positivity* assumption, which requires covariate overlap between treatment groups. Violations occur when a subgroup rarely or never receives treatment, leading to unreliable causal estimates (Karavani et al., 2019).

Positivity can be handled by estimating propensity scores and trimming extreme values (Petersen et al., 2012), but this reduces interpretability regarding excluded subjects and thus on generalizability. Other methods attempt to improve this by characterizing exclusions or directly assessing covariate overlap without propensity scores (Oberst et al., 2020; Wolf et al., 2021; Ackerman et al., 2020; Karavani et al., 2019).

Unlike these approaches, BICause Tree integrates pos-

itivity identification into the model, with an interpretable abstention mechanism for effect estimation.

2.3 Interpretability and causal inference

A critical challenge in many effect estimation methods is their lack of interpretability. A model is *interpretable* when its decisions are inherently transparent, like decision trees where outcomes are easily traced through logical conjunctions. In contrast, *explainable* models rely on post-hoc justifications using explanation tools like Shapley values (Lundberg and Lee, 2017) or LIME (Ribeiro et al., 2016). However, these techniques are often unstable and depend heavily on the model's accuracy (Mittelstadt et al., 2019). Therefore, many practitioners advocate for inherently interpretable models (Rudin, 2019), particularly for causal inference, where effect estimation impacts high-stakes decisions involving laypeople.

Beyond effect estimation, the characterization of positivity violations must also be interpretable for stakeholders like policymakers. Excluding samples can reduce external validity, potentially excluding a structurally biased subpopulation. Interpretable identification of the overlap helps policymakers understand to whom the study results apply (Oberst et al., 2020; Karavani et al., 2019).

BICauseTree directly generates a covariate-based description of the excluded subgroups, clarifying the target population for which the inferred effect is valid.

2.4 Resulting objectives

Our aim is to bridge the aforementioned gaps in the literature. In turn, our goals are: (i) unbiased estimation of causal effect, (ii) interpretability of both the balancing and positivity violation identification procedures, (iii) ability to handle high-dimensional datasets.

3 BICAUSETREE

Our BICauseTree approach balances observational datasets with the goal of estimating a causal effect in a subpopulation with sufficient overlap. BICauseTree utilizes the Absolute Standardized Mean Difference (ASMD) (Austin, 2009) frequently used for assessing potential confounding bias in observational data. Note that our balancing procedure is entirely interpretable, although it can be used in combination with arbitrary black-box outcome models or propensity score models. Finally, our method generates a covariate-based definition of the target population with with reasonable overlap, where our inference is valid.

3.1 Formal causal model

We consider a dataset of size n. For each individual sample i, we denote (X_i, T_i, Y_i) with $X_i \in \mathbb{R}^d$ a covariate vector for sample i measured prior to binary treatment allocation T_i. In Rubin (1973)'s potential outcomes framework, $Y_i(1)$ is the outcome under $T_i = 1$, and $Y_i(0)$ is the analogous outcome under $T_i = 0$. Then, assuming the consistency assumption, the observed outcome is defined as $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. Here, we aim to estimate the average treatment effect (ATE), defined as: $\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$.

3.2 Algorithm

The intuition for our algorithm is that, by partitioning the population to maximize treatment allocation heterogeneity, we may be able to find subpopulations that are natural experiments. We recursively partition the data according to the most imbalanced covariate between treatment groups. Using decision trees makes our approach transparent and non-parametric.

Splitting criterion The first step of our algorithm is to split the data until some stopping criterion is met. The tree recursively splits on the covariate that maximizes treatment allocation heterogeneity. To do so, we compute the Absolute Standardized Mean Difference (ASMD) for all covariates and select the covariate with the highest absolute value. The ASMD for X_j is:

$$\text{ASMD}_\text{j} = \frac{|\mathbb{E}[X_j|T{=}1] - \mathbb{E}[X_j|T{=}0]|}{\sqrt{Var([X_j|T{=}1]) + Var([X_j|T{=}0])}}$$

Since we assume all confounders are observed, the reason for choosing the one with the highest ASMD is that it is most likely to cause the most confounding bias, and therefore adjusting for it will likely minimize the residual confounding bias (Figures A1 and A2). Once that next splitting covariate j_{max} is chosen, we want to find a split that is most associated with treatment assignment, so that we may control for the effect of confounding. The tree finds the optimal splitting value by iterating over covariate values $x_{j_{max}}$ and taking the value associated with the lowest p-value from a Fisher's exact test or a χ^2 test, depending on the sample size.

Stopping criterion The tree-building process terminates when either: (i) the maximum ASMD is below a threshold, (ii) the minimum treatment group size falls below a threshold (iii) the total population falls below a threshold, or (iv) a maximum tree depth is reached. All thresholds are user-defined hyperparameters.

Pruning procedure Once the stopping criterion is met in all leaf nodes, the tree is pruned. A multi-

ple hypothesis test correction is first applied on the p-values of all splits. Following this, the splits with significant p-values or with at least one split with significant p-value amongst their descendants are kept. The implementation of the tree allows for user-defined multiple hypothesis test correction, with current experiments using Holm-Bonferonni (Holm, 1979; Abdi, 2010). The choice of the pruning and stopping criterion hyperparameters will guide the bias/variance trade-off of the tree. Deeper trees may have more power to detect treatment effect while shallower trees will be more likely to have biased effect estimation.

Positivity violation filtering Lastly, we evaluate the overlap in the resulting set of leaf nodes to identify those where inference isn't possible. Treatment balance is evaluated using a user-defined overlap estimation method, with the default method being the Crump procedure (Crump et al., 2009). The positivity-violating leaf nodes are tagged and excluded from inference.

Estimation Once a tree is contracted, it can be used to estimate both counterfactual outcomes and propensity scores. For each leaf, using the units propagated to that leaf, we can model the counterfactual outcome by taking the average outcome of those units in both treatment groups. Alternatively, we can fit any arbitrary causal model (e.g., IPW or an outcome regression) to obtain the average counterfactual outcomes in that leaf (Kang and Schafer, 2007; Quinlan et al., 1992). The ATE is then obtained by averaging the estimation across leaves. Similarly, we can estimate the propensity score in each leaf by taking the treatment prevalence or using any other estimator.

Algorithm 1 BICauseTree

Inputs: root node N_0, X, T, Y
Call *Build subtree*(N_0, X, T, Y)
Do multiple hypothesis test correction on all split p-values
Pruning procedure: keep splits with either (i) a significant p-value or (ii) at least one descendant with a significant p-value
Mark leaf nodes that violate positivity criterion

Algorithm 2 Build subtree

Inputs: current node N, X, T, Y
if Stopping criteria not met **then**
Find and record in N the covariate with maximum ASMD: $maxASMD := max_i(ASMD_i)$
Find and record in N the split value with the lowest p-value according to a Fisher test/χ^2 test
Record the p-value for this split in N
Split the data X, T, Y into $X_{left}, T_{left}, Y_{left}$ and $X_{right}, T_{right}, Y_{right}$ according to N's splitting covariate and value
Add two child nodes to N: N_{left} and N_{right}
Call *Build subtree*(N_{left}, X_{left}, T_{left}, Y_{left})
Call *Build subtree*(N_{right}, X_{right}, T_{right}, Y_{right})
end if

Decrease in treatment allocation imbalance

Lemma 1 (ASMD Decrease in Node Splitting)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_p)$ *be a vector of continuous and/or categorical variables,* T *a binary treatment indicator, and assume either multivariate normality or a Gaussian copula model with latent variables* $\mathbf{Z}$*. If the following conditions hold:*

1. $|\rho_{ij}| \leq \delta \leq \frac{1-K}{CM^2}$ *for all* $i \neq j$*, where* C *is a constant dependent on the split point*
2. $ASMD_i^{parent} \leq K \cdot ASMD_j^{parent}$ *for all* $i \neq j$*, with* $K < 1$
3. *For continuous variables,* $\frac{\sigma_i}{\sigma_j} \leq M$ *for all* $i \neq j$*, with* $M \geq 1$
4. *For continuous variables,* $Var(X_i|T = 1) = Var(X_i|T = 0) = \sigma_i^2$
5. *For categorical variables, category thresholds are constant across treatment groups*

Then, after splitting on the variable X_j *with the highest ASMD, the maximum ASMD in child nodes will not exceed the parent node's maximum ASMD. The ASMD imbalance in children nodes either decreases or remains the same. This extends to cases with unequal variances.*

Section A.3 provides a proof of convergence.

Code and implementation BICauseTree code is released open-source under: `https://github.com/IBM-HRL-MLHLS/BICause-Trees`. Our flexible implementation allows users to define a stopping criteria as well as a multiple hypothesis correction method. BICauseTree adheres to `causallib`'s API, and can accept various outcome and propensity models.

4 EXPERIMENTS AND RESULTS

4.1 Experimental settings

In all experiments–unless stated otherwise–the data was split into a training and testing set with a 50/50 ratio. The training set was used for the construction of the tree and for fitting the outcome models in leaf nodes, if relevant. Causal effects are estimated by taking a weighted average of the local treatment effects in each subpopulation. At the testing phase, the data

is propagated through the tree, and potential outcomes are evaluated using the previously fitted leaf outcome model.

We performed 50 random train-test splits, which we will refer to as *subsamples*, to avoid confusion with the tree partitions. For each subsample, effects are only computed on the non-violating samples of the population, as defined by BICauseTree. Specifically, we first used BICauseTree to filter overlap-violating observations and then applied all models on that same remaining sample. This ensured i) a fairer comparison, and ii) smaller likelihood of observations with extreme propensity scores.

To maintain a fair comparison, these samples are also excluded from effect estimation with other models and with ground truth.

Baseline comparisons We compare our method to the three most interpretable methods amongst existing approach –double Mahalanobis Matching, Inverse Probability Weighting (IPW), and Causal Tree (CT).

In **Mahalanobis Matching** (Rubin, 1980; Stuart, 2010), the nearest neighbor search operates on the Mahalanobis distance: $d(X_i, X_j) = (X_i - X_j)^T \Sigma^{-1} (X_i - X_j)$, where Σ is the estimated covariance matrix of the control or treatment group.

In **Inverse Probability Weighting** (Horvitz and Thompson, 1952), a propensity score model estimates the individual probability of treatment conditional on the covariates. The data is then weighted by the inverse propensities $P(T = t_i \mid X = \boldsymbol{x}_i)^{-1}$ to generate a balanced pseudo-population.

In **Causal Tree** (Athey and Imbens, 2016), the splitting criterion optimizes for treatment *effect* heterogeneity (see section A.4 for further details). We use a Causal Tree and not a Causal Forest to compare to an estimator which is equally interpretable as our estimator. Although both BICauseTree and CT utilize decision trees, BICauseTree splits are optimized for balancing treatment *allocation* whilst CT splits are optimized for balancing treatment *effect*, under assumed exchangeability. In other words, CT *assumes* exchangeability while BICauseTree *finds* exchangeability. We also compare our results to an unadjusted marginal outcome estimator, which will act as our "dummy" baseline model.

As using a single Causal Tree for our interpretability goal gives rise to high estimation bias, Causal Tree was excluded from the main manuscript for scaling purposes. Finally, BiCause Tree also shares some similarities with Coarsedened Exact Matching (CEM) Iacus et al. (2012a), but the strength of our approach lies in performing "blocking" in a principled way, based on ASMD.

We thus refer the reader to sections A.6, A.7 and A.8 for a comparison with both CT and CEM. For synthetic experiments, we use the simplest version of our tree which we term BICauseTree(Marginal) where the effect is estimated by taking average outcomes in leaf nodes. For real-world experiments, we compare BICauseTree(Marginal) with BICauseTree(IPW), an augmented version in which an IPW model is fitted in each leaf node.

To compare estimation methods, we compute the difference between the estimated ATE and the true ATE for each subsample (or train-test partition) and display the resulting distribution of estimation biases in a box plot. Further experimental details, including hyperparameters, can be found in the Appendix under sections A.11 and A.5.

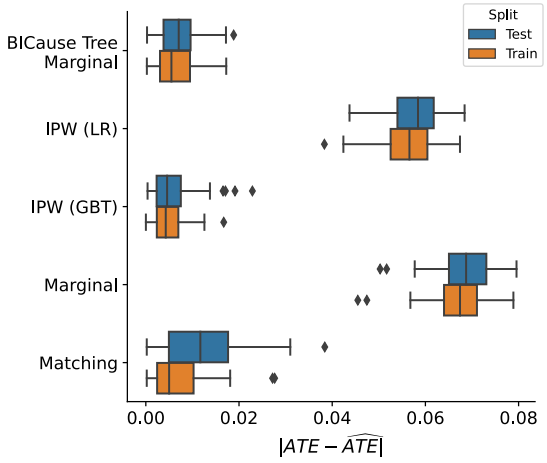

Figure 1: Estimation bias for the natural experiment dataset across 50 subsamples, with $N = 20,000$.

4.2 Synthetic datasets

We evaluate BICauseTree on two synthetic datasets: the "natural experiment dataset" to demonstrate its ability to identify subgroups with lower treatment imbalance, and the "positivity violations dataset" to showcase its identification of positivity-violating samples. Due to the interaction-based nature of the data generation procedure, we additionally compare our approach to an IPW estimator with a Gradient Boosting classifier propensity model, referred to as IPW (GBT) in both synthetic experiments. This choice ensures a fair comparison across estimators.

Identifying natural experiments For the natural experiment dataset, we used a Death outcome D, binary treatment T and two covariates: Sex S and Age A, thus $X = (S, A) \in \mathbb{R}^2$. We defined four subpopulations, each constituting a natural experiment with a truncated normal propensity distribution centered around a pre-defined mean and variance (see Section A.5.1). Then, individual treatment propensities were sampled from the corresponding distribution, and observed treatment values were sampled from a Bernoulli parameterized with the individual propensities. No positivity violation was modeled. The sample size is $N = 20,000$. The marginal distribution of covariates follows: $S \sim \text{Ber}(0.5), A \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu = 50, \sigma = 20$.

The partition obtained from training BICauseTree on the entire dataset is shown in Figure A4. Our tree successfully identifies the subpopulations in which a natural experiment was simulated. Figure 1 shows the estimation bias across subsamples. In addition to being transparent, BICauseTree has lower bias in causal effect estimation compared to all other methods, excluding IPW(GBT) which has comparable performance.

Despite its higher estimation variance, Matching has low bias, probably due to covariate space being well-posed and low-dimensional. Contrastingly, the logistic regression in IPW(LR) is not able to model treatment allocation as the true propensities are generated from a noisy piecewise constant function of the covariates resulting in a threshold effect that explains its poor performance. The non-parametric, local nature of both Matching and BICauseTree thus contrasts with the parametric estimation by IPW(LR).

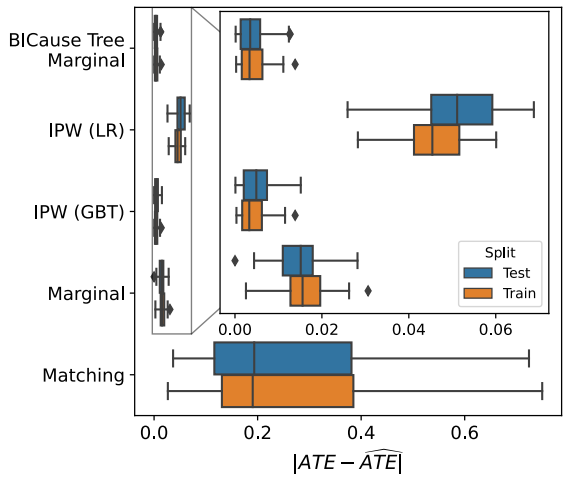

Figure 2: Estimation bias across 50 subsamples, positivity violations dataset, excluding positivity violating leaf nodes.

Identifying positivity violations For the positivity violations dataset, we consider a synthetic dataset with a Death outcome D, a binary treatment of interest T, and three Bernoulli covariates –Sex S, cancer C and arrhythmia A– such that $X = (S, C, A)$ (see Section A.5.2 for further details). As for the natural experiment dataset, we modeled treatment allocation with stochasticity by sampling propensities from a truncated Gaussian distribution first. Treatment allocation was simulated to ensure that overlap is very limited in two subpopulations: females with no cancer and no arrhythmia are rarely treated, while males with cancer and arrhythmia are almost always treated.

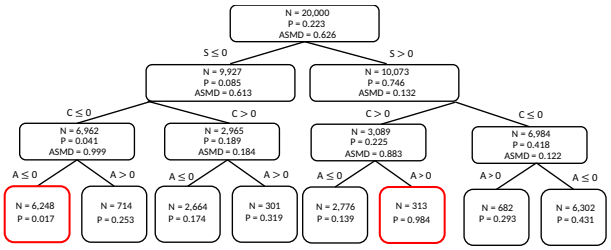

Figure 3: Tree structure after training on the entire positivity violations dataset. Violating leaf nodes are in red. P is for treatment prevalence, ASMD is for maximum ASMD.

Figure 3 shows the partition obtained from training BICauseTree on the entire dataset, where BICauseTree excludes the subgroups where positivity violations were modeled. This further shows the interpretability of the covariate-based definition of our target population, i.e. the overall cohort, excluding men without cancer and arythmia, and women with cancer and arythmia. On average, 67.1% of the cohort remained after positivity filtering with very little variability across subsamples. Maximum ASMD is null in all leaf nodes.

As seen in Figure 2, BICauseTree's effect estimation remains unbiased with low variance after filtering. Our estimator compares favorably with IPW(GBT) while maintaining interpretability. IPW(LR) is more biased, likely due to extreme weights in the initial cohort. Despite filtering samples with overlap violations, the remaining propensity weights may still induce bias. Variance is comparable across methods, except Matching, which is both more biased and has higher variance. Further calibration results for BICauseTree are in the Appendix, section A.6.2.

