Shrinkage Bayesian Causal Forest with Instrumental Variable

Lennard Maßmann*
University of Duisburg-Essen
Faculty of Business Administration and Economics
Universitätsstraße 12, 45117 Essen, Germany
lennard.massmann@uni-due.de

Jens Klenke
University of Duisburg-Essen
Faculty of Business Administration and Economics
Universitätsstraße 12, 45117 Essen, Germany
jens.klenke@vwl.uni-due.de

Abstract

We propose Shrinkage Bayesian Causal Forest with Instrumental Variable (SBCF-IV), a method for discovering and estimating subgroups with heterogeneous Complier Average Causal Effects (CACE) in sparse high-dimensional settings with imperfect compliance. SBCF-IV places a sparsity-inducing Dirichlet prior on splitting probabilities within Bayesian Additive Regression Trees used to estimate the conditional intention-to-treat and the complier share, concentrating posterior mass on the few covariates that genuinely moderate the complier effect. The resulting split frequencies then enter a downstream CART as variable-level costs, directing it toward the relevant moderator variables and yielding an interpretable partition of the covariate space. The Bayesian feature-selection mechanism thus regularizes effect estimation and simultaneously steers subgroup discovery. Monte Carlo experiments show that, as the share of irrelevant covariates grows, SBCF-IV recovers the true partition more reliably than BCF-IV at both the tree and unit level, and retains nominal coverage in regimes where BCF-IV's intervals deteriorate. We apply the method to the Oregon Health Insurance Experiment and the 401(k) eligibility data.

1 Introduction

The heterogeneous treatment effect (HTE) literature has expanded along two complementary dimensions: estimating the conditional average treatment effect (CATE) and discovering interpretable subgroups whose effects deviate from the population average (Lipkovich et al. 2017, Künzel et al. 2019, Dwivedi et al. 2020). For CATE estimation, nonparametric machine learning methods have become standard including causal forests (Athey et al. 2019), Bayesian Additive Regression Trees (BART) (Hill 2011), Bayesian Causal Forest (BCF) (Hahn et al. 2020, Caron et al. 2022), and doubly robust meta-learners (Kennedy 2023, Semenova & Chernozhukov 2021). In parallel, a growing strand of work has focused on the data-driven discovery of interpretable subgroups, with decision-tree-based methods (Athey & Imbens 2016, Bargagli-Stoffi & Gnecco 2020, Bargagli-Stoffi et al. 2022, Lee et al. 2021) being particularly prominent due to their interpretability. Ensemble-based extensions such as the causal rule ensemble (Bargagli-Stoffi et al. 2024) and causal distillation trees (Huang et al.

*RGS Econ - Ruhr Graduate School in Economics, Hohenzollernstraße 1-3, 45128 Essen, Germany

Preprint.

2025) address shortcomings of instability of single-tree methods by aggregating decision rules across many trees, yielding more stable and expressive subgroup representations when estimating CATE.

A limitation of recent work on HTE is that the identification for CATE rests on the assumption of regular assignment mechanisms, which is rarely defensible in observational studies with unobserved heterogeneity in treatment uptake. When a valid instrument is available, a relevant target estimand becomes the Complier Average Causal Effect (CACE), or Local Average Treatment Effect (LATE), nonparametrically identified for the subpopulation of compliers under the standard Instrumental Variables (IV) assumptions (Imbens & Angrist 1994, Angrist et al. 1996). The interpretation of complier effects has been the subject of ongoing debate: because compliance status is a counterfactual quantity that is never directly observed, critics have argued that the CACE pertains to an unidentified subgroup and is therefore of limited policy relevance (Deaton 2010, Swanson & Hernán 2014). In contrast, the LATE is often the most one can learn nonparametrically in the presence of unmeasured confounding without imposing restrictive effect homogeneity assumptions, and it remains informative about the underlying causal structure. Crucially, when covariates explain a substantial share of the variation in compliance, complier effects effectively coincide with conditional effects in identifiable subgroups, and the concerns about an unknown target population largely dissolve (Kennedy et al. 2020). This observation provides direct motivation for studying the conditional Complier Average Causal Effect (cCACE). The cCACE is the CACE as a function of observed characteristics. Characterizing how complier effects vary along interpretable covariate profiles transforms the CACE from a property of an unobserved subgroup into a set of policy-relevant statements about identifiable populations. Several methods have been developed to recover heterogeneous IV effects, each targeting a distinct inferential object. Forest-based estimators such as the instrumental variable forests of Wang et al. (2022) and the instrumental forest within Generalized Random Forests (Athey et al. 2019) target a unit-level conditional IV function, delivering pointwise-consistent estimates with valid asymptotic inference but no explicit partition of the covariate space; the drivers of heterogeneity are recovered post hoc through variable importance scores or best linear projections. Partition-based procedures like the IV tree of Wang et al. (2022) and the matching procedure of Johnson et al. (2022) return an interpretable partition over which subgroup-level effects can be read off directly, paired with closed-testing inference.