4.3 Realistic datasets

Causal benchmark datasets We use two causal benchmark datasets to show the value of our approach. The twins dataset illustrates the high applicability of our procedure to clinical settings. It is based on real-world records of $N = 11,984$ pairs of same-sex twin

births and has 75 covariates. It tests the effect of being born the heavier twin (i.e. the treatment) on death within one year (i.e. the outcome). We use the dataset generated by Neal et al. (2020), which simulates an observational study from the initial data by selectively hiding one of the twins with a generative approach. We also ran our analysis on the *2016 Atlantic Causal Inference Conference* (ACIC) semisynthetic dataset with simulated outcomes (Hahn et al., 2019). For ACIC, given that trees are data greedy, and due to the smaller sample size ($N = 4,802$) relative to the number of covariates ($d = 79$), the models were trained on 70% of the dataset.

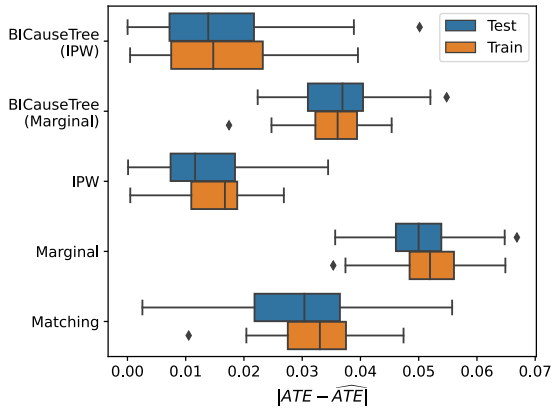

Figure 4: Estimation bias for the twins dataset across 50 subsamples, excluding positivity violating leaf nodes.

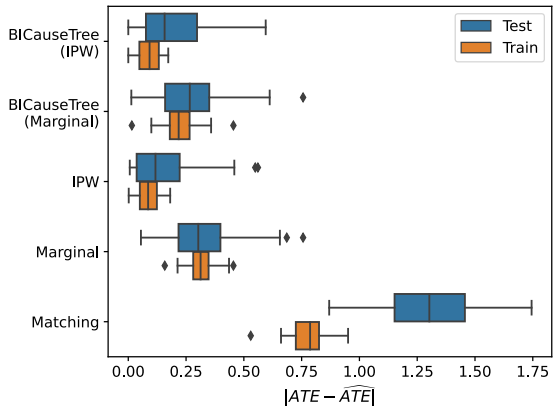

Figure 5: Estimation bias for the ACIC dataset across 50 subsamples, excluding positivity-violating leaf nodes.

Effect estimation Figure 4 shows the distribution of estimation bias across subsamples on the twins dataset, compared to baseline models. Our BICauseTree(Marginal) estimator is less biased than the marginal estimator. Augmenting it with an IPW outcome model (BICauseTree(IPW)) further reduces bias, making it comparable to IPW in both bias and variance. Figure 5 compares estimation bias on the ACIC dataset, showing that both BICauseTree models perform similarly to IPW in bias and variance.

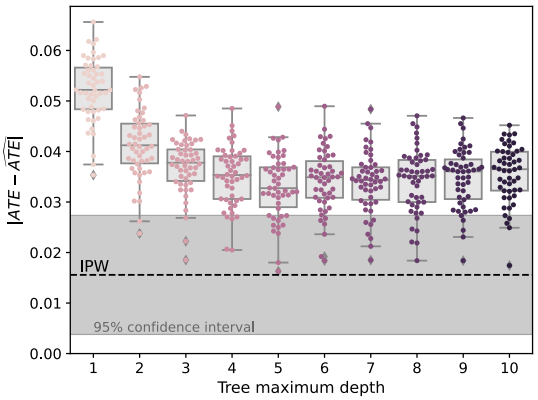

Figure 6: Estimation bias for *BICauseTree(Marginal)* with varying maximum depth parameter, and average bias of IPW (dotted), on the twins training set.

Bias-interpretability tradeoff We expect a bias-interpretability tradeoff: deeper trees are less biased but harder to interpret, while shallower trees are more transparent but less accurate. Figure 6 shows how leaf node bias decreases as we increase the maximum depth of BICauseTree(Marginal) in the twins dataset (see Figure A23 for ACIC). Each circle represents bias for a subsample, with the dotted line showing average bias using IPW and the shaded area representing its 95% CI. The overlap suggests this tradeoff, with deeper trees requiring complex outcome models to reduce bias. Augmenting BICauseTree with IPW lowers bias but reduces transparency.

Ultimately, figure 6 shows that bias reduction is consistent beyond a maximum depth parameter of 5. This robustness w.r.t the maximum depth hyperparameter may be due to the pruning step.

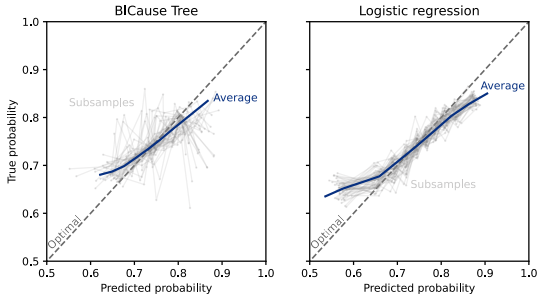

Figure 7: Calibration of propensity score, twins dataset

Interpretable positivity violations filtering As previously discussed, BICauseTree provides a built-in method for identifying positivity violations in the covariate space directly. After positivity filtering, the effect was computed on an average of 99.5% ($\sigma = 0.006$)

of the population on the twins dataset, and an average of 85.9% ($\sigma = 0.093$) of the ACIC dataset.

Figure A16 shows the tree partition for the twins dataset, identifying one leaf node with positivity violations ($N = 106$). This example highlights the real-world relevance of defining a covariate-based non-violating subpopulation; we can assert that the estimated effect of being born heavier may not apply to newborns in this violating node. BICauseTree's ability to make such distinctions is crucial in safety-sensitive settings. For instance, if the newborns in that violating subgroup are at higher risk, applying the ATE to the entire cohort could deny them necessary follow-up, increasing their risk of death.

Knowing who these newborns are would not be possible using an alternative non-interpretable model that provides opaque exclusion criteria, like propensity trimming. Note that, unlike for effect estimation, the positivity identification remains transparent regardless of the chosen propensity or outcome model at the leaves.

User study for interpretability We further evaluate the interpretability of our method compared to our baselines by conducting interviews. Participants had expertise in Statistics, but no expertise in causal inference. On average, results showed increased intuitiveness and clarity of our method as well as lower time to understand our method compared to others. Details are given in the Appendix under section A.9.

Propensity score estimation BICauseTree can also act as a propensity model, modelling the treatment at its leaf nodes. Given the importance of calibrated propensity scores (Gutman et al., 2022), Figure 7 compares the calibration of the propensity score estimation of BICauseTree with the one from logistic regression (IPW) on the testing set of the twins dataset (Figure A22 for ACIC). As expected, logsitic regression, which has better data efficiency, has less-noisy calibration. However, BICauseTree still shows satisfying calibration on average. See Appendix for propensity score calibration for the synthetic natural experiments and calibration of BICauseTree for the *outcome* variable.

Tree consistency To evaluate the consistency of our clustering across subsamples, we train our tree on 70% of the dataset and compute the adjusted Rand index (Rand, 1971) (further details in section A.2). We chose not to train on 50% of the data here as most of the inconsistency would then be due to the variance between subsamples. For the twins dataset, the Rand index across 50 subsamples of size $N = 8,388$, is 0.633 ($\sigma = 0.208$). For the ACIC dataset, the Rand index across 50 subsamples of size $N = 3,361$, is 0.314 ($\sigma = 0.210$) which shows that our tree is not consistent across subsamples if the sample size is not substantial. However, we exemplify consistent identification of the positivity population, with the variance of the percentage of positive samples being $\sigma = 0.006$ and $\sigma = 0.093$ (see paragraph 4.3) in the twins and ACIC dataset respectively. Throughout our experiments, we noticed how consistency starts to decrease if the maximum depth hyperparameter increases past a certain threshold, and thus recommend users test the tree consistency across subsamples when tuning this hyperparameter.

5 DISCUSSION

We have introduced a decision tree-based model, with a customized splitting criterion, able to estimate the ATE in a transparent, interpretable way. Additionally, the model can also identify subgroups violating the positivity assumption, characterize them in covariate space, and abstain from estimating an effect on them. Our experimental results on both synthetic data and realistic data show our method has comparable accuracy to existing baselines. Our user study confirms its interpretability is either far superior, or comparable with methods that aren't efficient in high dimensions.

Real-world relevance of our model While our model may retain residual confounding bias, its high interpretability meets the needs of conservative sectors like healthcare, where causal methods face skepticism due to the lack of ground truth (Department of Health and Human Services, Food and Drug Administration, 2023).

In high-stakes applications like risk assessment and policy decisions, our approach balances performance with transparency, enabling expert scrutiny and fostering trust. Its simplicity also makes it practical for low-resource settings or real-time scenarios, such as generating risk scores for patient prioritization in hospitals based on causal effect measures, even with limited computational resources or time (Erasmus et al., 2021; Kellett et al., 2021).

Overlap filtering Additionally, BICauseTree can detect, characterize, and exclude subgroups violating the overlap assumption, improving internal validity by ensuring outcome extrapolation is data-driven rather than model-based. It also enhances external validity by informing policymakers about the population—explicitly defined in covariate space—to which the estimated effect is expected to generalize.

On top of being intuitive, BICauseTree inherits additional desirable strengths from decision trees, like identifying complex interactions in a flexible, data-adaptive,

non-parameteric way in high-dimensional data (see Figure A3), with a computational expense comparable to IPW and Causal Forest (Sani et al., 2018).

Limitations While our method inherits the interpretability of decision trees, it also inherits their weaknesses, such as lower data efficiency compared to parametric estimators like IPW. Theoretically, this leads to higher variance, but since IPW is a low-precision model, our experiments show comparable variance.

Additionally, decision trees use fewer samples the deeper they grow, decreasing the precision of the estimated splitting criterion. Therefore, while the mean of the estimated ASMD may be independent of sample size, its variance is dependent and can become less precise (Austin, 2009).

This holds true to our ASMD-based splitting criterion. Furthermore, choosing covariates for split is based solely on ASMD, which, first, is a marginal measure, and second, does not regard the association of covariate and outcome (see A.6.1).

Future work Additional research could perfect some of the limitations presented above. First, the pruning procedure correcting for multiple hypotheses can avoid assuming independence and leverage the hierarchical tree structure (Behr et al., 2020; Allen, 1997; Malek et al., 2017). Second, following the "honest effect" framework (Athey and Imbens, 2016) we can further split the sample, using one split to structure the tree and another to fit the node models; though this will require even more data. Third, we currently fit models at each leaf independently, but since our target estimand is the ATE, we can partially pool estimates across leaves (clusters), fitting a multilevel outcome model with varying intercepts or slopes for treatment coefficients (Feller and Gelman, 2015). Lastly, future work may further explore excluding terminal leaf nodes with high imbalance (as they might not enclose natural experiments), or experiment with partially applying additional covariate adjustments in those unbalanced leaves, while balanced uses a simple, interpretable average for estimation.

Limitations notwithstanding, we claim BICauseTree is very relevant when causation is examined in high-stake, public-facing domains. Its transparency can inspire trust among laypeople, while any concerns about its accuracy may be mitigated by fitting an ensemble of causal models and examining if the BICauseTree estimator is an outlier. We provide an open-source implementation of BICauseTree in Python adhering to `causallib`'s API to maximize ease of use and encourage practitioners to adopt it in their work.

Acknowledgements

This work was conducted while LTM was a PhD student at the University of Oxford and an intern at IBM Research Israel. We sincerely thank our reviewers for their valuable feedback, which helped improve this manuscript. Additionally, we are grateful to Michael Danziger for his careful proofreading of this article.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.

Abdi, H. (2010). Holm's sequential bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8.

Ackerman, S., Farchi, E., Raz, O., Zalmanovici, M., and Dube, P. (2020). Detection of data drift and outliers affecting machine learning model performance over time. *arXiv preprint arXiv:2012.09258*.

Allen, M. P. (1997). Testing hypotheses in nested regression models. *Understanding regression analysis*, pages 113–117.

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

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Ann. Statist*, 47:1148–1178.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., and Syrgkanis, V. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.x.

Behr, M., Ansari, M. A., Munk, A., and Holmes, C. (2020). Testing for dependence on tree structures. *Proceedings of the National Academy of Sciences*, 117(18):9787–9792.

Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., and Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Crump, R. K., Hotz, V. J., Imbens, G., and Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Department of Health and Human Services, Food and Drug Administration (2023). Using artificial intelligence and machine learning in the development of drug and biological products; availability. `https://www.fda.gov/media/167973/download?attachment`. [Docket No. FDA–2023–N–0743].

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Erasmus, A., Brunet, T. D., and Fisher, E. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, 47(5):329–335.

Feller, A. and Gelman, A. (2015). Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource*, pages 1–16.

Gutman, R., Karavani, E., and Shimoni, Y. (2022). Propensity score models are better when post-calibrated. *arXiv preprint arXiv:2211.01221*.

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

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

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

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.

Iacus, S. M., King, G., and Porro, G. (2012a). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Iacus, S. M., King, G., and Porro, G. (2012b). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*.

Karavani, E., Bak, P., and Shimoni, Y. (2019). A discriminative approach for finding and characterizing positivity violations using decision trees. *arXiv preprint arXiv:1907.08127*.

Kellett, J., Wasingya-Kasereka, L., Brabrand, M., and Opio, M. (2021). Two simple replacements for the triage early warning score to facilitate the south african triage scale in low resource settings. *African Journal of Emergency Medicine*, 11(1):45–51.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political analysis*, 27(4):435–454.

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

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems*, pages 6449–6459.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Malek, A., Katariya, S., Chow, Y., and Ghavamzadeh, M. (2017). Sequential multiple hypothesis testing with type i error control. In *Artificial Intelligence and Statistics*, pages 1468–1476. PMLR.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Neal, B., Huang, C.-W., and Raghupathi, S. (2020). Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*.

Oberst, M., Johansson, F., Wei, D., Gao, T., Brat, G., Sontag, D., and Varshney, K. (2020). Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54.

Quinlan, J. R. et al. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203.

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

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, pages 293–298.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sani, H. M., Lei, C., and Neagu, D. (2018). Computational complexity analysis of decision tree algorithms. In *Artificial Intelligence XXXV: 38th SGAI International Conference on Artificial Intelligence, AI 2018, Cambridge, UK, December 11–13, 2018, Proceedings 38*, pages 191–197. Springer.

Seaman, S. R. and Vansteelandt, S. (2018). Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184.

Shalit, U., Johansson, F. D., and Sontag, D. (2017a). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shalit, U., Johansson, F. D., and Sontag, D. (2017b). Representation learning for treatment effect estimation from observational data. In *Proceedings of the*

31st International Conference on Neural Information Processing Systems, pages 6088–6098.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

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

Wolf, G., Shabat, G., and Shteingart, H. (2021). Positivity validation detection and explainability via zero fraction multi-hypothesis testing and asymmetrically pruned decision trees. *arXiv preprint arXiv:2111.04033*.

Xu, R., Li, S., Wu, Y., Wang, L., and Ravikumar, P. (2020). Optimal transport for treatment effect estimation. In *Proceedings of the 37th International Conference on Machine Learning*.

Yampolskiy, R. V. (2018). *Artificial intelligence safety and security*. CRC Press.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. *arXiv preprint arXiv:1804.07464*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, we provide assumptions, notations and pseudo-code for our algorithm.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, we explicit compute time and sample size, and explain that BICauseTree inherits the computational complexity from decision trees, all in section A.9. of the Appendix.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Not Applicable.
   (b) Complete proofs of all theoretical results. Not Applicable.
   (c) Clear explanations of any assumptions. Not Applicable.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, we have provided notebooks to reproduce our experiments.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, all experimental details can be found in the Appendix section A.9.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, this description is provided in the Appendix section A.9.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A APPENDIX

A.1 Causal inference

Confounder A confounder is a variable that is associated with both the treatment and the outcome, causing a spurious correlation. For instance, summer is associated with eating ice cream and getting sunburns, but there is no causal relationship between the two.

Propensity score model A propensity score model is a function that predicts treatment from the observed covariates i.e. $P(T = 1|C = c)$ for a binary treatment T and a covariate vector C.

Potential outcome As defined by the Rubin causal model (Rubin, 2005), a potential outcome $Y(t)$ is the value that Y would take if T were set by (hypothetical) intervention to the value t.

Identification assumptions Inference is possible under three identification assumptions.

- **No interference** For a given individual i, this assumption implies that $Y_i(t)$ represents the value that Y would have taken for individual i if T had been set to t for individual i, i.e the potential value of Y_i if T_i had been set to t.
- **Consistency** For a given individual i, $T_i = t \Rightarrow Y_i = Y_i(t)$. This means that for individuals who actually received treatment level t, their observed outcome is the same as what it would have been had they received treatment level t via a hypothetical intervention.
- **Conditional exchangeability** For a given individual i, we assume that conditional on C, the actual treatment level T is independent of each of the potential outcomes:
  $Y(t) \perp T \mid \mathbf{C}, \forall t$

A.2 Evaluation of the tree consistency

We evaluated the consistency of our clustering across subsamples using an Adjusted Rand Index (Rand, 1971). Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of S into r subsets, and $Y = \{Y_1, \ldots, Y_s\}$, a partition of S into s subsets, define the following:

- a, the number of pairs of elements in S that are in the same subset in X and in the same subset in Y.
- b, the number of pairs of elements in S that are in different subsets in X and in different subsets in Y.
- c, the number of pairs of elements in S that are in the same subset in X and in different subsets in Y.
- d, the number of pairs of elements in S that are in different subsets in X and in the same subset in Y.

The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a + b$ can be considered as the number of agreements between X and Y and $c + d$ as the number of disagreements between X and Y. The adjusted Rand index is the corrected-for-chance version of the Rand index (Hubert and Arabie, 1985).