We propose Shrinkage Bayesian Instrumental Variable Causal Forest (SBCF-IV), a generalization of Bayesian Instrumental Variable Causal Forest (BCF-IV) tailored to settings in which the share of covariates that drive effect heterogeneity is small relative to P. The BCF-IV algorithm of Bargagli-Stoffi et al. (2022) combines a Bayesian Additive Regression Trees (BART)-based (Chipman et al. 2010) sum-of-trees estimator for the conditional Intention-To-Treat (cITT) with a shallow Classification and Regression Trees (CART) post-processing step (Breiman et al. 1984), yielding interpretable subgroup-level estimates of cCACE within a two-step procedure based on stratification and IV estimation. By doing so, it reconciles the predictive accuracy of ensemble methods (Athey et al. 2019, Hartford et al. 2017) with the interpretability of single-tree approaches (Athey & Imbens 2016, Bargagli-Stoffi & Gnecco 2020, Johnson et al. 2022). However, BCF-IV inherits BART's uniform split-variable prior and therefore performs no targeted feature selection: when the covariate vector contains many irrelevant variables, as is typical in modern administrative or biomedical data, both the ensemble and the downstream CART can dilute their attention across spurious moderators that weaken the quality of the discovered subgroups. Our contribution is twofold. First, we replace the BCF estimator of the cITT with the Shrinkage Bayesian Causal Forest (SBCF) of Caron et al. (2022), which augments BART's uniform split-variable prior with a sparsity-inducing Dirichlet prior in the spirit of SoftBART (Linero & Yang 2018, Linero 2018), delivering fully Bayesian feature shrinkage and allowing the data to concentrate posterior mass on the few covariates that actually moderate the treatment effect. Second, we feed the resulting posterior split frequencies into the subgroup-discovery CART as variable-level costs, steering the partition toward the covariates the ensemble has identified as most relevant and thereby improving the discovery of heterogeneous subgroups. Throughout, our scope follows that of Bargagli-Stoffi et al. (2022): a binary randomized instrument, a binary treatment, and standard IV identification assumptions. Section 4 accordingly benchmarks SBCF-IV against BCF-IV as the direct ancestor and natural comparator under this setting.

The paper proceeds as follows. Section 2 sets up the potential outcomes framework and identifying assumptions for the cCACE under an irregular assignment mechanism. Section 3 introduces SBCF-IV. Section 4 reports Monte Carlo evidence on its performance relative to BCF-IV in high-dimensional settings based on tree-level and unit-level performance criteria. Section 5 applies SBCF-IV to two

empirical studies: the Oregon Health Insurance Experiment (OHIE) (Finkelstein et al. 2012, Johnson et al. 2022) and the 401(k) retirement plans dataset (Poterba et al. 1992, 1995, Chernozhukov et al. 2018). Section 6 concludes.

2 Potential outcomes and irregular assignment

We follow Bargagli-Stoffi & Gnecco (2020) and Bargagli-Stoffi et al. (2022) and adopt Rubin's causal model, working from the outset within the irregular assignment framework of Imbens & Rubin (2015). For N units indexed $i = 1, \dots, N$, let $Y_i \in \mathbb{R}$ denote the observed outcome, $Z_i \in \{0, 1\}$ a binary instrument (assignment), $W_i \in \{0, 1\}$ the actual treatment received, and $X_i \in \mathbb{R}^P$ the i-th row of an $N \times P$ matrix X of pre-treatment covariates. Each unit is endowed with potential outcomes $Y_i(Z_i = z, W_i = w)$ and potential treatments $W_i(z)$ for $z, w \in \{0, 1\}$, related to observed quantities by the consistency relations $Y_i = Y_i(Z_i, W_i)$ and $W_i = W_i(Z_i)$. The instrument Z_i is unconfounded but the receipt W_i may be confounded. This is the canonical IV setting, and our inferential target is the cCACE in Definition 2.2, based on the latent subpopulation of compliers.

Definition 2.1 (Compliance subgroups)**.** Each unit belongs to one of four latent compliance subgroups, defined by the joint values of its potential treatments:

$$G_i = \begin{cases} C, & W_i(0) = 0,\ W_i(1) = 1 \quad \text{(compliers)} \\ D, & W_i(0) = 1,\ W_i(1) = 0 \quad \text{(defiers)} \\ AT, & W_i(0) = 1,\ W_i(1) = 1 \quad \text{(always-takers)} \\ NT, & W_i(0) = 0,\ W_i(1) = 0 \quad \text{(never-takers)}, \end{cases}$$

with conditional subgroup proportions $\pi_G(x) = \Pr(G_i = G \mid X_i = x)$ for $G \in \{C, D, AT, NT\}$.

Definition 2.2 (Conditional CACE)**.** The conditional Complier Average Causal Effect is the estimand

$$\tau^{\text{CACE}}(x) \;:=\; \mathbb{E}\big[Y_i(1, W_i(1)) - Y_i(0, W_i(0)) \;\big|\; G_i = C,\, X_i = x\big],$$

which, under the exclusion restriction in Assumption 2.1(d), reduces to $\mathbb{E}[Y_i(1) - Y_i(0) \mid G_i = C, X_i = x]$.