A.3 Proof of convergence

A.3.1 Proof of ASMD Decrease with Multivariate Gaussian X

Lemma (Multivariate Normal Case)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_p)$ *be multivariate normal random variables and* T *a binary treatment indicator. If the following conditions hold:*

1. $|\rho_{ij}| \leq \delta \leq \frac{1-K}{CM^2}$ *for all* $i \neq j$*, where* C *is a constant dependent on the split point*
2. $ASMD_i^{parent} \leq K \cdot ASMD_j^{parent}$ *for all* $i \neq j$*, with* $K < 1$
3. $\frac{\sigma_i}{\sigma_j} \leq M$ *for all* $i \neq j$*, with* $M \geq 1$
4. $Var(X_i|T=1) = Var(X_i|T=0) = \sigma_i^2$ *for all* i

Then, after splitting on the variable X_j *with the highest ASMD, the maximum ASMD in the resulting child nodes will not exceed the original maximum ASMD in the parent node. Put differently, the ASMD imbalance in children nodes either decreases or remains the same.*

Note: *While the proof assumes equal variances for simplicity, similar reasoning can be extended to cases where variances differ between treatment groups.*

Problem Statement

Given:

- A set of continuous random variables $X = (X_1, X_2, \ldots, X_p)$.
- A binary variable T.
- X follows a multivariate Gaussian (normal) distribution.
- The Absolute Standardized Mean Difference (ASMD) for variable X_i is defined as:

$$\text{ASMD}_i = \frac{|\mathbb{E}[X_i \mid T=1] - \mathbb{E}[X_i \mid T=0]|}{\sqrt{\text{Var}(X_i \mid T=1) + \text{Var}(X_i \mid T=0)}}$$

Procedure:

- Identify X_j with the maximum ASMD in the parent node.
- Split the data on X_j at the value x_j that yields the lowest p -value from Fisher's exact test.

Goal: Prove that under certain assumptions, the maximum ASMD over all variables X_i in the child nodes does not exceed the original maximum ASMD in the parent node.

Assumptions

1. **Bounded Correlation Between Variables:** The correlations between X_j and other variables X_i (for $i \neq j$) are bounded:

   $$|\rho_{ij}| \leq \delta, \quad \text{for all } i \neq j$$

   where δ is a small positive constant (e.g., $\delta \leq 0.1$).

2. **Initial ASMDs of other variables:** The initial ASMDs of other variables X_i are lower than that of X_j :

   $$\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}, \text{ for all } i \neq j$$

   where K is a constant less than 1 (e.g., $K \leq 0.5$).

3. **Equal Variances Across Groups:** The variances of X_i conditional on T are equal:

   $$\text{Var}(X_i \mid T=1) = \text{Var}(X_i \mid T=0) = \sigma_i^2$$

4. **Bounded Ratio of Standard Deviations:** There exists a constant $M \geq 1$ such that:

$$\frac{\sigma_i}{\sigma_j} \leq M, \quad \text{for all } i \neq j$$

5. **Linearity and Normality:** Utilizes properties of the multivariate normal distribution, implying linear relationships between variables.

Proof Outline

1. Express ASMDs in terms of parameters.
2. Derive the ASMD of X_i in the child nodes.
3. Bound the change in ASMD of X_i Using δ.
4. Demonstrate that under the given assumptions, the ASMD of X_i in the child nodes does not exceed the original maximum ASMD.

Detailed Proof 1. Express ASMDs in Terms of Parameters

For Variable X_i : In the parent node:

$$\text{ASMD}_i^{\text{parent}} = \frac{|\mu_{i1} - \mu_{i0}|}{\sqrt{2}\sigma_i} = \frac{|\Delta\mu_i|}{\sqrt{2}\sigma_i}$$

where: $\mu_i = \mathbb{E}[X_i \mid T = t]$ and $\Delta\mu_i = \mu_{i1} - \mu_{i0}$

Similarly for variable X_j :

$$\text{ASMD}_j^{\text{parent}} = \frac{|\Delta\mu_j|}{\sqrt{2}\sigma_j}$$

2. Derive the ASMD of X_i in the child nodes

After splitting on X_j at x_j, consider one of the child nodes (e.g., $X_j \leq x_j$). Conditional Mean of X_i for $T = t$:

$$\mathbb{E}[X_i \mid X_j \leq x_j, T = t] = \mu_{it} + \rho_{ij}\frac{\sigma_i}{\sigma_j}\left(\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt}\right)$$

Difference in Conditional Means between $T = 1$ and $T = 0$:

$$\Delta\mu_i^{\text{child}} = \Delta\mu_i + \rho_{ij}\frac{\sigma_i}{\sigma_j}\left(\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0}\right)$$

where: $\Delta\mu_j^{\text{trunc},t} = \mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt}$

3. Bound the change in ASMD of X_i using δ

Absolute difference in adjustments:

$$\left|\Delta\mu_i^{\text{child}} - \Delta\mu_i\right| = \left|\rho_{ij}\frac{\sigma_i}{\sigma_j}\left(\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0}\right)\right|$$

Since $|\rho_{ij}| \leq \delta$ and $\frac{\sigma_i}{\sigma_j} \leq M$,

$$\left|\Delta\mu_i^{\text{child}} - \Delta\mu_i\right| \leq \delta \cdot \frac{\sigma_i}{\sigma_j}\left|\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0}\right| \leq \delta \cdot M\left|\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0}\right|$$

Bounding $\left|\Delta\mu_j^{\text{trunc, }1} - \Delta\mu_j^{\text{trunc, },0}\right|$, we assume a constant C such that:

$$\left|\Delta\mu_j^{\text{trunc, }1} - \Delta\mu_j^{\text{trunc, }0}\right| \leq C\,|\Delta\mu_j|$$

Note that this term is proportional to $|\Delta\mu_j|$ **because X is multivariate Gaussian.** (proof in the following subsection)

Therefore:

$$\left|\Delta\mu_i^{\text{child}} - \Delta\mu_i\right| \leq CM\delta\,|\Delta\mu_j|$$

ASMD of X_i in the child node:

$$\text{ASMD}_i^{\text{child}} = \frac{\left|\Delta\mu_i^{\text{child}}\right|}{\sqrt{2}\sigma_i} \leq \frac{|\Delta\mu_i| + CM\delta\,|\Delta\mu_j|}{\sqrt{2}\sigma_i}$$

or

$$\text{ASMD}_i^{\text{child}} \leq \text{ASMD}_i^{\text{parent}} + \frac{CM\delta\,|\Delta\mu_j|}{\sqrt{2}\sigma_i}$$

4. Demonstrate that $\text{ASMD}_i^{\textbf{child}} \leq \text{ASMD}_j^{\textbf{parent}}$

Express $|\Delta\mu_j|$ in Terms of $\text{ASMD}_j^{\text{parent}}$

$$|\Delta\mu_j| = \sqrt{2}\sigma_j \cdot \text{ASMD}_j^{\text{parent}}$$

Substitute back into the ASMD of X_i :

$$\text{ASMD}_i^{\text{child}} \leq \text{ASMD}_i^{\text{parent}} + CM\delta \cdot \frac{\sqrt{2}\sigma_j \cdot \text{ASMD}_j^{\text{parent}}}{\sqrt{2}\sigma_i} = \text{ASMD}_i^{\text{parent}} + \frac{CM\delta\sigma_j}{\sigma_i} \cdot \text{ASMD}_j^{\text{parent}}$$

Using $\frac{\sigma_i}{\sigma_j} \leq M$:

$$\text{ASMD}_i^{\text{child}} \leq \text{ASMD}_i^{\text{parent}} + CM^2\delta \cdot \text{ASMD}_j^{\text{parent}}$$

To ensure $\text{ASMD}_i^{\text{child}} \leq \text{ASMD}_j^{\text{parent}}$ we require:

$$\text{ASMD}_i^{\text{parent}} + CM^2\delta \cdot \text{ASMD}_j^{\text{parent}} \leq \text{ASMD}_j^{\text{parent}}$$

Subtract $\text{ASMD}_i^{\text{parent}}$ from both sides:

$$CM^2\delta \cdot \text{ASMD}_j^{\text{parent}} \leq \text{ASMD}_j^{\text{parent}} - \text{ASMD}_i^{\text{parent}}$$

Divide both sides by $\text{ASMD}_j^{\text{parent}}$:

$$CM^2\delta \leq 1 - \frac{\text{ASMD}_i^{\text{parent}}}{\text{ASMD}_j^{\text{parent}}}$$

Using the assumption $\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}$:

$$CM^2\delta \leq 1 - K$$

Therefore, to satisfy the inequality:

$$\delta \leq \frac{1-K}{CM^2}$$

Since C is a constant dependent on the truncation and properties of the normal distribution, and $K < 1$, this inequality provides an upper bound on δ.

Under the given assumptions, this recursive process will lead to a decrease in ASMD imbalance as we continue to split.

Note: The proof assumes equal variances for simplicity, but similar reasoning can be extended when variances differ.

A.3.2 Proof of ASMD Decrease with Mixed Data Types X Using Gaussian Copula

Lemma (Mixed Data Types Case using Gaussian Copula)**.** *Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector of mixed continuous and categorical variables, T a binary treatment indicator, and assume a Gaussian copula model with latent variables $\mathbf{Z}$. If the following conditions hold:*

1. *$|\rho_{ij}| \leq \delta \leq \frac{1-K}{M^2 C}$ for all $i \neq j$, where C is a constant dependent on the split point*
2. *$ASMD_i^{parent} \leq K \cdot ASMD_j^{parent}$ for all $i \neq j$, with $K < 1$*
3. *For continuous variables, $\frac{\sigma_i}{\sigma_j} \leq M$ for all $i \neq j$, with $M \geq 1$*
4. *For continuous variables, $Var(X_i|T = 1) = Var(X_i|T = 0) = \sigma_i^2$*
5. *For categorical variables, the thresholds determining categories are the same across treatment groups*

Then, after splitting on the variable X_j with the highest ASMD, the maximum ASMD in the resulting child nodes will not exceed the original maximum ASMD in the parent node. Put differently, the ASMD imbalance in children nodes either decreases or remains the same.

Note: *While the proof assumes equal variances for simplicity, similar reasoning can be extended to cases where variances differ between treatment groups.*

Assumptions

1. **Variables and Treatment:**
   - Let $X = (X_1, X_2, \ldots, X_p)$ be a vector of variables consisting of both continuous and categorical variables.
   - Let T be a binary treatment indicator ($T = 0$ or 1).
2. **Gaussian Copula Model:**
   - There exists a latent variable vector $Z = (Z_1, Z_2, \ldots, Z_p)$ such that:

     $$(Z, T) \sim \mathcal{N}_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

     where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix capturing the dependence structure between variables and the treatment.
   - The observed variables X_i are obtained by applying the inverse of their marginal cumulative distribution functions to the standard normal cumulative distribution of the latent variables:

     $$X_i = F_i^{-1}(\Phi(Z_i)),$$

     where F_i^{-1} is the inverse CDF (quantile function) of X_i, and Φ is the standard normal CDF.

3. **Latent Variable Representation for Categorical Variables:**
   - For each categorical variable X_i, there exists a latent continuous variable X_i^*.
   - The observed categorical variable X_i is determined by thresholding X_i^*:
   $$X_i = k \quad \text{if} \quad X_i^* \in (\tau_{i,k-1}, \tau_{i,k}],$$
   where $\{\tau_{i,k}\}$ are threshold parameters with $-\infty = \tau_{i,0} < \tau_{i,1} < \ldots < \tau_{i,K_i} = \infty$.
4. **Bounded Correlations:**
   - The correlations between the splitting variable X_j and other variables X_i are bounded:
   $$|\rho_{ij}| \leq \delta, \quad \forall i \neq j,$$
   where δ is a small positive constant.
5. **Initial ASMDs:**
   - The initial Absolute Standardized Mean Differences (ASMDs) of other variables X_i are significantly lower than that of X_j:
   $$\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}, \quad \text{with} \quad K < 1.$$
6. **Standard Deviations:**
   - For continuous variables, the standard deviations satisfy:
   $$\frac{\sigma_i}{\sigma_j} \leq M, \quad \forall i \neq j,$$
   where $M \geq 1$ is a constant.
7. **Equality of Variances Across Treatment Groups:**
   - For continuous variables:
   $$\text{Var}(X_i \mid T = 1) = \text{Var}(X_i \mid T = 0) = \sigma_i^2.$$
8. **Thresholds for Categorical Variables:**
   - The thresholds $\{\tau_{i,k}\}$ used to determine the observed categories from latent variables are the same across treatment groups.

—

Proof Step 1: Express ASMDs in Terms of Latent Variables

Continuous Variables:

For a continuous variable X_i, the ASMD in the parent node is:

$$\text{ASMD}_i^{\text{parent}} = \frac{|\mu_{i1} - \mu_{i0}|}{\sqrt{2}\sigma_i},$$

where $\mu_{it} = \mathbb{E}[X_i \mid T = t]$.

Categorical Variables:

For a categorical variable X_i, modeled via a latent variable X_i^*:

- The probability that X_i takes category k in group $T = t$ is:
$$p_{it,k} = \mathbb{P}(X_i = k \mid T = t) = \Phi\left(\frac{\tau_{i,k} - \mu_{it}^*}{\sigma_i^*}\right) - \Phi\left(\frac{\tau_{i,k-1} - \mu_{it}^*}{\sigma_i^*}\right).$$

- The ASMD for category k can be defined as:

$$\text{ASMD}_{i,k}^{\text{parent}} = \frac{2(p_{i1,k} - p_{i0,k})}{\sqrt{p_{i1,k}(1-p_{i1,k}) + p_{i0,k}(1-p_{i0,k})}}.$$

Step 2: Introduction of Constant C in the Gaussian Copula

In the context of the Gaussian copula, the dependence between variables is captured by the covariance matrix $\mathbf{\Sigma}$. When we split on variable X_j, we consider the conditional distribution of other variables given X_j.

For Continuous Variables:

The conditional expectation of X_i given $X_j \leq x_j$ and $T = t$ is:

$$\mathbb{E}[X_i \mid X_j \leq x_j, T = t] = \mu_{it} + \rho_{ij}\frac{\sigma_i}{\sigma_j}\left(\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt}\right).$$

The term $\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt}$ can be expressed using properties of the truncated normal distribution:

$$\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt} = -\sigma_j \lambda_t,$$

where λ_t is defined as:

$$\lambda_t = \frac{\phi(z_t)}{\Phi(z_t)},$$

with $z_t = \frac{x_j - \mu_{jt}}{\sigma_j}$, and ϕ and Φ are the standard normal PDF and CDF, respectively.

Difference Between Treatment Groups:

The difference in these terms between the two treatment groups is:

$$\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0} = -\sigma_j(\lambda_1 - \lambda_0).$$

Using a Taylor expansion around $z = \frac{x_j - \mu_j}{\sigma_j}$, where $\mu_j = \frac{\mu_{j1} + \mu_{j0}}{2}$, we approximate λ_t as:

$$\lambda_t \approx \lambda - \lambda'\left((-1)^{t+1}\frac{\Delta\mu_j}{2\sigma_j}\right),$$

where:

$\lambda = \frac{\phi(z)}{\Phi(z)}$, and $\lambda' = \frac{d}{dz}\left(\frac{\phi(z)}{\Phi(z)}\right) = -\lambda(z + \lambda)$.

Therefore, the difference $\lambda_1 - \lambda_0$ is:

$$\lambda_1 - \lambda_0 \approx \lambda'\left(\frac{\Delta\mu_j}{\sigma_j}\right).$$

Substituting back, we have:

$$|\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0}| = |-\sigma_j(\lambda_1 - \lambda_0)| = |\lambda'||\Delta\mu_j|.$$

Define the constant C as:

$$C = |\lambda'| = \lambda|z + \lambda|.$$

This constant C depends on the truncation point x_j and captures the effect of truncation on the mean differences.

Result:

Under the given assumptions, and per the previous result in the continuous (multivariate Gaussian) case we have:

$$\delta \leq \frac{1-K}{M^2 C},$$

we have that the ASMDs of both continuous and categorical variables in the child nodes after splitting do not exceed $\text{ASMD}_j^{\text{parent}}$:

$$\text{ASMD}_i^{\text{child}} \leq \text{ASMD}_j^{\text{parent}}, \quad \forall i \neq j.$$

Calculation of C in the Multivariate Gaussian and Gaussian Copula Cases

Multivariate Gaussian Case Derivation of C:

Consider a continuous variable X_j that follows a normal distribution within each treatment group $T = t$:

$$X_j \mid T = t \sim \mathcal{N}(\mu_{jt}, \sigma_j^2).$$

When we split on X_j at a threshold x_j, the truncated means for each treatment group are:

$$\mu_{jt}^{\text{trunc}} = \mathbb{E}[X_j \mid X_j \leq x_j, T = t] = \mu_{jt} - \sigma_j \lambda_t,$$

where λ_t is the inverse Mills ratio:

$$\lambda_t = \frac{\phi(z_t)}{\Phi(z_t)},$$

with $z_t = \dfrac{x_j - \mu_{jt}}{\sigma_j}$, $\phi(\cdot)$ is the standard normal PDF, and $\Phi(\cdot)$ is the standard normal CDF.

The difference in truncated means between treatment groups is:

$$\Delta\mu_j^{\text{trunc}} = \mu_{j1}^{\text{trunc}} - \mu_{j0}^{\text{trunc}} = (\mu_{j1} - \sigma_j\lambda_1) - (\mu_{j0} - \sigma_j\lambda_0) = \Delta\mu_j - \sigma_j(\lambda_1 - \lambda_0),$$

where $\Delta\mu_j = \mu_{j1} - \mu_{j0}$.

To approximate λ_t, we perform a first-order Taylor expansion around $\bar{\mu}_j = \dfrac{\mu_{j1} + \mu_{j0}}{2}$:

$$\lambda_t \approx \lambda - \lambda' \left((-1)^{t+1} \frac{\Delta\mu_j}{2\sigma_j} \right),$$

where:

- $z = \dfrac{x_j - \bar{\mu}_j}{\sigma_j}$,
- $\lambda = \dfrac{\phi(z)}{\Phi(z)}$,
- $\lambda' = \dfrac{d}{dz}\left(\dfrac{\phi(z)}{\Phi(z)}\right) = -\lambda(z + \lambda$

Then, the difference $\lambda_1 - \lambda_0$ is:

$$\lambda_1 - \lambda_0 \approx \lambda' \left(\frac{\Delta\mu_j}{\sigma_j} \right).$$

Substituting back:

$$\Delta\mu_j^{\text{trunc}} = \Delta\mu_j - \sigma_j \left(\lambda' \frac{\Delta\mu_j}{\sigma_j} \right) = \Delta\mu_j - \lambda'\Delta\mu_j = (1 - \lambda')\Delta\mu_j.$$

The magnitude of the change in the truncated mean difference is:

$$\left| \Delta\mu_j^{\text{trunc}} - \Delta\mu_j \right| = \left| -\lambda'\Delta\mu_j \right| = \left| \lambda' \right| \left| \Delta\mu_j \right|.$$

Define the Constant C:

$$C = |\lambda'| = \lambda|z + \lambda|.$$

Gaussian Copula Case Derivation of C:

In the Gaussian copula framework, the dependence structure between variables is modeled using a multivariate normal distribution of latent variables Z:

$$Z = (Z_1, Z_2, \ldots, Z_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is the correlation matrix. The observed variables X_j are obtained by transforming the latent variables using their marginal distributions:

$$X_j = F_j^{-1}\left(\Phi(Z_j)\right),$$

where F_j^{-1} is the inverse CDF of X_j.

When splitting on X_j, we consider the conditional distribution of Z_j given $Z_j \leq z_j$. The truncated mean of Z_j is:

$$\mu_{Z_j}^{\text{trunc}} = \mathbb{E}[Z_j \mid Z_j \leq z_j] = -\lambda,$$

where $\lambda = \dfrac{\phi(z_j)}{\Phi(z_j)}$

The difference in truncated means between treatment groups (assuming different thresholds for Z_j) is similarly approximated. Using Taylor expansion as in the multivariate Gaussian case, we arrive at:

$$\left| \Delta\mu_{Z_j}^{\text{trunc}} - \Delta\mu_{Z_j} \right| = \left| \lambda' \right| \left| \Delta\mu_{Z_j} \right|,$$

where $\Delta\mu_{Z_j}$ is the difference in means of Z_j between treatment groups, and:

$$C = |\lambda'| = \lambda|z_j + \lambda|.$$

Since the observed X_j is a function of Z_j through F_j^{-1}, the change in $\Delta\mu_j^{\text{trunc}}$ can be related back to $\Delta\mu_{Z_j}^{\text{trunc}}$, and C serves the same role in bounding the change.

A.4 Further related work

Causal inference provides a wide range of methods for estimating causal effect from data with unbalanced treatment allocation. In balancing methods such as matching or weighting methods, the data is pre-processed to create subgroups with lower treatment imbalance or "natural experiments".

A.4.1 Effect estimation

Matching *Matching* methods consist of clustering similar units from the treatment and control groups to reduce imbalance. In general, a matching procedure generates weights w_{ij} denoting the assignment of one or many control units j to a treated unit i ((Morgan and Winship, 2015), Chapter 5). Exact matching only assigns control units to treatment units with the exact same set of covariate values. But typically, matched control units j are chosen based on a nearest-neighbors search according to some distance metric. However, matching procedures induce a bias-variance trade-off as discarding unmatched samples reduces estimation error at the cost of increased variance. However, all matching methods suffer from the curse of dimensionality, making them impractical in high-dimensional datasets. Exact matching and coarsened exact matching (Iacus et al., 2012b) find exponentially fewer matches as the input dimension grows (Abadie and Imbens, 2006). Alternative methods include Propensity score matching (Austin, 2011), where distance is computed from an estimate of the propensity score $P(T = 1 \mid X = \boldsymbol{x})$, and Mahalanobis distance matching (Rubin, 1980) (see more details below in 4.1). However, compression into a single dimension can lead to highly unrelated matches with very different characteristics in the original covariate space, and can ultimately increase estimation bias (King and Nielsen, 2019). *Clivio et. al* overcome this issue by developing a multivariate balancing score to perform matching for high-dimensional causal inference. Nevertheless, this approach is not interpretable.

Weighting methods An alternative to matching are *weighting* methods, where sample weights are estimated, generalizing the problem formulation of matching. In Inverse Probability Weighting (IPW) (Horvitz and Thompson, 1952) which is the most popular alternative in that category, the samples are weighted according to their *propensity* score, i.e. the estimated probability of treatment conditional on their covariates.

Outcome models *Outcome* models estimate the causal effect from regression models where both treatment and covariates act as predictors of the outcome. These regressions can be fitted through various methods like linear regression (Imbens and Rubin, 2015), neural networks (Shi et al., 2019; Shalit et al., 2017a), or tree-based models (Athey and Imbens, 2016). Common alternatives include Doubly Robust estimators (Seaman and Vansteelandt, 2018), Double Debiased Machine Learning (Chernozhukov et al., 2018) and metalearners such as the T-learner and X-learner (Künzel et al., 2019). These methods have the advantage of being very data-adaptive. More particularly, the Causal Tree approach (Athey and Imbens, 2016) builds on regression tree methods, and splits the data to optimize for goodness of fit in treatment effects. Causal Tree separates the training dataset into two subsamples: a splitting subsample and an estimating subsample. The splitting subsample is used to build a causal tree while the estimating subsample is used to generate unbiased conditional treatment effect estimates. This procedure is called "honest estimation" and is anticipated to avoid overfitting. However, most outcome models are not interpretable.

Representation learning Representation learning has emerged as a key approach for causal effect estimation, especially in observational settings where confounding is a significant challenge. By learning low-dimensional representations of covariates, these methods aim to reduce bias and improve the estimation of treatment effects. For instance, neural networks have been used to capture complex relationships between variables, as shown in (Shalit et al., 2017b). Optimal transport methods have further enhanced these estimates by ensuring balance between treatment groups (Xu et al., 2020). Additionally, decomposed representations, which separate confounding information from treatment-specific effects, allow for more accurate estimates (Louizos et al., 2017), while disentangled representations help isolate factors affecting potential outcomes (Yoon et al., 2018). However, by definition using representation learning prevents any interpretability of the effect estimation procedure.

A.4.2 Positivity violations

Causal inference is only possible under the *positivity* assumption, which requires covariate distributions to overlap between treatment arms. Thus, positivity violations (also referred to as no overlap) occur when certain subgroups

in a sample do not receive one of the treatments of interest or receive it too rarely (Karavani et al., 2019). Overlap is essential as it guarantees data-driven outcome extrapolation across treatment groups. Having no common support means there are subjects in one group with no counterparts from the other group, and, therefore, no reliable way to pool information on their outcome had they been in the other group. Non-violating samples are thus the only ones for which we can guarantee some validity of the inferred causal effect.

There are three common ways to characterize positivity. The most common one consists in estimating propensity scores and excluding the samples associated with extreme values (known as "trimming") (Petersen et al., 2012). The threshold for propensity scores can be set arbitrarily or dynamically (Crump et al., 2009). However, since samples are excluded on the basis of their propensity scores and not their covariate values, these methods lack interpretability about the excluded subjects and how it may affect the target population on which we can generalize the inference. Thus, other methods have been developed to overcome this challenge by characterizing the propensity-based exclusion (Oberst et al., 2020; Wolf et al., 2021; Ackerman et al., 2020). Lastly, the third way tries to characterize the overlap from covariates and treatment assignment directly, without going through the intermediate propensity score e.g. PositiviTree (Karavani et al., 2019). In PositiviTree, a decision tree classifier is fitted to predict treatment allocation. In contrast to their approach, BICause Tree implements a tailor-made optimization function where splits are chosen to maximize balancing in the resulting sub-population, whereas PositiviTree uses off-the-shelf decision trees maximizing separation. Ultimately, the above-mentioned methods for positivity identification and characterization are model agnostic. In our model, BICauseTree, positivity identification *and characterization* are inherently integrated into the model, and effect estimation comes with a built-in interpretable abstention prediction mechanism.

A.4.3 Further comparison with Causal Tree

Causal Trees (CT) (Athey and Imbens, 2016) are another tree-based model for causal inference that (i) leverages the inherent interpretability of decision trees, and (ii) has a custom objective function for recursively splitting the data. Although both utilize decision trees, BICauseTree and CT serve distinct purposes. BICauseTree splits are optimized for balancing treatment *allocation* while CT splits are optimized for balancing treatment *effect*, under assumed exchangeability. In other words, CT *assumes* exchangeability while BICauseTree *finds* exchangeability. As such, our approach is only suited for ATE estimation while CT is better suited for Conditional Average Treatment Effect estimation (Athey and Imbens, 2016). Furthermore, in practice, causal effects are often averaged over multiple trees into a Causal *Forest* (Wager and Athey, 2018; Athey et al., 2019) that is no longer interpretable, and users are encouraged to use post-hoc explanation methods (Battocchi et al., 2019).

A.5 Experimental details: synthetic datasets

A.5.1 Natural experiment dataset

For the natural experiment dataset, we consider a Death outcome D, a binary treatment of interest T and two covariates such that $X = (S, A)$ with S the sex and A the continuous age such that:

$$S \sim \text{Bernoulli}(0.5)$$
$$A \sim \text{Normal}\left(50, 20^2\right)$$

The sample size was chosen to be $N = 20,000$.

We defined four sub-populations, each constituting a natural experiment, with a different propensity distribution $P(T = 1|X = x)$:

$$\Pr(T = 1|S = 1, A \geq 50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.5, 0.1^2\right)$$
$$\Pr(T = 1|S = 1, A < 50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.3, 0.1^2\right)$$
$$\Pr(T = 1|S = 0, A \geq 50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.1, 0.1^2\right)$$
$$\Pr(T = 1|S = 0, A < 50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.4, 0.1^2\right)$$

Individual treatment propensities were sampled from the corresponding distributions above and observed treatment values were sampled from a Bernoulli distribution parameterized with the individual propensities. The outcome

probabilities were not modeled as a distribution. Instead, observed outcome values were sampled directly from a Bernoulli distribution parameterized with a constant value that depended on both X and T.

- $\Pr(Y|T=1,S=1,A\geq 50) \sim \mathcal{B}(0.1)$
- $\Pr(Y|T=1,S=1,A< 50) \sim \mathcal{B}(0.2)$
- $\Pr(Y|T=1,S=0,A\geq 50) \sim \mathcal{B}(0.4)$
- $\Pr(Y|T=1,S=0,A< 50) \sim \mathcal{B}(0.15)$
- $\Pr(Y|T=0,S=1,A\geq 50) \sim \mathcal{B}(0.2)$
- $\Pr(Y|T=0,S=1,A< 50) \sim \mathcal{B}(0.4)$
- $\Pr(Y|T=0,S=0,A\geq 50) \sim \mathcal{B}(0.8)$
- $\Pr(Y|T=0,S=0,A< 50) \sim \mathcal{B}(0.3)$

No positivity violation was modeled in this experiment.

A.5.2 Positivity violations dataset

For this second synthetic experiment, we build a dataset with two positivity-violating subgroups. Let us consider a synthetic example of a dataset with a Death outcome D, a binary treatment of interest T, and three binary covariates –sex, cancer, and arrhythmia– such that $X = (S, C, A)$. For the marginal distributions, we set:

$$S \sim \text{Ber}(0.5)$$

$$C \sim \text{Ber}(0.3)$$

$$A \sim \text{Ber}(0.1)$$

The sample size was chosen to be $N = 20,000$.

- $\Pr(T=1|S=1,C=1,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(1.00, 0.02^2\right)$
- $\Pr(T=1|S=1,C=0,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(0.32, 0.10^2\right)$
- $\Pr(T=1|S=1,C=1,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.12, 0.10^2\right)$
- $\Pr(T=1|S=1,C=0,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.42, 0.10^2\right)$
- $\Pr(T=1|S=0,C=1,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.17, 0.10^2\right)$
- $\Pr(T=1|S=0,C=1,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(0.30, 0.10^2\right)$
- $\Pr(T=1|S=0,C=0,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(0.24, 0.10^2\right)$
- $\Pr(T=1|S=0,C=0,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.00, 0.02\right)$

The observed outcome values were sampled from a Bernoulli distribution parameterized with a constant value that depended on both X and T. For the treated:

- $\Pr(Y|T=1,S=1,C=1,A=1) \sim \mathcal{B}(0.13)$
- $\Pr(Y|T=1,S=1,C=1,A=0) \sim \mathcal{B}(0.08)$
- $\Pr(Y|T=1,S=1,C=0,A=1) \sim \mathcal{B}(0.21)$
- $\Pr(Y|T=1,S=1,C=0,A=0) \sim \mathcal{B}(0.1)$
- $\Pr(Y|T=1,S=0,C=1,A=1) \sim \mathcal{B}(0.36)$
- $\Pr(Y|T=1,S=0,C=1,A=0) \sim \mathcal{B}(0.29)$
- $\Pr(Y|T=1,S=0,C=0,A=1) \sim \mathcal{B}(0.24)$
- $\Pr(Y|T=1,S=0,C=0,A=0) \sim \mathcal{B}(0.09)$

For the untreated:

- $\Pr(Y|T=0,S=1,C=1,A=1) \sim \mathcal{B}(0.31)$
- $\Pr(Y|T=0,S=1,C=1,A=0) \sim \mathcal{B}(0.4)$
- $\Pr(Y|T=0,S=1,C=0,A=1) \sim \mathcal{B}(0.29)$
- $\Pr(Y|T=0,S=1,C=0,A=0) \sim \mathcal{B}(0.45)$
- $\Pr(Y|T=0,S=0,C=1,A=1) \sim \mathcal{B}(0.4)$

- $\Pr(Y|T=0,S=0,C=1,A=0) \sim \mathcal{B}(0.51)$
- $\Pr(Y|T=0,S=0,C=0,A=1) \sim \mathcal{B}(0.43)$
- $\Pr(Y|T=0,S=0,C=0,A=0) \sim \mathcal{B}(0.73)$

A.6 Further experiment results: synthetic experiments

A.6.1 Natural experiment dataset

Adding Coarsened Exact Matching Coarsened Exact Matching (CEM) Iacus et al. (2012a) is a matching method for causal inference that temporarily coarsens covariates, matches treated and control units exactly on these coarsened values, and then retains the original (uncoarsened) values for analysis, reducing model dependence and improving balance. On the natural experiment, CEM exhibits performance comparable to Mahalanobis Matching (MM).

Table A1: Mean estimation bias and estimation variance for the natural experiment dataset. MM: Mahalanobis Matching, IPW: Inverse Propensity Score Weighting, CEM: Coarsened Exact Matching

Method	Sample	Mean Bias	Variance
BICause Tree	Test	0.012	0.004
BICause Tree	Train	0.009	0.004
IPW (LR)	Test	0.060	0.012
IPW (LR)	Train	0.058	0.010
IPW (GBT)	Test	0.008	0.006
IPW (GBT)	Train	0.007	0.005
Marginal	Test	0.072	0.011
Marginal	Train	0.067	0.010
MM	Test	0.012	0.018
MM	Train	0.007	0.012
CEM	Test	0.013	0.018
CEM	Train	0.011	0.012

The importance of selecting candidate features by max ASMD The heuristic for choosing the feature with the maximum absolute standardized mean difference (ASMD) among all features to split on can be empirically justified. We can do that by comparing two BICauseTrees: one that selects the feature with maximal ASMD at each recursive iteration (the model used throughout this work), to a second tree whose only difference is choosing candidate features randomly. Figure A1 demonstrates how this design decision contributes to a reduction in estimation bias.

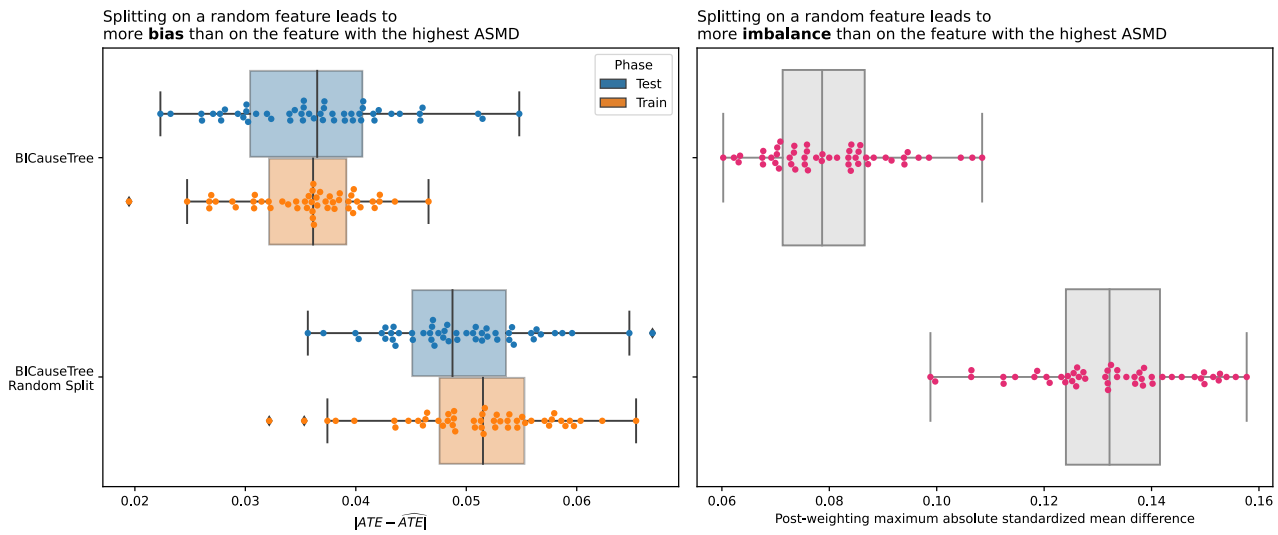

Figure A1: Comparison of estimation bias (left, absolute difference) and imbalance (right, maximum absolute standardized mean difference [ASMD]) using two flavors of BICauseTree: First, the original version from the manuscript, wherein each recursion the feature to split on is chosen by the one with maximal ASMD. And second, a variant where features are selected at random, with all other hyperparameters held constant. We see that selecting the most imbalanced feature for stratification leads to better balancing and, more importantly, better estimation - justifying the rationale for selecting features based on the highest ASMD.