Definition 2.2 fixes the target as a property of the latent complier subpopulation and it is not, by itself, an object computable from the observed distribution of (Y_i, W_i, Z_i, X_i). Identification proceeds through the conditional intention-to-treat effect and the conditional complier share, both of which admit clean expressions in terms of observed conditional means.

Definition 2.3 (Conditional ITT and complier share)**.** The conditional intention-to-treat effect and the conditional complier share are

$$\begin{aligned} \text{ITT}_Y(x) \;&:=\; \mathbb{E}[Y_i \mid Z_i = 1, X_i = x] - \mathbb{E}[Y_i \mid Z_i = 0, X_i = x], \\ \pi_C(x) \;&:=\; \Pr(G_i = C \mid X_i = x). \end{aligned}$$

By a mixture argument over the compliance subgroups of Definition 2.1,

$$\text{ITT}_Y(x) \;=\; \pi_C(x)\,\text{ITT}_{Y,C}(x) + \pi_D(x)\,\text{ITT}_{Y,D}(x) + \pi_{AT}(x)\,\text{ITT}_{Y,AT}(x) + \pi_{NT}(x)\,\text{ITT}_{Y,NT}(x),$$

where $\text{ITT}_{Y,G}(x)$ denotes the conditional ITT among units of compliance types G in Definition 2.1.

Definition 2.3 makes explicit that $\text{ITT}_Y(x)$ is a covariate-weighted mixture and isolating $\tau^{\text{CACE}}(x) = \text{ITT}_{Y,C}(x)$ requires assumptions that neutralize the contributions of defiers, always-takers, and never-takers. The classical Angrist et al. (1996) conditions for IV estimation deliver point identification.

Assumption 2.1 (IV identification under irregular assignment)**.**

(a) *SUTVA / consistency:* $Y_i = Y_i(Z_i, W_i)$ and $W_i = W_i(Z_i)$, with no interference between units and no hidden treatment variants.

(b) *Relevance:* $\pi_C(x) > 0$ almost surely.

(c) *Unconfounded instrument:* $Z_i \perp\!\!\!\perp \big(\{Y_i(z, w)\}_{z,w \in \{0,1\}^2},\, W_i(0),\, W_i(1)\big) \;\big|\; X_i$.

(d) *Exclusion restriction:* $Y_i(z, w) = Y_i(w)$ for all $z, w \in \{0, 1\}$.

(e) *Monotonicity:* $W_i(1) \geq W_i(0)$.

Assumption 2.1(a) ensures consistency and rules out interference and hidden treatment variants. Relevance (b), together with monotonicity (e), guarantees $\pi_C(x) > 0$ almost surely so that the identification ratio below is well-defined; monotonicity also rules out defiers ($\pi_D(x) = 0$), which must be defended on substantive grounds or enforced through a one-sided non-compliance design. The exclusion restriction (d) confines the effect of Z_i on Y_i to the channel through W_i. Under Assumption 2.1, the complier estimand of Definition 2.2 is identified from the observed distribution.

Proposition 2.1 (Identification of $\tau^{\text{CACE}}(x)$)**.** *Under Assumption 2.1,*

$$\tau^{CACE}(x) \;=\; \frac{ITT_Y(x)}{\pi_C(x)} \;=\; \frac{\mathbb{E}[Y_i \mid Z_i = 1, X_i = x] - \mathbb{E}[Y_i \mid Z_i = 0, X_i = x]}{\mathbb{E}[W_i \mid Z_i = 1, X_i = x] - \mathbb{E}[W_i \mid Z_i = 0, X_i = x]}.$$

The proof, adapting Angrist et al. (1996) to the conditional target as in Bargagli-Stoffi et al. (2022), is given in Appendix A.

Given a partition $\{\mathbb{X}_j\}_j$ of the covariate space, Proposition 2.1 motivates a sample-moment estimator targeting the subgroup-averaged complier effect

$$\tau^{\text{CACE}}_{\mathbb{X}_j} \;\coloneqq\; \mathbb{E}\big[Y_i(W=1) - Y_i(W=0) \;\big|\; G_i = C,\, X_i \in \mathbb{X}_j\big], \tag{2.1}$$

obtained by replacing population conditional means in Proposition 2.1 with subgroup sample analogues.

Definition 2.4 (Subgroup-wise 2SLS estimator)**.** For $X_i = x \in \mathbb{X}_j$, with $N_{z,j} = \sum_{l:\, X_l \in \mathbb{X}_j} \mathbf{1}\{Z_l = z\}$ the count of $\mathbb{X}_j$-units assigned to $Z_l = z \in \{0, 1\}$,

$$\widehat{\tau}^{\text{CACE}}(x) \coloneqq \widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j} \;=\; \frac{\dfrac{1}{N_{1,j}} \sum_{l:\, X_l \in \mathbb{X}_j} Y_l Z_l \;-\; \dfrac{1}{N_{0,j}} \sum_{l:\, X_l \in \mathbb{X}_j} Y_l (1 - Z_l)}{\dfrac{1}{N_{1,j}} \sum_{l:\, X_l \in \mathbb{X}_j} W_l Z_l \;-\; \dfrac{1}{N_{0,j}} \sum_{l:\, X_l \in \mathbb{X}_j} W_l (1 - Z_l)}$$

targets $\tau^{\text{CACE}}_{\mathbb{X}_j}$.