Additionally, adjusting for the maximal ASMD can be justified in simplistic scenarios. Consider the following data generating process:

$$
\begin{aligned}
Y &= A + X_1 + X_2 \\
A &\sim Bernoulli(0.5) \\
X_1, X_2 &\sim Normal(0, 1) \\
X_2|_{A=1} &\sim Normal(1, 1)
\end{aligned}
$$

Where X_1 is balanced (independent of A), but X_2 is not since the mean of the treated samples is shifted one unit upward.

We can then consider four simple linear models:
1) Null-model: $Y \sim 1 + A$.
2) Full-model: $Y \sim 1 + A + X_1 + X_2$.
3) X_1-model: $Y \sim 1 + A + X_1$.
4) X_2-model: $Y \sim 1 + A + X_2$.
And observe that the X_2-model recovers the true effect much better than the X_1-model (see Figure A2).

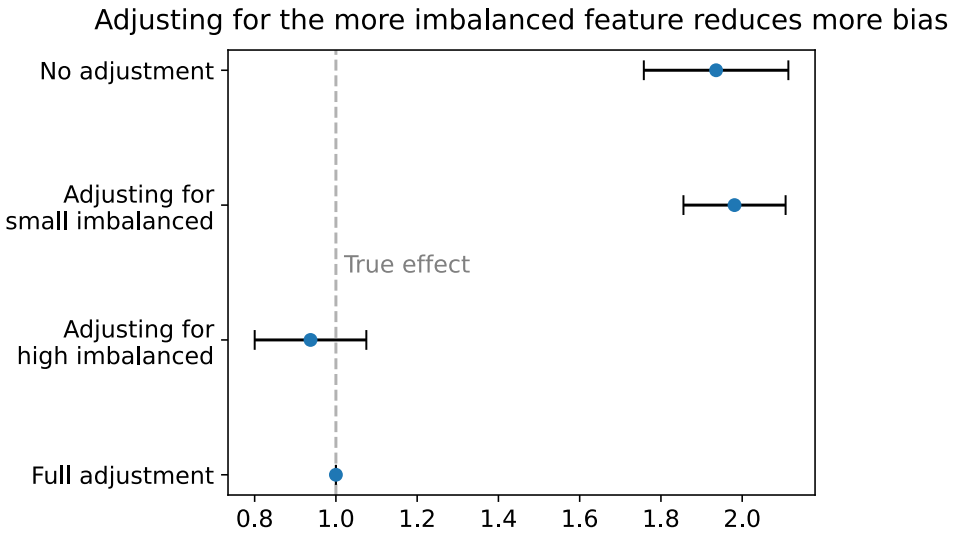

Figure A2: Adjusting for the more imbalanced predictor reduces more estimation bias of the treatment effect. The data contains a treatment assignment and two predictor variables – one balanced across treatment groups and one imbalanced. We examine four models, all adjusting for the treatment assignment but differing in their adjustment of the predictors. Adjusting for the more imbalanced predictor almost retrieves the treatment effect perfectly (dashed grey line), while adjusting the less imbalanced predictor still suffers from large residual confounding.

Note that this implicitly assumes all predictors have an equal association with the outcome. Therefore, as suggested in the main text, future work can create a combine score of ASMD and predictor-outcome association. One such approach can be as follows.

1. We calculate the ASMDs, denoted s.

2. We then calculate any variable importance measure, preferably fitted jointly using all predictors and treatment assignment, as long as its values ranges in the positives. Denote it as q.
   Examples can be the absolute coefficients for regression models or permutation importance for tree-based models.

3. We then normalize both s and q, separately so each sum to 1, depicting relative imbalance and relative predictor-outcome importance among predictors.

4. Finally, we use the combined measurement $s \cdot q$ to select the predictor to split on in each iteration.

Examining $s \cdot q$ it is easy to see that if a variable j is balanced (near-zero s_j) or unrelated to the outcome (near-zero q_j), the combined score will be small and less likely to be chosen for splitting. The predictor j itself is also less likely to cause bias because it is either balanced, or, if imbalanced, affects the outcome very little.

Applicability in high-dimensional data Being based on decision tree logic allows BICauseTree to scale properly for high-dimensional data. BICauseTree, like decision trees, scans all the features before deciding which to split on. This allows it to focus on the most biasing feature at each partition, regardless of their number, enabling them to find natural experiments in noisy high-dimensional data. To demonstrate that, we augmented our natural experiment dataset with an increasing number of noisy features. Figure A3 shows that the error of BICauseTree's ATE estimation stays consistent regardless of the number of added features.

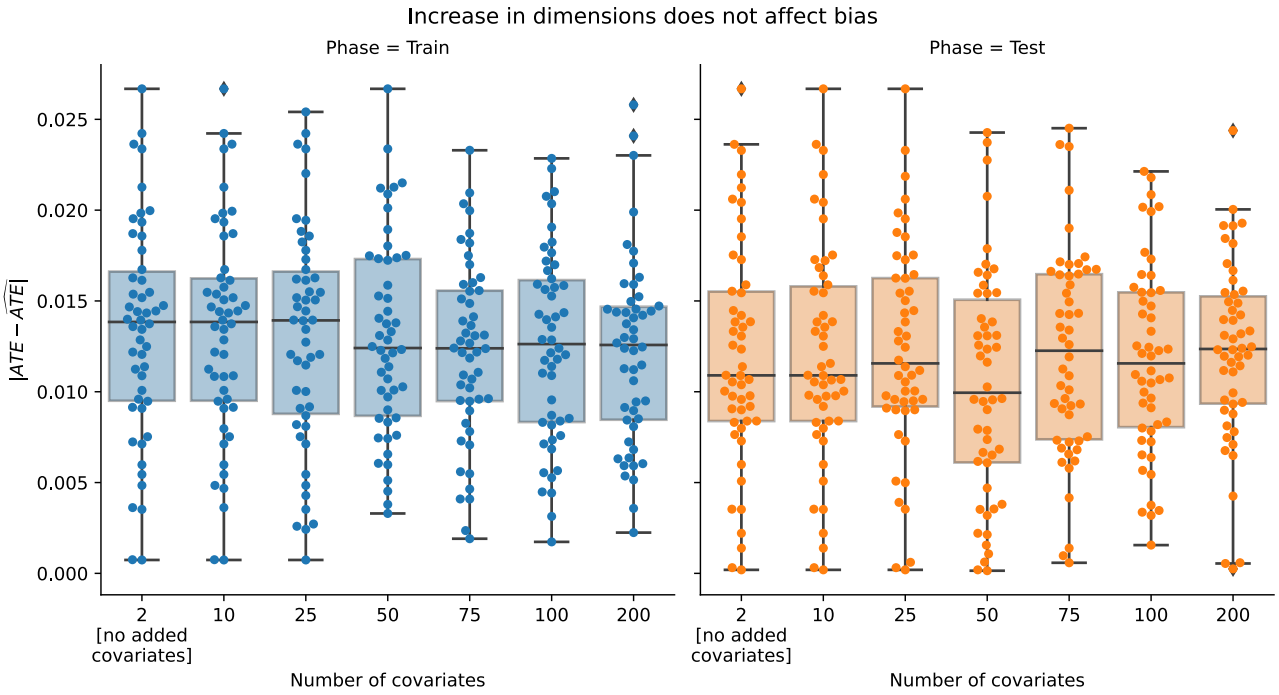

Figure A3: Comparison of estimation bias as a function of number of covariates. We take a synthetic dataset that adheres to conditional exchangeability and increasingly add noisy covariates to it. We see the estimation bias remains the same, throughout the train set (left) and the test set (right) regardless of the number of covariates added. This better strengthens our claim for the suitability of our method to discover unbiased natural experiments in high-dimensional data.

Partitioning The tree partitions so to recreate the four intended sub-populations where a natural experiment was simulated. The average Rand index was equal to 0.901 across the 50 subsamples ($\sigma = 0.263$). Violating leaf nodes are marked in red.

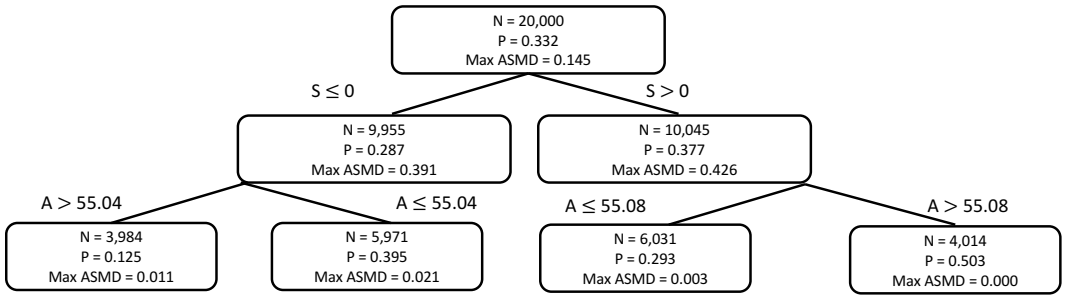

Figure A4: Tree structure after training on the entire natural experiment dataset ($N = 20,000$). Violating leaf nodes are marked in red. P is for node prevalence.

Causal effect estimation Here, Causal Tree has both higher estimation bias and higher estimation variance compared to other methods. The tree-like nature of our data structure may be incompatible with the optimization function of Causal Tree, which maximizes treatment effect heterogeneity. Causal Forest, which has multiple Causal Tree models, however, overcomes this difficulty and performs well.

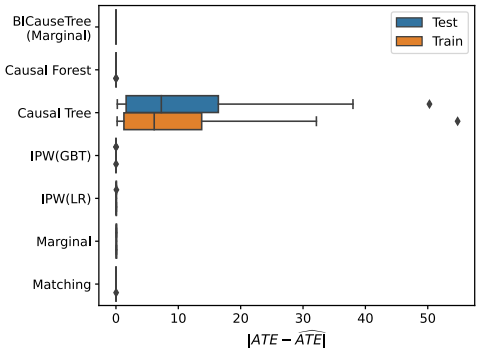

Figure A5: Estimation bias for the natural experiment dataset across 50 subsamples with $N = 20,000$.

Propensity score estimation BICauseTree(Marginal)'s propensity estimation is well calibrated. It is closer to the identity line than logistic regression- (IPW(LR)) and the gradient boosting trees- (IPW(GBT)) based models. The calibration however remains satisfactory across all models. This further shows our ability to identify natural experiments in a simple data setting.

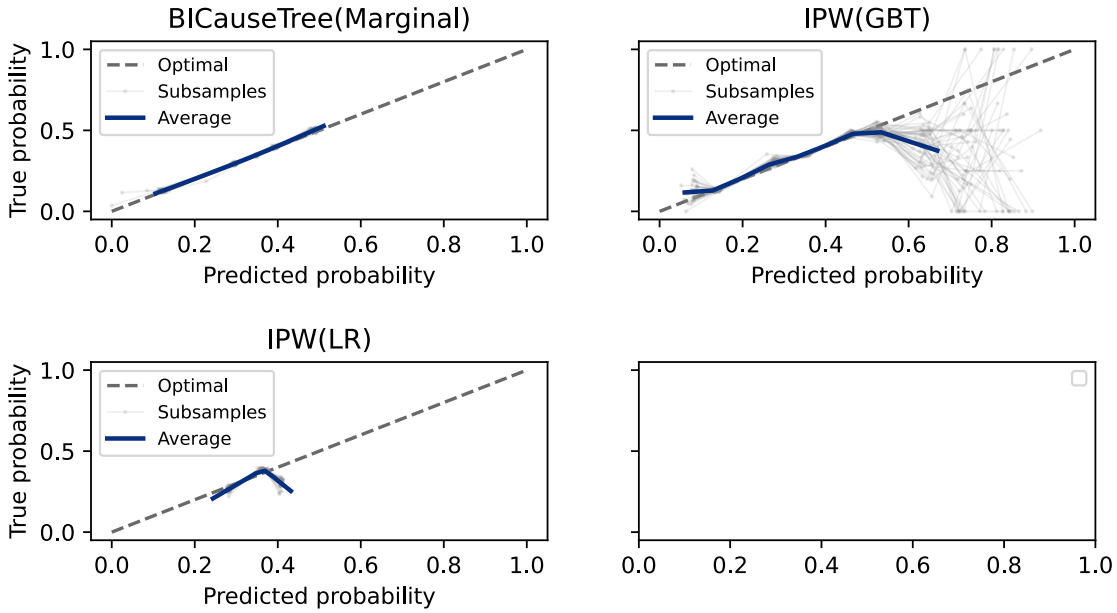

Figure A6: Calibration of the propensity score estimation for the natural experiment dataset across 50 subsamples, on the natural experiment testing set ($N = 10,000$).

Outcome estimation The performance of BICauseTree w.r.t outcome estimation is only satisfactory in this experiment. This shows our ability to estimate the ATE despite a somehow lower calibration of our outcome estimation. Ultimately, the performance would likely have been improved had we used a more complex outcome model. The calibration of the predicted outcomes by BICauseTree however remains better than the one by Matching.

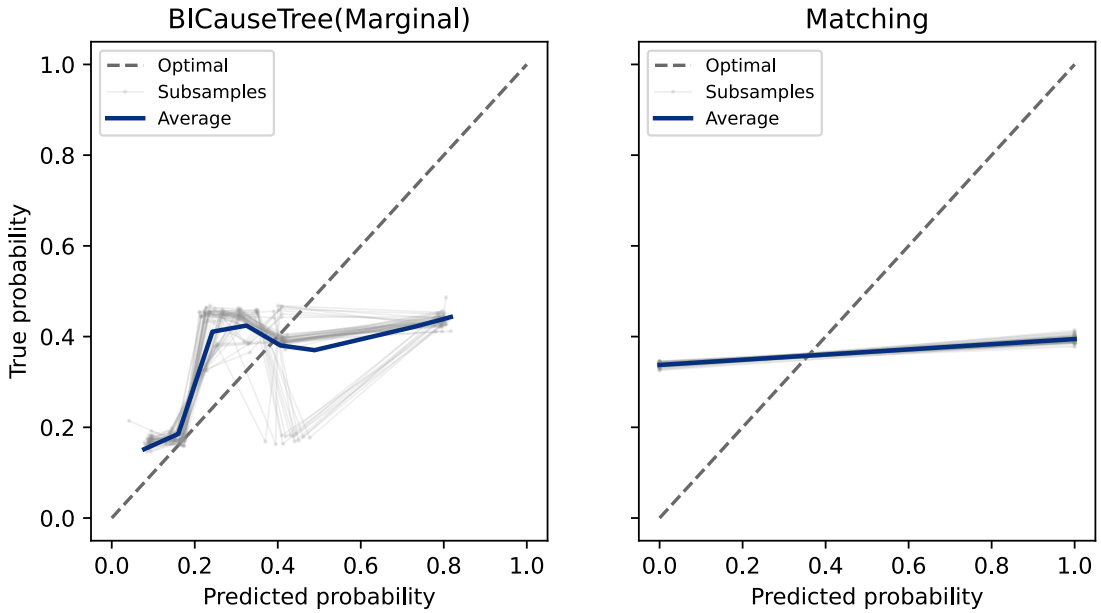

Figure A7: Calibration of the outcome estimation for the natural experiment dataset across 50 subsamples on the natural experiment testing set ($N = 10,000$).

Effect estimation bias reduction with tree depth Figure A8 shows how estimation bias decreases as we increase the maximum depth hyperparameter of our BICauseTree(Marginal) on the natural experiment training dataset ($N = 10,000$). Here, each circle in the plot represents the estimation bias for a subsample. The dotted line shows the average bias with an IPW estimator. The shaded area represents the 95% confidence interval (CI) for IPW. This result shows that our estimation is robust to the choice of a maximum tree depth hyperparameter. Above a certain threshold, the effect estimate from BICauseTree remains unbiased.

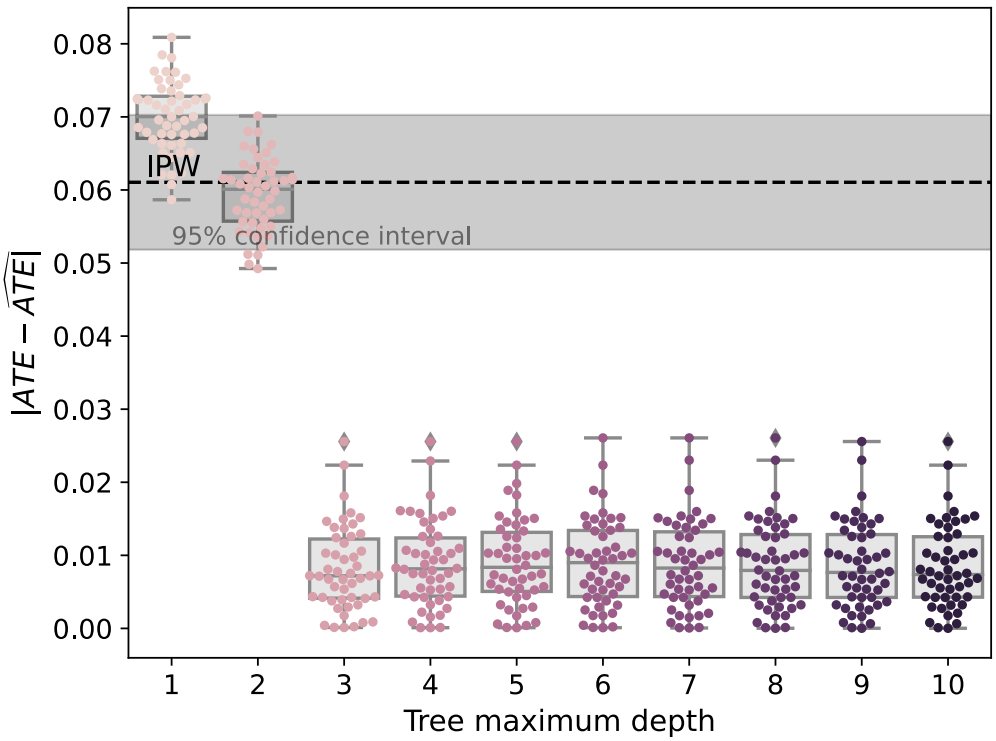

Figure A8: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the natural experiment training set ($N = 10,000$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure A9 below shows the weighted ASMD of both covariates S and A in BICauseTree models with varying maximum tree depth hyperparameters. It illustrates the reduction of treatment allocation bias as the tree grows, but also shows that this reduction is a heuristic as the ASMD might not decrease monotonically.

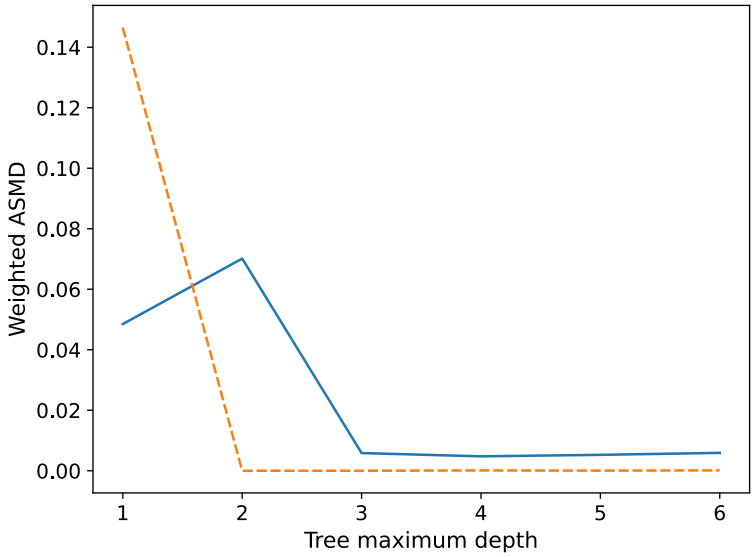

Figure A9: Weighted ASMD for all covariates applying *BICauseTree(Marginal)* models with varying maximum tree depths on the natural experiment training set ($N = 10,000$) across 50 subsamples.

A.6.2 Positivity violations dataset

Partitioning BICauseTree coherently flags the two subpopulations where positivity violations were simulated, as shown in the example partition on the entire dataset below in Figure A10 (violating leaves are marked in red). The average Rand index was equal to 0.986 across the 50 subsamples ($\sigma = 0.026$).

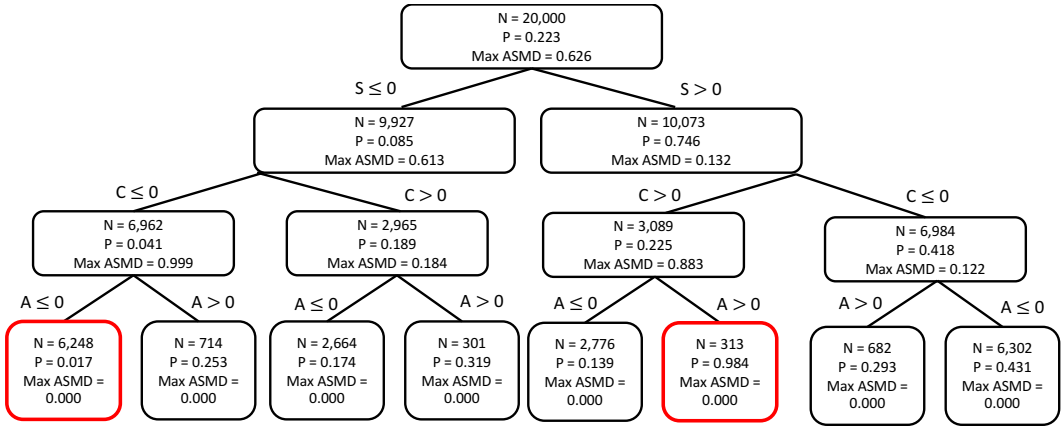

Figure A10: Tree structure after training on the entire positivity violations dataset ($N = 20,000$). Violating leaf nodes are marked in red. P is for node prevalence.

Causal effect estimation Our tree compares with all existing alternatives. Matching has both higher estimation bias and higher estimation variance compared to other methods.

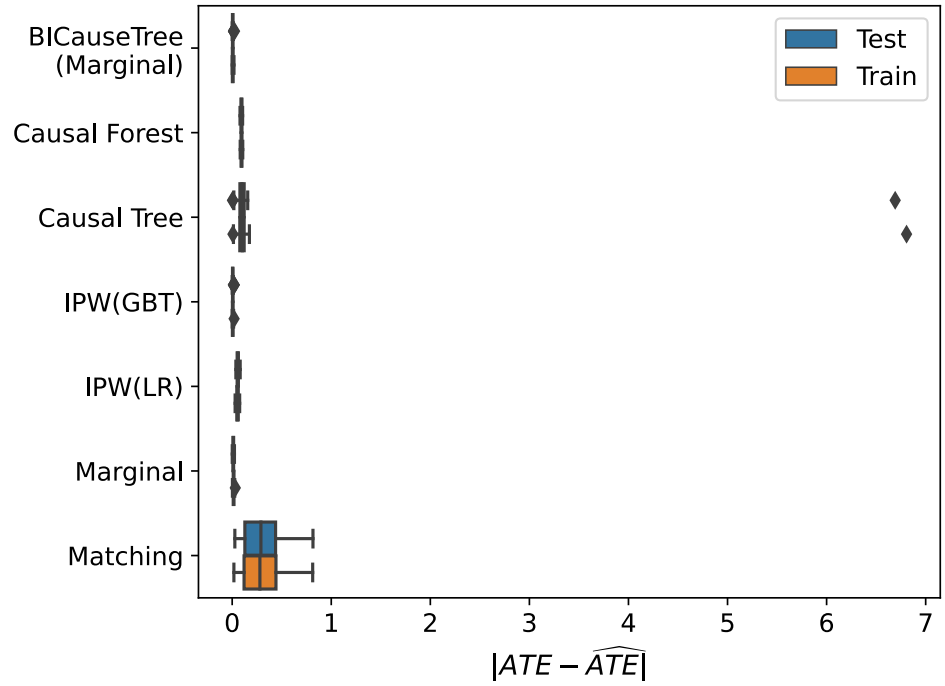

Figure A11: Estimation bias on the positivity violations dataset ($N = 20,000$) across 50 subsamples

Propensity score estimation BICauseTree(Marginal) shows good calibration since it indeed clearly identified the correct subpopulations. The gradient-boosting tree (IPW(GBT)) also shows good calibration, as the tree-based modeling captures the underlying structure of the data. On the other hand, the logistic regression-based model (IPW(LR)) is ill-specified for the data and therefore presents relatively poorer calibration.

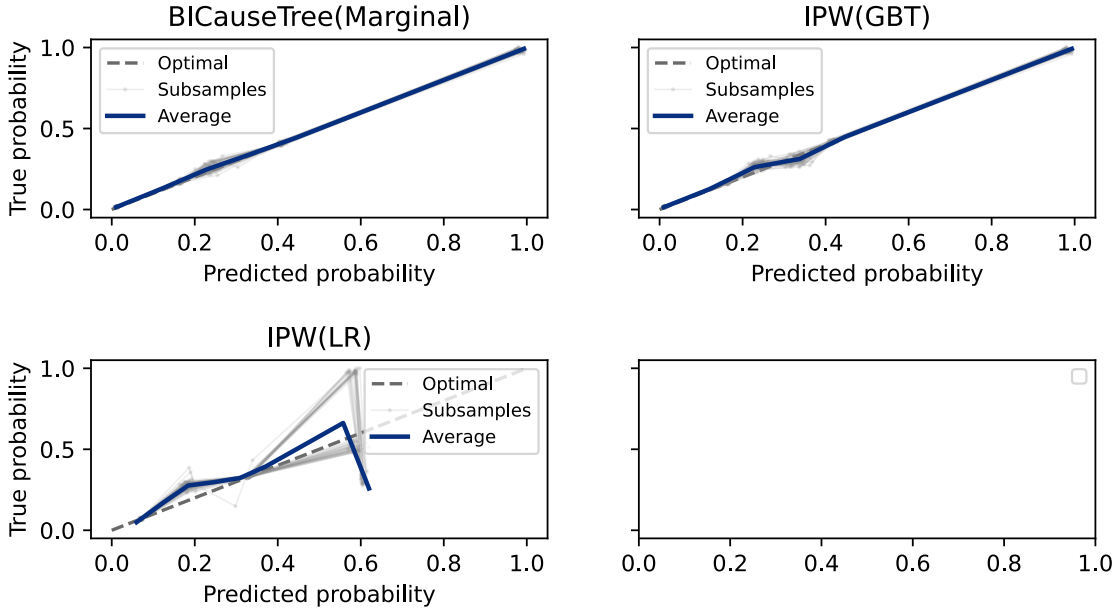

Figure A12: Propensity score calibration comparing our approach to existing alternatives on the testing set of the positivity violation dataset ($N = 10,000$) across 50 subsamples

Outcome estimation BICauseTree shows descent calibration on outcome prediction in this experiment. Matching shows poor calibration.

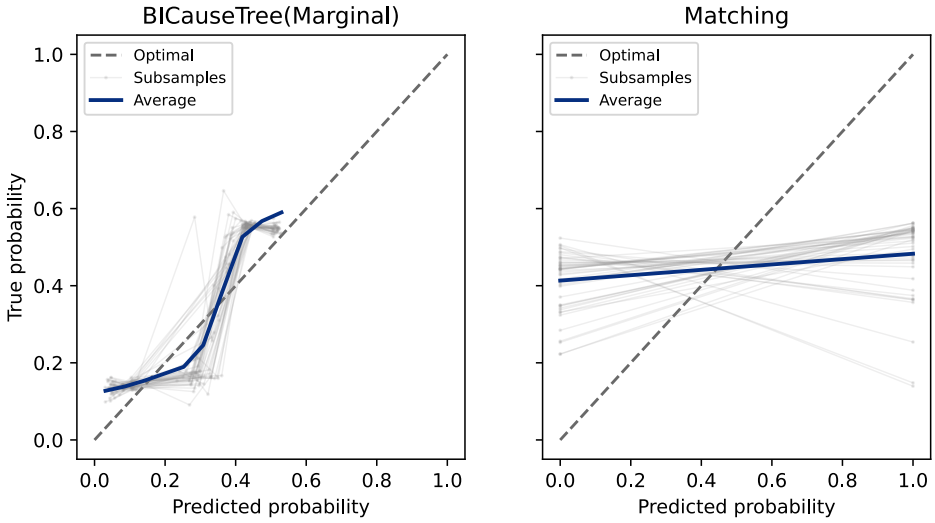

Figure A13: Calibration of the outcome estimation comparing our approach to existing alternatives on the positivity violations testing dataset ($N = 10,000$) across 50 subsamples

Effect estimation bias reduction with tree depth We see in this case, that having a max depth that is too large leads to an increase in the bias. This might indicate that a hyperparameter tuning procedure would benefit cases where the tree might become too deep and overfit to the training data.

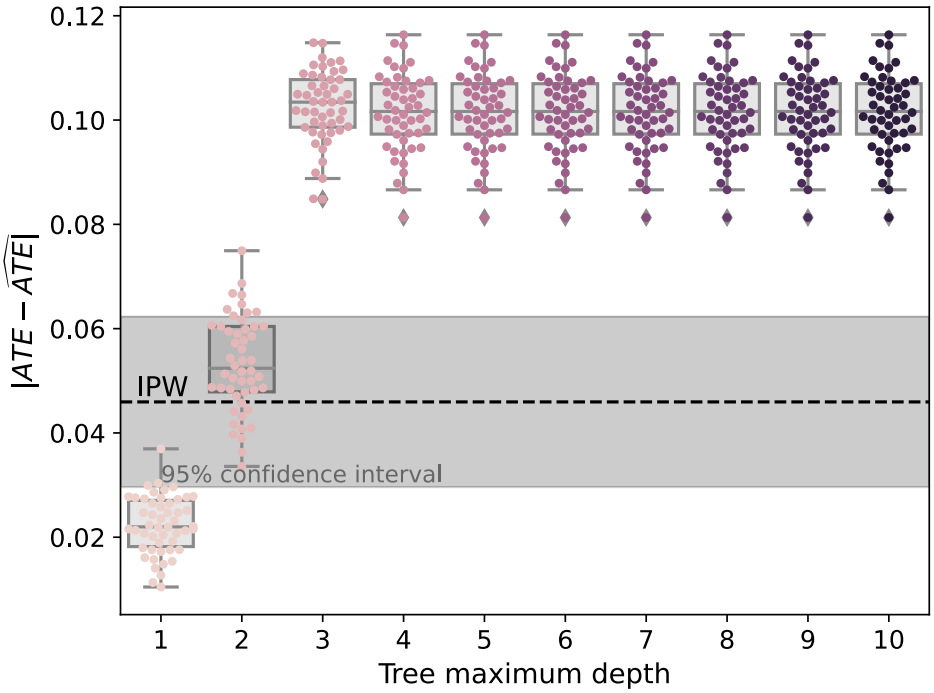

Figure A14: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the positivity violations experiment training set ($N = 10,000$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure A15 below shows the weighted ASMD of all three covariates in BICauseTree models with varying maximum tree depth hyperparameters. We see that the ASMD generally decreases with tree depth, showing that the splitting criteria we use leads to balanced subpopulations, as expected.

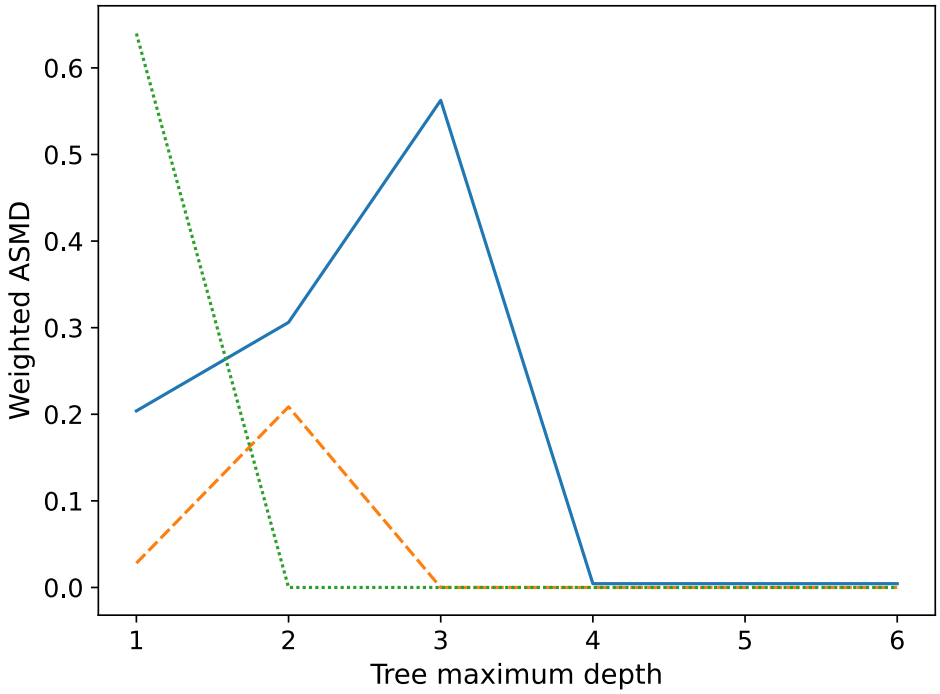

Figure A15: Weighted ASMD for all covariates applying *BICauseTree(Marginal)* models with varying maximum tree depths on the positivity violations dataset training set ($N = 10,000$) across 50 subsamples

A.7 Further experiment results: the twins dataset

Partitioning The following plot shows the final tree built for the twins dataset. Red nodes indicate positivity violation population.

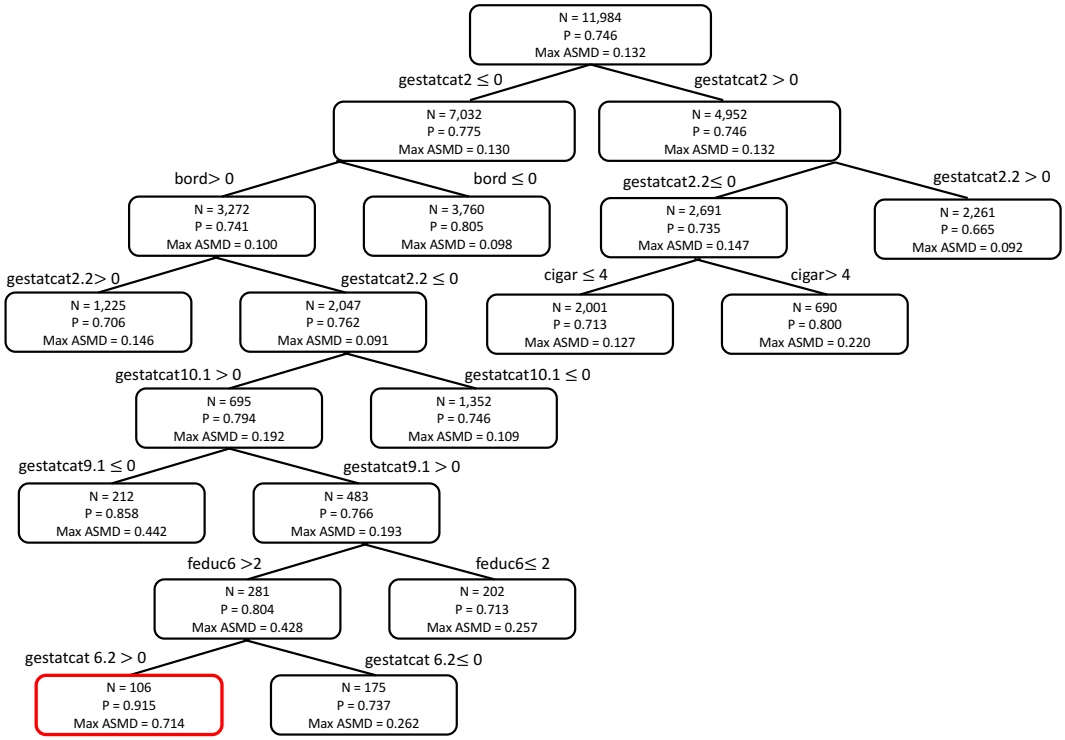

Figure A16: Tree structure after training on the entire twins dataset ($N = 11,984$). Violating leaf nodes are marked in red. P is for node prevalence.