Equivalently, $\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j}$ is the Two-Stage Least Squares estimator on the subgroup-restricted simultaneous system

$$Y_{i,\mathbb{X}_j} \;=\; \alpha_{\mathbb{X}_j} + \tau^{\text{CACE}}_{\mathbb{X}_j} W_{i,\mathbb{X}_j} + \varepsilon_{i,\mathbb{X}_j}, \qquad W_{i,\mathbb{X}_j} \;=\; \pi_{0,\mathbb{X}_j} + \pi_{C,\mathbb{X}_j} Z_{i,\mathbb{X}_j} + \eta_{i,\mathbb{X}_j}, \tag{2.2}$$

with $\mathbb{E}[\varepsilon_{i,\mathbb{X}_j}] = \mathbb{E}[\eta_{i,\mathbb{X}_j}] = 0$ and $\mathbb{E}[Z_{i,\mathbb{X}_j}\, \eta_{i,\mathbb{X}_j}] = 0$. Under Assumption 2.1 and a sufficient number of i.i.d. observations within each $\mathbb{X}_j$, $\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j}$ is consistent and asymptotically normal for $\tau^{\text{CACE}}_{\mathbb{X}_j}$, with the reduced form and formal asymptotic results collected in Appendix B.

Definition 2.4 presumes that the partition $\{\mathbb{X}_j\}_j$ is known. The contribution of BCF-IV (Bargagli-Stoffi et al. 2022) is to discover this partition from the data through (i) an honest split of the sample into disjoint discovery and inference subsets $\mathcal{I}_{\text{disc}}$ and $\mathcal{I}_{\text{inf}}$; (ii) interpretable discovery of heterogeneity on $\mathcal{I}_{\text{disc}}$; and (iii) inference for $\tau^{\text{CACE}}_{\mathbb{X}_j}$ on $\mathcal{I}_{\text{inf}}$. Our SBCF-IV algorithm in Algorithm 1 inherits this three-step structure of Bargagli-Stoffi et al. (2022) with adaptations for a high-dimensional covariate setup described in the next section.

3 Shrinkage Bayesian Causal Forests with Instrumental Variables (SBCF-IV)

We propose SBCF-IV, an extension of BCF-IV (Bargagli-Stoffi et al. 2022) to settings with many irrelevant covariates. The overall structure of honest sample splitting, data-driven discovery of heterogeneous subgroups on $\mathcal{I}_{\text{disc}}$, and 2SLS inference on $\mathcal{I}_{\text{inf}}$ is preserved. Our adaptations concern the discovery step where we replace BCF with the sparsity-inducing Shrinkage BCF (SBCF) of Caron et al. (2022) and feed its posterior variable-selection frequencies into the subgroup-finding tree as variable-level costs. Algorithm 1 summarizes the full procedure.

Working on $\mathcal{I}_{\text{disc}}$, we separately model the numerator and denominator of the identification ratio in Proposition 2.1. Following Hahn et al. (2020), we adopt the semi-parametric specification[2]

$$\mathbb{E}[Y_i \mid Z_i = z, X_i = x] \;=\; \mu\big(e(x), x\big) \;+\; \text{ITT}_Y(x)\, z, \tag{3.1}$$

[2]Background on CART, BART, BCF, and SBCF is collected in Appendix C.

where $e(x) = \Pr(Z_i = 1 \mid X_i = x)$ is the instrument's propensity score, included as a covariate in the control function $\mu(e(x), x)$ to mitigate regularization-induced confounding and targeted selection. Independent BART priors are placed on $\mu(e(x), x)$ and $\mathrm{ITT}_Y(x)$, with depth-penalty parameters $(\eta, \beta) = (3, 0.25)$ on $\mathrm{ITT}_Y(x)$ favoring shallow trees and hence simpler heterogeneity patterns. The compliance component is modeled analogously via

$$\mathbb{E}[W_i \mid Z_i = z, X_i = x] \;=\; \delta(z, x), \tag{3.2}$$

with a BART probit prior on $\delta(z, x)$ (Hill 2011). Combining estimates yields a pointwise complier-share estimate $\widehat{\pi}_C^{\text{SBCF}}(x) = \widehat{\delta}(1, x) - \widehat{\delta}(0, x)$ and a pointwise complier-effect estimate $\widehat{\tau}^{\text{SBCF}}(x) = \widehat{\mathrm{ITT}}_Y^{\text{SBCF}}(x)/\widehat{\pi}_C^{\text{SBCF}}(x)$. We emphasize that $\widehat{\tau}^{\text{SBCF}}(x)$ is a posterior estimate on $\mathcal{I}_{\text{disc}}$ used solely as a heterogeneity signal for the tree-fitting step, while we conduct final inference for the subgroup target $\tau_{\mathbb{X}_j}^{\text{CACE}}$ with $\widehat{\tau}_{\mathbb{X}_j}^{\text{2SLS}}$ from Definition 2.4 on $\mathcal{I}_{\text{inf}}$.

SBCF-IV departs from BCF-IV in the prior on the split-variable selection probabilities $s = (s_1, \ldots, s_P)$. BCF uses a uniform $s_j = 1/P$. Instead, SBCF imposes a sparsity-inducing Dirichlet prior,

$$s \sim \text{Dirichlet}\big(\tfrac{\alpha}{P}, \ldots, \tfrac{\alpha}{P}\big), \qquad \frac{\alpha}{\alpha + \rho} \sim \text{Beta}(a, b), \tag{3.3}$$

with defaults $(a, b, \rho) = (0.5, 1, P)$ (Caron et al. 2022). Small α concentrates mass on few covariates, and the hyperprior on $\alpha/(\alpha + \rho)$ lets the data determine the degree of sparsity, with the preference for sparsity strengthening as P grows. We place separate Dirichlet priors on the split probabilities s_μ and s_{ITT_Y} of the two components of (3.1), using $\rho_\mu = P + 1$ (to accommodate the propensity score as an extra covariate in μ) and $\rho_{\mathrm{ITT}_Y} = P/2$ (to further concentrate mass on few active splits in the treatment effect component). The same sparsity-inducing prior in (3.3) is applied in the SoftBART probit model used for $\delta(z, x)$ in (3.2), with $\rho_\delta = P + 1$. Appendix D provides more information on the full sparsity-inducing prior specifications. We fit a shallow CART (Breiman et al. 1984) to the pointwise estimates $\widehat{\tau}^{\text{SBCF}}(X_i)$ to recover an interpretable partition $\{\mathbb{X}_j\}_j$. The posterior split frequencies $\widehat{s}_{\mathrm{ITT}_Y}$ mainly indicate which covariates drive heterogeneity and we pass these split frequencies to `rpart` (Therneau & Atkinson 2026) through the `cost` argument, setting variable-level costs to

$$c_{\text{psp}} \;=\; \frac{\max\{\widehat{s}_{\mathrm{ITT}_Y}\}}{\widehat{s}_{\mathrm{ITT}_Y}}. \tag{3.4}$$

Because `rpart` divides split-improvement by the candidate variable's cost, (3.4) upweights covariates with a higher posterior inclusion probability at each split, in a manner analogous to variable-importance weighting in random forests (Breiman 2001). The cost vector c_{psp} is defined via $\hat{s}_{\mathrm{ITT}_Y}$, the posterior split frequencies of the cITT component, mirroring the Intention-to-Treat (ITT)-anchored discovery step of BCF-IV (Bargagli-Stoffi et al. 2022). When within-leaf variation of $\pi_C(x)$ is small, leaf-level heterogeneity in $\tau^{\text{CACE}}(x)$ is dominated by heterogeneity in $\mathrm{ITT}_Y(x)$, making $\hat{s}_{\mathrm{ITT}_Y}$ the natural cost weight. Aggregating $\hat{s}_{\mathrm{ITT}_Y}$ and $\hat{s}_\delta$ for regimes with substantial denominator-driven heterogeneity is a natural extension and is left to future work.

Inference on the discovered partition follows BCF-IV without modification. For each node $\mathbb{X}_j$ of the tree learned on $\mathcal{I}_{\text{disc}}$, we compute $\widehat{\tau}_{\mathbb{X}_j}^{\text{2SLS}}$ on $\mathcal{I}_{\text{inf}}$ via the simultaneous system in (2.2). Consistency and asymptotic normality under subgroup-level moment conditions are collected in Appendix B. To guard against spurious heterogeneity, nodes flagged by a first-stage F-test for weak instruments are discarded, and p-values across leaves are adjusted for the familywise error rate using Holm-corrections for adjusted p-values (Holm 1979, Bargagli-Stoffi et al. 2022).

4 Simulation study

We combine the design of Bargagli-Stoffi et al. (2022) with the high-dimensional setup of Caron et al. (2022) to assess SBCF-IV and BCF-IV in settings with many irrelevant covariates. Appendix E.1 presents a detailed description of the complete simulation design. For each of $N = 1{,}000$ units, we generate a binary instrument $Z_i \sim \text{Bin}(0.5)$, a covariate vector $X_i \in \mathbb{R}^P$ with $P \in \{10, 50, 100\}$ (half binary, half continuous), and potential outcomes and treatments according to

$$\begin{aligned} W_i(0) = 0, \quad W_i(1) \sim \text{Bin}(\pi_{\text{comp}} = 0.75), \\ Y_i(0) = \mu(X_i) + \epsilon_i, \quad Y_i(1) = Y_i(0) + W_i(1)\,\tau^{\text{CACE}}(X_i), \end{aligned} \tag{4.1}$$

Algorithm 1 Shrinkage Bayesian Causal Forest with Instrumental Variable (SBCF-IV)

Require: N units $\{(X_i, Z_i, W_i, Y_i)\}_{i=1}^{N}$.
Ensure: Tree-structured partition of the covariate space with node-level CACE estimates.