Causal effect estimation Figure A17 shows a comprehensive comparison of the different models. Causal Tree shows poor performance, whereas Causal Forest performs comparably to BICauseTree(Marginal) and IPW.

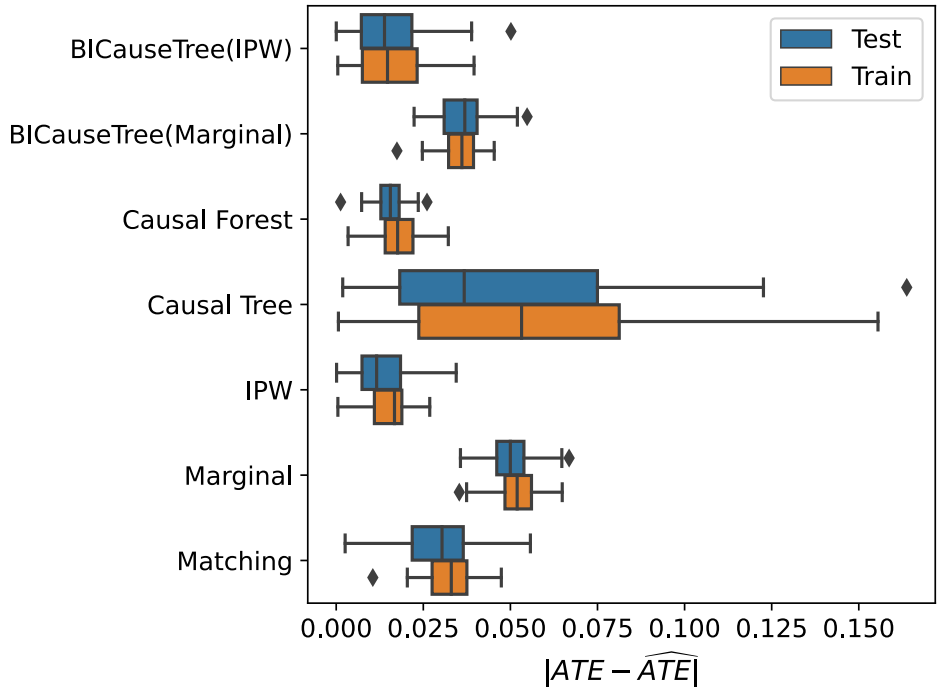

Figure A17: Estimation bias on the twins dataset ($N = 11,984$) across 50 subsamples

Adding Coarsened Exact Matching When comparing Mahalanobis Matching and Coarsened Exact Matching on the realistic benchmark twins dataset, we see that MM performs slightly better. This is likely due to the higher dimensionality, which makes exact matching difficult, as well as the presence of continuous variables and potential inter-feature correlations that MM can address. BiCause Tree shares some similarities with CEM, but the strength of our approach lies in performing "blocking" in a principled way, based on ASMD. We believe this additionnal comparison enhances our comparative analysis and strengthens the validity of our approach.

Table A2: Mean estimation bias and estimation variance for the twins dataset

Method	Sample	Bias	Variance
BICauseTree (IPW)	Test	0.017	0.011
BICauseTree (IPW)	Train	0.018	0.012
BICauseTree (Marginal)	Test	0.038	0.009
BICauseTree (Marginal)	Train	0.035	0.008
IPW	Test	0.014	0.013
IPW	Train	0.018	0.010
Marginal	Test	0.050	0.000
Marginal	Train	0.052	0.009
MM	Test	0.031	0.019
MM	Train	0.034	0.016
CEM	Test	0.025	0.015
CEM	Train	0.024	0.012

Outcome estimation BICauseTree(Marginal) shows good calibration on outcome estimation. BICauseTree(IPW) seems to have some bias in outcome estimation. This is possibly due to the parametric nature of our IPW model which uses a Logistic Regression. Matching shows poor outcome estimation ability.

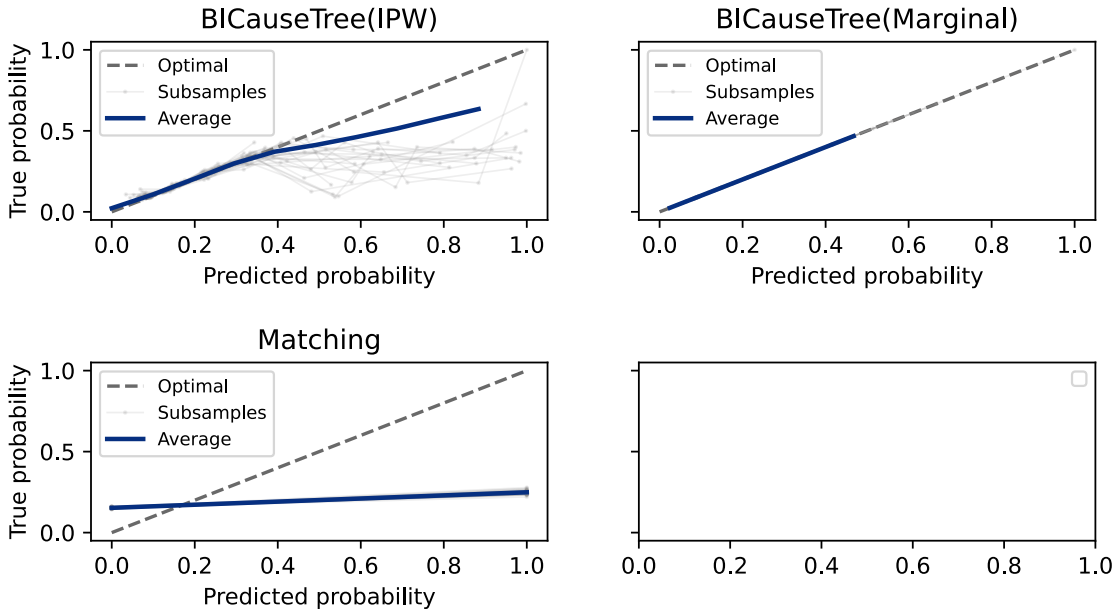

Figure A18: Calibration of the outcome estimation across 50 subsamples, on the twins testing set ($N = 5,992$).

Treatment allocation bias reduction with tree depth Figure A19 below shows the reduction in ASMD for the top 10 most imbalanced features in the entire population. It compares BICauseTree models with varying maximum tree depth hyperparameters. In all 10 covariates, we notice how the ASMD reduces as the maximum tree depth increases.

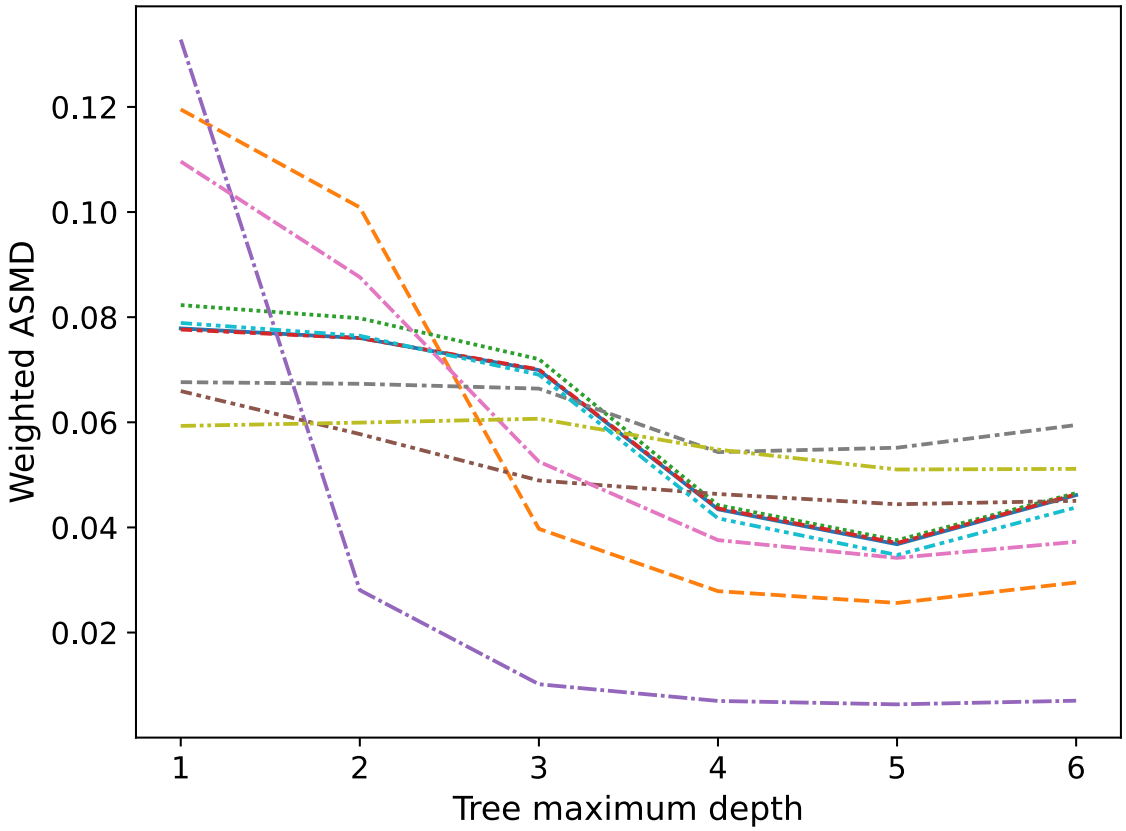

Figure A19: Weighted ASMD for the top 10 covariates with the highest ASMD in the initial population for *BICause-Tree(Marginal)* models with varying maximum tree depths on the twins dataset training set ($N = 5,992$) across 50 subsamples

A.8 Further experiment results: the ACIC dataset

Partitioning The following shows the tree built for the ACIC dataset.

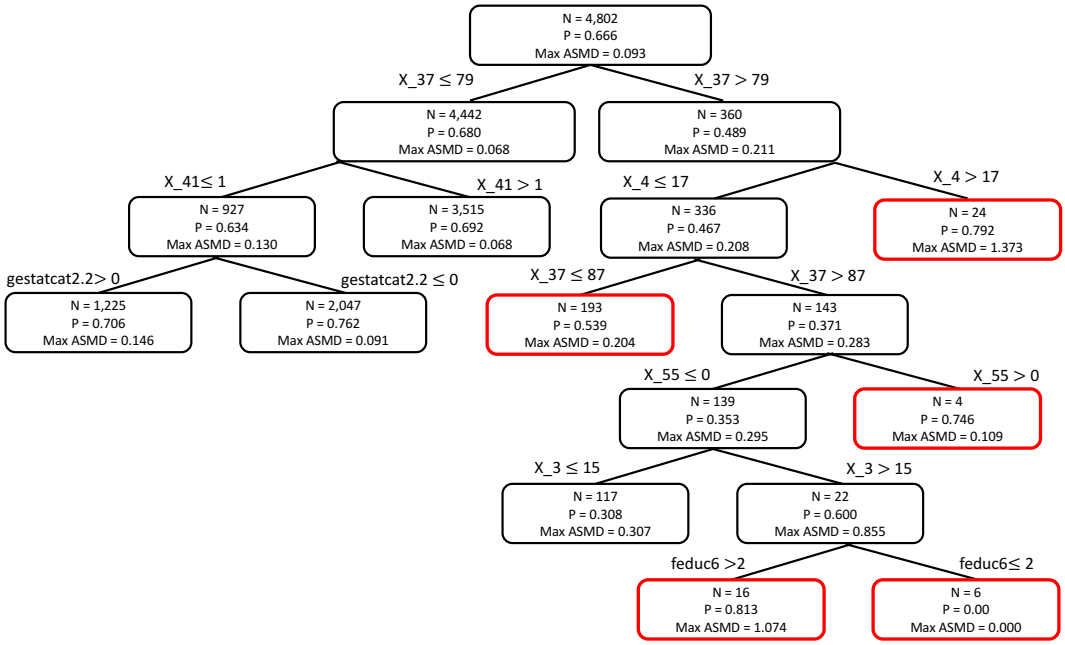

Figure A20: Tree structure after training on the entire ACIC dataset ($N = 4,802$). Violating leaf nodes are marked in red. P is for node prevalence.

Causal effect estimation Figure A17 shows a comprehensive comparison of the different models. The performance of both BICauseTree models compares with the one from IPW or Causal Forest. Causal Tree however shows poor performance. Matching shows high estimation bias compared to all other models, including the Marginal estimator.

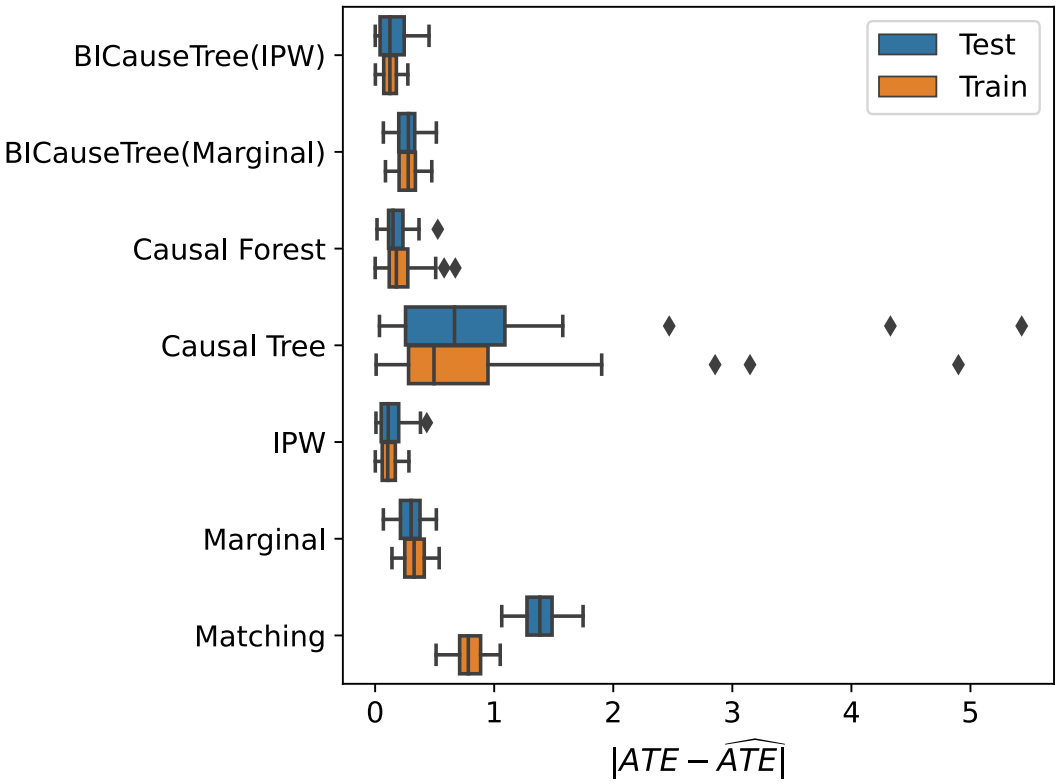

Figure A21: Estimation bias on the ACIC dataset testing set ($N = 960$) across 50 subsamples

Propensity score estimation We show here the calibration of propensity scores in the ACIC dataset. BICauseTree performs less well than IPW, which is more tailored for propensity estimation. As the outcome is non-binary in ACIC, we are not able to generate a calibration plot comparing the outcome estimation of BICauseTree with the one from existing approaches.

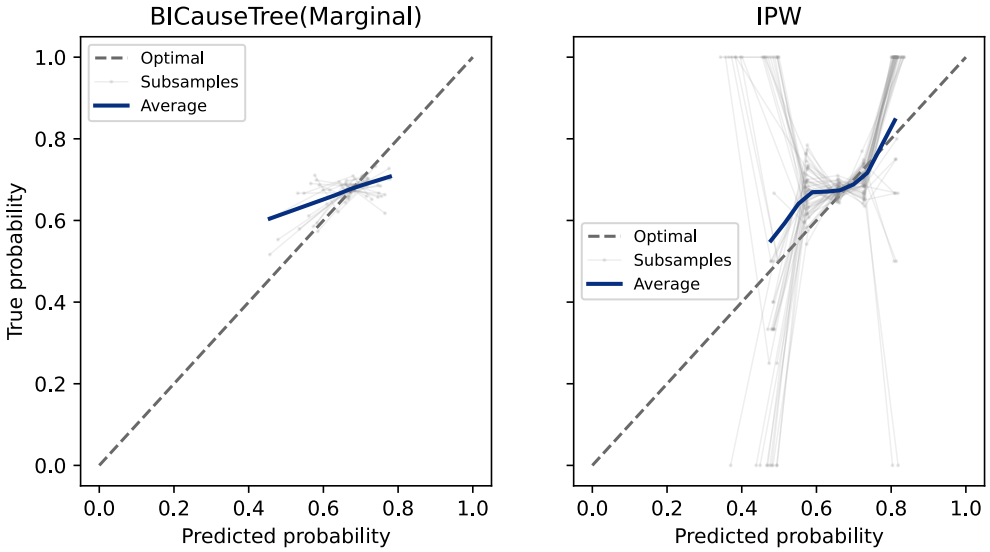

Figure A22: Propensity score calibration comparing our approach to existing alternatives on the ACIC dataset testing set ($N = 960$) across 50 subsamples

Effect estimation bias reduction with tree depth Figure A23 below shows the estimation bias for various maximum tree depth hyperparameters. We notice some overlap to the IPW confidence interval, and bias reduces

for depths up to max depth 6, then the variance across subsamples starts to increase. This may be due to the limited sample size of this dataset.