Step 1: Honest splitting.
1: Randomly partition the sample into $\mathcal{I}_{\text{disc}}$ and $\mathcal{I}_{\text{inf}}$ (defaults to half-size splits).

Step 2: Discovery (on $\mathcal{I}_{\text{disc}}$).
2: Estimate $\widehat{\text{ITT}}_Y^{\text{SBCF}}(x)$ via SBCF under the sparsity prior (3.3); save posterior split frequencies $\widehat{s}_{\text{ITT}_Y}$.
3: Estimate $\widehat{\pi}_C^{\text{SBCF}}(x)$ via a SoftBART probit with similar sparsity prior.
4: Form the pointwise heterogeneity signal $\widehat{\tau}^{\text{SBCF}}(x) = \widehat{\text{ITT}}_Y^{\text{SBCF}}(x)/\widehat{\pi}_C^{\text{SBCF}}(x)$.
5: Fit a shallow CART to $\{(\widehat{\tau}^{\text{SBCF}}(x), X_i)\}$ with variable-level costs c_{psp} as in (3.4); the resulting partition is $\{\mathbb{X}_j\}_j$.

Step 3: Inference (on $\mathcal{I}_{\text{inf}}$).
6: For every node $\mathbb{X}_j$, compute $\widehat{\tau}_{\mathbb{X}_j}^{\text{2SLS}}$ via (2.2), targeting $\tau_{\mathbb{X}_j}^{\text{CACE}}$.
7: Run first-stage weak-instrument tests and adjust leaf-level p-values for the familywise error rate.

8: **return** The pruned tree with node-level CACE estimates and adjusted inference.

with observed values $W_i = Z_i W_i(1)$ and Y_i defined analogously under SUTVA. The Gaussian error term reads $\epsilon_i \sim \mathcal{N}(0,1)$. Three features of (4.1) are central to the estimand of interest. First, $W_i(0) = 0$ imposes one-sided non-compliance, ruling out defiers and always-takers by design and making monotonicity (Assumption 2.1(e)) hold exactly. Second, we choose the compliance rate $\pi_{\text{comp}} = 0.75$ to control the strength of the instrument. Third, heterogeneity in the conditional CACE is confined to the first two binary covariates,

$$\tau^{\text{CACE}}(X_i) = \begin{cases} k, & X_i \in l_1 = \{X_{i,1} = 0,\ X_{i,2} = 0\}, \\ -k, & X_i \in l_2 = \{X_{i,1} = 1,\ X_{i,2} = 1\}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.2}$$

with effect size $k \in \{0, 0.2, 0.4, ..., 2\}$.[3] The remaining covariates, including all continuous ones, are irrelevant for $\tau^{\text{CACE}}(x)$, so the true sparsity level in the treatment effect component grows with P. This isolates the setting SBCF-IV is designed for: relevance concentrated on two covariates while $P - 2$ noise variables compete for splits. The control function $\mu(X_i)$ is adopted from Caron et al. (2022) and depends only on three continuous covariates through a nonlinear combination of sine, quadratic, and absolute-value terms. Its explicit form is given in Appendix E.1.

We evaluate SBCF-IV along three complementary dimensions: tree-level subgroup detection, unit-level classification, and unit-level estimation precision with uncertainty quantification. Tree-level recovery is captured by the Detection Rate (DR) and False Discovery Rate (FDR), which record whether the true heterogeneity regions l_1, l_2 are recovered. Unit-level classification is assessed on $\mathcal{I}_{\text{inf}}$ via Recall, Precision, False Positive Rate (FPR), and F-score, translating structural detection into how reliably individual observations are sorted into statistically significant leaves. Estimation precision is then quantified by the per-unit bias, MSE, and 95% coverage of the resulting conditional CACE estimates. Formal definitions and further discussions are deferred to Appendix E.2.

Figure 4.1 shows that SBCF-IV dominates BCF-IV on both tree-level criteria across all three covariate dimensions. For detection, SBCF-IV's DR rises steeply once $k \geq 0.8$ and saturates near one for $k \geq 1.4$, essentially uniformly in P: adding irrelevant covariates does not meaningfully degrade the algorithm's ability to recover l_1 and l_2 as leaves. BCF-IV, by contrast, only begins to detect the true subgroups for $k \geq 1$ in the low-dimensional case ($P = 10$), and its DR flattens near 0.3 at $k = 2$; for $P \in \{50, 100\}$, BCF-IV's DR stays close to zero throughout the grid, indicating that the uniform

[3] We use l_1, l_2 to denote the true heterogeneity subgroups in the DGP, distinct from the discovered subgroups $\{\mathbb{X}_j\}_j$ produced by the CART step of SBCF-IV. A successful run recovers l_1 and l_2 as leaves of the tree, i.e., $\mathbb{X}_j = l_j$ for $j \in \{1, 2\}$ up to labeling.

split-variable prior fails to concentrate on the two binary covariates that drive heterogeneity once they are buried among many noise variables. The ordering on FDR is equally clear. SBCF-IV maintains an FDR near zero across the entire (k, P) grid, so its detection gains do not come at the cost of spurious discoveries. BCF-IV, by contrast, exhibits an FDR that grows monotonically with k and with P, reaching roughly 0.75 at $k = 2$ with $P = 10$ and remaining above 0.25 for larger P. Most of BCF-IV's rare "discoveries" in high-dimensional settings are false positives. Together, the two panels indicate that the sparsity-inducing Dirichlet prior is doing what it is designed to do: it sharpens both sides of the discovery problem simultaneously, recovering true heterogeneity where BCF-IV misses it while controlling false flags where BCF-IV generates them.

Figure 4.2 confirms that SBCF-IV's tree-level advantages in Figure 4.1 translate directly into sharper unit-level sorting. Here, we report Precision and the F-score. The full set of classification metrics is reported in Appendix E. Precision follows the same pattern across both algorithms: for $k \geq 1$, SBCF-IV's Precision rises steeply and saturates near one by $k = 1.4$ across all three values of P, meaning that nearly every unit assigned to a significant leaf is in a truly heterogeneous subgroup. BCF-IV reaches roughly Precision $= 0.4$ at $k = 2$ under $P = 10$ and stays close to zero for $P \in \{50, 100\}$. In the high-dimensional regime, nearly all of BCF-IV's unit-level positive verdicts are misclassifications, which is a consequence of FDR results in Figure 4.1. The F-score mirrors this ranking and is the more informative single summary since it penalizes both missed heterogeneity and spurious flags. SBCF-IV's F-score converges to one for $k \geq 1.4$ uniformly in P, indicating that the algorithm both identifies the correct units and avoids misclassification. BCF-IV's F-score remains below 0.4 throughout the (k, P) grid and flatlines near zero once $P \geq 50$.

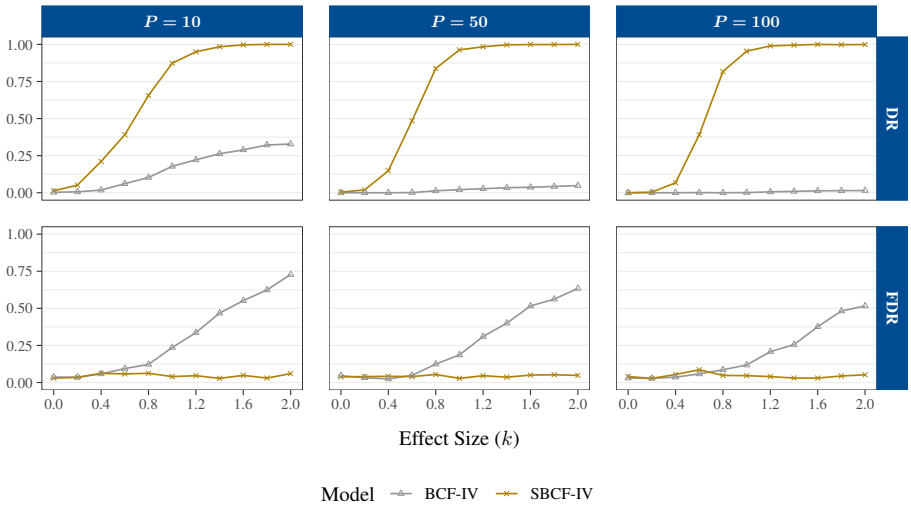

Figure 4.1: Tree-level subgroup detection as a function of effect size k, across covariate dimensions $P \in \{10, 50, 100\}$. The top row reports the Detection Rate (DR), defined in (E.3) as the average share of true heterogeneity subgroups l_1, l_2 recovered as leaves of the discovered tree. The bottom row reports the False Detection Rate (FDR), defined in (E.4) as the share of replications in which at least one spurious leaf is flagged as significant at $\alpha = 0.05$. Results are averaged over M Monte Carlo replications with $N = 1{,}000$, and compare SBCF-IV (orange) with BCF-IV (grey).