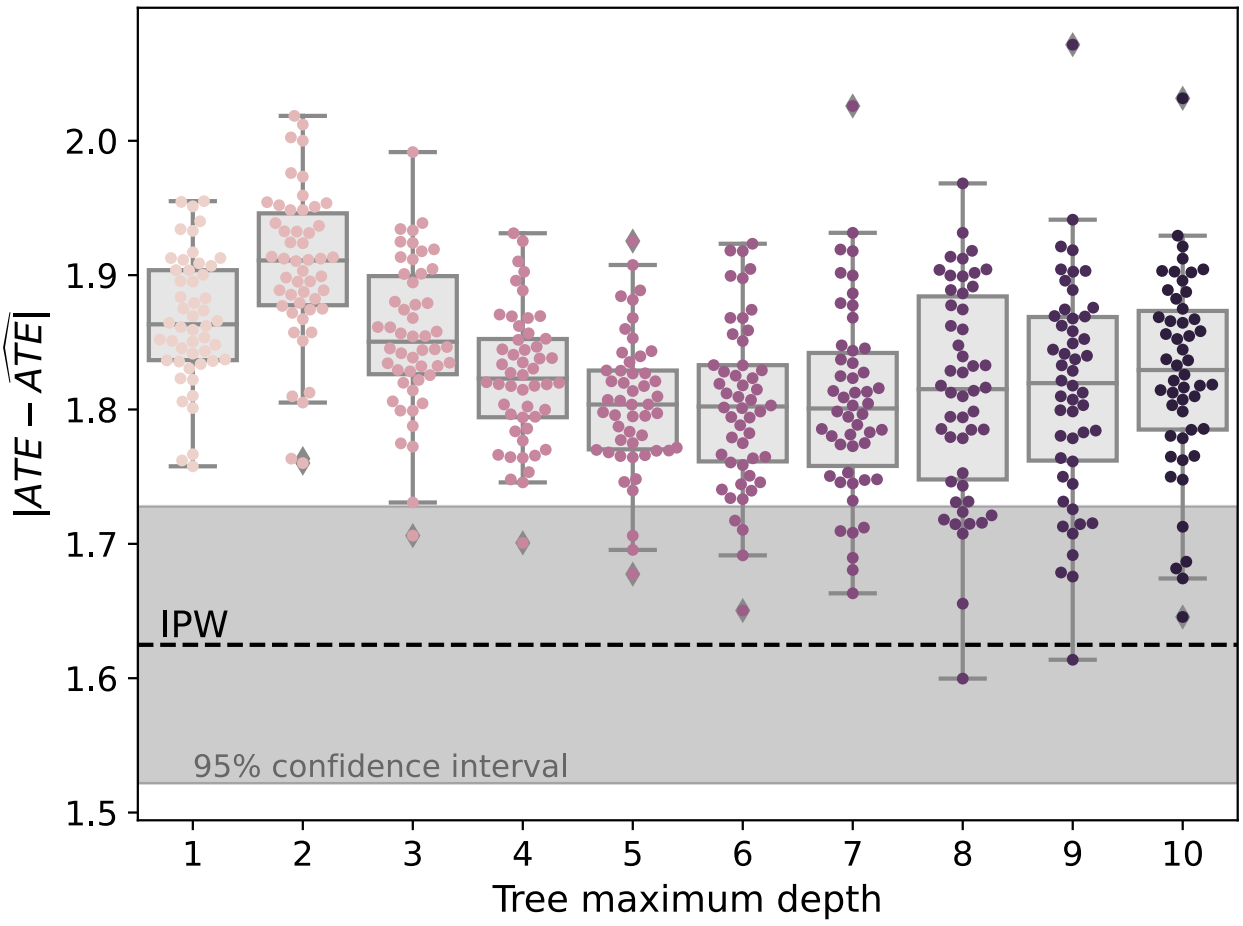

Figure A23: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the ACIC training set ($N = 3,842$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure A24 below shows the reduction in ASMD for the top 10 most imbalanced features in the entire population. It compares BICauseTree models with varying maximum tree depth hyperparameters. All 10 covariates show a reduction in ASMD with depth.

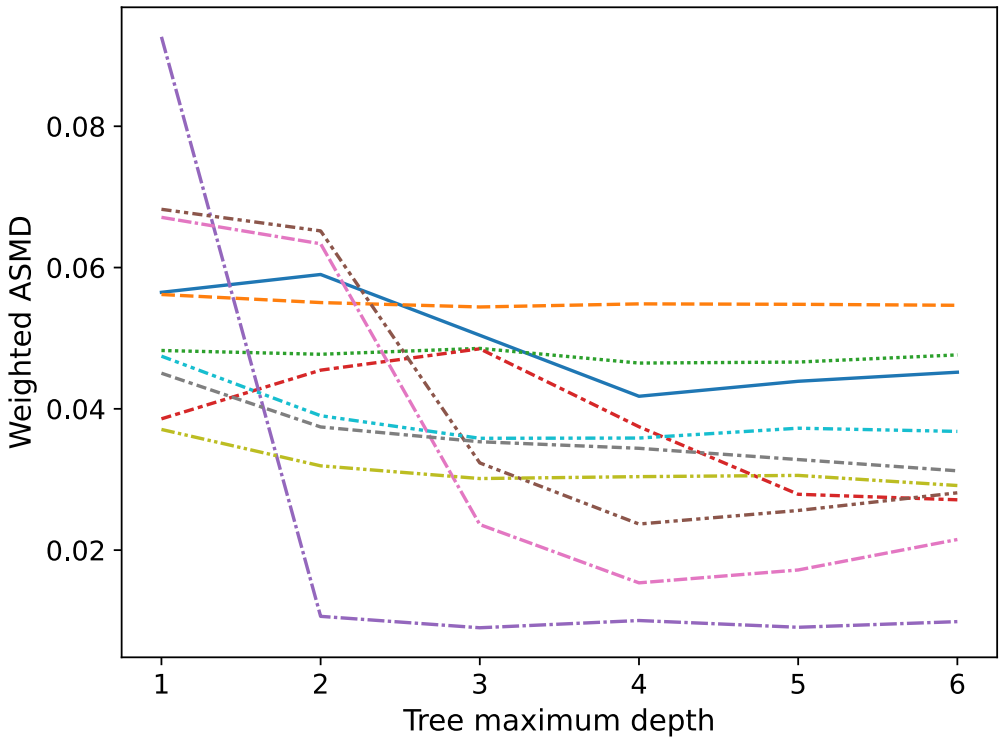

Figure A24: Weighted ASMD across maximum tree depths for the top 10 covariates with the highest ASMD in the initial population, applying *BICauseTree(Marginal)* models with varying maximum tree depths on the ACIC dataset training set ($N = 3,842$) across 50 subsamples

A.9 User Study Questions for Average Treatment Effect (ATE) Estimation

We interviewed 25 participants had expertise in Statistics, but no expertise in causal inference. Non-binary responses were given on a Likert scale from 1 (Very unclear) to 5 (Very clear.

1. **Clarity**
   Question: How clear was the explanation of the average treatment effect provided by this method?
   Follow-up: Please describe any part of the explanation you found unclear.

2. **Trustworthiness**
   Question: How trustworthy do you find the estimated average treatment effect based on the explanation provided?
   Follow-up: What, if anything, made you question the trustworthiness of the explanation?

3. **Actionability**
   Question: Based on the explanation, how confident are you that you could make a decision or policy change involving the treatment effect?
   Follow-up: What additional information, if any, would improve your ability to act on this explanation?

4. **Causal Understanding**
   Task-based Question: Summarize the relationship between the treatment and the outcome as described in the explanation. What is the estimated average treatment effect?
   Scoring: Does the treatment increase or decrease the outcome, and by how much?

5. **Low cognitive Load**
   Question: How easy was it to understand the explanation of the treatment effect?
   Follow-up: Which part of the explanation, if any, did you find most challenging?

6. **Time to Understand**
   Task-based Measurement: Track the time participants take to read and summarize the explanation.

Question: Do you feel you had enough time to fully understand the explanation?
Scale: Yes/No
Follow-up: How much more time would you need to feel confident in your understanding?

7. **Intuitiveness**
Question: How intuitive did you find the method used to compute the average treatment effect in this explanation?
Follow-up: What aspects of the method made it more or less intuitive to understand the treatment effect estimation?

Method Comparison Table

Section	Matching	IPW	Causal Tree	BICause Tree
Clarity	3.84	3.08	2.84	4.68
Trustworthiness	3.88	4.28	4.0	3.32
Actionability	3.76	4.04	3.36	3.6
Causal Understanding	4. 76	4.8	3.88	4.72
Low cognitive Load	4.64	3.76	2.48	4.64
Time to Understand	1 min	2 min	3 min	1 min
Intuitiveness	4.08	3.72	2.92	4.8

Table A3: Comparison of interpretability metrics across four methods, on a Likert scale from Not very likely (1) to very likely (5). For Time to Understand, we report the most common response. Averaged across $N = 25$ participants.

A.10 Implementation

A.10.1 Positivity violations evaluations

We implemented two positivity violations definition procedures: the Crump method and a method we introduced, which we call the *symmetric prevalence* threshold method. The Crump method is a data-driven method that defines a threshold for extreme propensity scores (i.e., positivity violations), based on the distribution of propensity scores in the data (see further results in (Crump et al., 2009)).

The symmetric prevalence procedure generates upper and lower cutoff values that are adjusted for the prevalence. The cutoffs are computed such that if the overall prevalence was 0.5 they would be symmetrical (e.g. for $\alpha = 0.05$ the cutoffs would be 0.05 and 0.95). We may consider this as class reweighting of the propensities, within the entire population.

If we denote the overall prevalence μ, and consider α as the cutoff had the distribution been symmetric (we recommend taking $\alpha = 0.1$), the cutoffs are computed as follows:

$$\textit{Upper cutoff} = \frac{(1-\alpha) * \mu}{(1-\alpha) * \mu + \alpha * (1-\mu)}$$

$$\textit{Lower cutoff} = \frac{\alpha * \mu}{\alpha * \mu + (1-\alpha) * (1-\mu)}$$

A.11 Experimental details and computation

For BICauseTree, most hyperparameters were set to their default value. Multiple hypothesis test correction was done following a step-down method using Holm-Bonferroni adjustments (Holm, 1979; Abdi, 2010), with $\alpha = 0.05$. The threshold for weight trimming for positivity violations was computed using the Crump procedure (Crump et al., 2006, 2009) with 10000 segments. Throughout all experiments the minimum treatment group size was set to 2 patients. The maximum depth is the only parameter which varied depending on the experiments. It was set to 5 for both synthetic datasets and 10 for the experiments on the twins dataset. A long-standing practice has been to define any covariate with $ASMD \geq 0.10$ as a potentially problematic confounder (Austin, 2009), so we set this as our default threshold and used it in our experiments.
For IPW(LR), we used a Logistic Regression with a *saga* solver, no penalty and a maximum number of iterations equal to 500. For comparison purposes, BICauseTree(IPW) used similar hyperparameters for its internal IPW

outcome model. For IPW(GBT) we used default hyperparameters.
We applied double matching based on the Mahalanobis distance for all datasets but ACIC, on which we used an Euclidean distance to avoid non-invertable matrices caused by the sparsity of non-null column values.
Causal Tree with a single estimator and a subforest size of 1. For Causal Forest, we used 50 estimators and a subforest size of 1.

For including Causal Tree into our experiments, we used the code available at `https://github.com/py-why/EconML`. Outcome and propensity models were trained using `sklearn` with default parameters and 500 maximum iterations for Logistic Regression when relevant. Statistical testing was implemented using the `statsmodel` package.

A.11.1 Compute and Runtime

In general, BICauseTree inherits its computational efficiency from decision trees. Namely the fitting complexity is $O\left(M \cdot V \cdot N \log(N)\right)$, Where N it the number of observations, M is the number of features, and $V = \max_{m \in M}(V_m)$ the maximal number of unique levels among all M features. Briefly, recursively splitting the tree is $O(N \log(N))$, Additionally, for every recursive iteration we consider all M features (calculating ASMD is $O(N)$ for each feature), and, among the select feature we consider a split cutoff over all its unique values (note this is 1 for binary features, calculating a statistical test takes $O(N)$ for each level). Therefore the resulting complexity is $O\left(MVN \log(N)\right)$.

Table A4 shows the average runtime it took for each model to fit. We see BICauseTree compares with IPW, Causal Tree and Causal Forest in terms of wall-clock run-time, while Matching has higher compute time than all other models.

Experiment	Amount of Compute
Natural experiment dataset	
BICauseTree(Marginal)	286
IPW	134
Matching	882
Causal Tree	183
Causal Forest	390
Positivity violations dataset	
BICauseTree(Marginal)	258
IPW	128
Matching	929
Causal Tree	137
Causal Forest	384
Twins	
BICauseTree(Marginal)	420
BICauseTree(IPW)	567
IPW	329
Matching	2283
Causal Tree	403
Causal Forest	672
ACIC	
BICauseTree(Marginal)	329
BICauseTree(IPW)	376
IPW	239
Matching	1092
Causal Tree	354
Causal Forest	439

Table A4: Total amount of compute in seconds for model fitting across 50 train-test splits, for selected experiments. All experiments were run on a 10 core CPU Apple M1 Pro.

We further tested the runtime of BICauseTree as a function of dimensionality. Figure A25 shows the time, averaged over 10 random splits, it took for BICauseTree to fit—including filtering for positivity—as a function of the number of covariates in the input data (following the experiment design presented earlier in the Appendix in Figure A3.

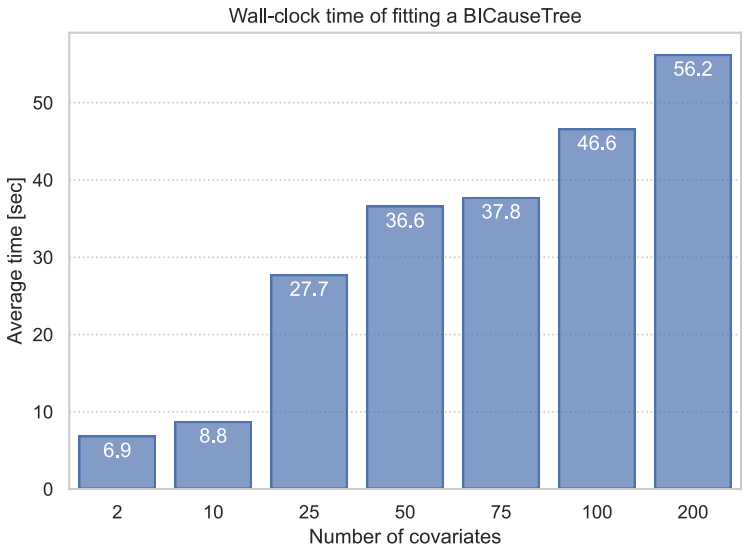

Figure A25: Runtime of BICauseTree as a function of increasing the dimensionality of the input. The bars show the average wall-clock time, in seconds, it took to fit a BICauseTree for different numbers of input covariates.

A.11.2 Causal benchmark datasets

The **twins dataset** was originally taken from the denominator file at `https://www.nber.org/research/data/linked-birthinfant-death-cohort-data`. However, we use data generated by *Neal et. al* (Neal et al., 2020), which simulates an observational study from the initial data by selectively hiding one of the twins with a generative approach. The sample size is $N = 11,984$ pairs of twins, with the essential inclusion criterion being that both individuals were born weighing less than 2kg. The mortality rate amongst the lighter twins is 18.9%, and for the heavier 16.4%, for an average treatment effect of -2.5% (which we thus consider as ground truth). A total of 75 covariates were recorded, relating to the parents' socio-demographic features and medical history, the pregnancy and the birth.

The **ACIC dataset** was generated using the *causallib* package at `https://github.com/BiomedSciAI/causallib/blob/master/causallib/datasets/data/acic_challenge_2016/README.md`. It contains covariates, simulated treatment, and simulated response variables for the causal inference challenge in the 2016 Atlantic Causal Inference Conference (Dorie et al., 2019). For each of 20 conditions, treatment and response data were simulated from real-world data corresponding to 4802 individuals and 58 covariates. After one-hot encoding, a total of 79 covariates was included. More specifically, we used the set of treatment and response variables *zymu 174570858*, with the two expected potential outcomes (*mu0*, *mu1*).

A.12 Social impact of our work: further details

Recent years have seen a surge in the Explainable AI (XAI) literature, motivated by rising ethical concerns around artificial intelligence. Model scrutiny is particularly relevant in sensitive environments such as healthcare, which require high safety standards considering the major consequences of invalid predictions on individual trajectories (Yampolskiy, 2018). Calls for model transparency are further motivated by evidences that supervised machine learning is inclined to reproduce inherent bias and prejudice against discriminated groups (Challen et al., 2019). Ultimately, XAI aims at building trust in models that ought to be deployed. In recent years, the demand for transparency expanded beyond the research community, notably with incentives from high institutions. The European Union General Data Protection Regulation legislation has mandated a "right to explanation" for individual predictions that can "significantly affect" users (Holzinger et al., 2017). In other words, algorithmic results should be re-traceable on demand. In parallel, quality control frameworks such as the U.S. Food and Drug Administration guidance have been introduced to ensure the safety of clinical AI (Benjamens et al., 2020). By providing interpretable effect estimation, our BICauseTree approach aligns with the mission of increasing transparency for downstream users. As such, it is most likely to have positive social impact. We however caution against relying exclusively on causal inference when data accuracy or volumes are insufficient, as misleading effect

estimates may have negative social impact.