Table 1 reports estimation precision and uncertainty quantification for $\widehat{\tau}^{\text{CACE}}(x)$ and $k \in \{0, 1, 2\}$, while qualitatively similar results for the full k-range are deferred to Appendix E.4. Three patterns emerge. First, point estimation accuracy diverges sharply with effect size. For $k \geq 1$, SBCF-IV delivers consistently lower MSE and absolute bias than BCF-IV, with the gap widening as k grows: at $(k, P) = (2, 10)$, SBCF-IV's MSE is roughly a four-fold improvement; at $(k, P) = (2, 100)$, the contrast is a ten-fold improvement. BCF-IV's MSE explodes with P, while SBCF-IV's remains essentially flat, confirming that the dimension-invariance seen in Figures 4.1 and 4.2 extends to point estimation accuracy. Second, the absolute bias tells a complementary story: BCF-IV's absolute bias grows from 0.29 at $(k, P) = (0, 10)$ to 1.5 at $(k, P) = (2, 100)$, because its uniform split-variable prior dilutes the estimated effect across noise covariates. SBCF-IV's absolute bias stays below 0.5 throughout the grid. Third, uncertainty quantification degrades with increasing k and P for

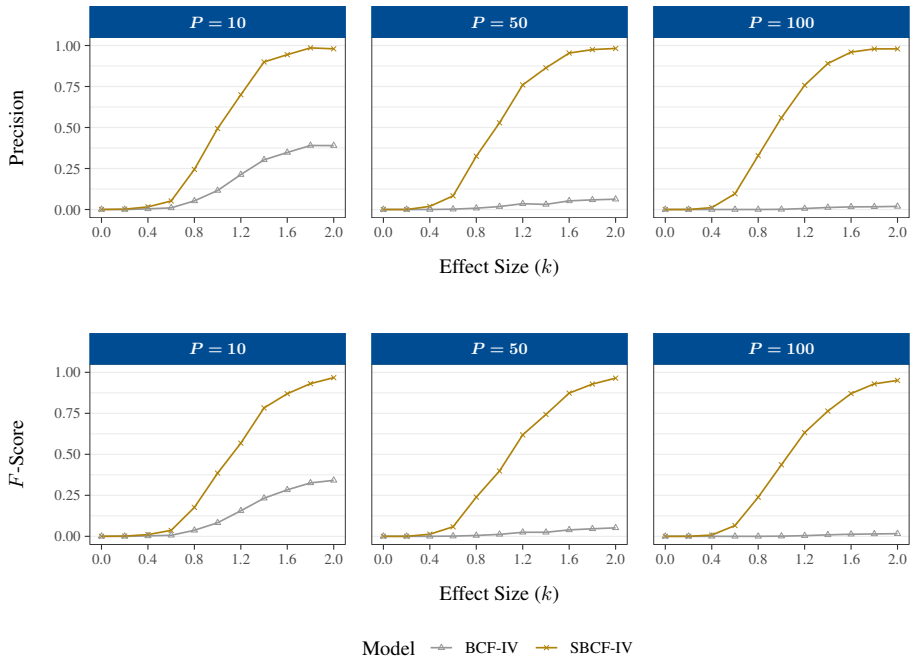

Figure 4.2: Unit-level classification performance as a function of effect size k, across covariate dimensions $P \in \{10, 50, 100\}$. The top row reports Precision and the bottom row the F-score, both defined in (E.5) and computed over all units in $\mathcal{I}_{\text{inf}}$ with significance evaluated at $\alpha = 0.05$ using Holm-adjusted p-values. Results are averaged over M Monte Carlo replications with $N = 1{,}000$, and compare SBCF-IV (orange) with BCF-IV (grey).

BCF-IV. SBCF-IV maintains coverage near the nominal 0.95 level across the entire (k, P) grid, so its confidence intervals remain honest even when the true effect is large or the covariate space is high-dimensional. BCF-IV's coverage, by contrast, drops sharply with k. Taken together, the three simulation dimensions point to a consistent conclusion: the sparsity prior improves discovery, classification, point estimation, and inference simultaneously, with SBCF-IV delivering tight and honest confidence intervals.

Table 1: Estimation precision and uncertainty quantification for $\widehat{\tau}^{\text{CACE}}(x)$ at $P \in \{10, 50, 100\}$. MSE, Bias, Absolute Bias, 95% CI Coverage, and CI Length, averaged over $M = 500$ Monte Carlo replications with $N = 1{,}000$. Results pool subgroups l_1 and l_2 and compare BCF-IV (left) with SBCF-IV (right). Standard deviations are reported in parentheses. We focus on $k \in \{0, 1, 2\}$ for visibility. Full results are in Appendix E.4. Note that standard deviations across Monte Carlo replications are reported in parentheses.

		BCF-IV					**SBCF-IV**				
P	k	MSE	Bias	Abs. Bias	Coverage	CI Length	MSE	Bias	Abs. Bias	Coverage	CI Length
10	0	0.188 (0.163)	-0.004 (0.225)	0.286 (0.140)	0.948 (0.176)	1.434 (0.219)	0.191 (0.161)	-0.004 (0.222)	0.314 (0.131)	0.957 (0.131)	1.600 (0.176)
	1	0.473 (0.333)	0.000 (0.286)	0.569 (0.240)	0.695 (0.337)	1.600 (0.210)	0.298 (0.327)	-0.002 (0.308)	0.426 (0.256)	0.899 (0.235)	1.794 (0.160)
	2	0.931 (0.729)	0.014 (0.446)	0.811 (0.354)	0.478 (0.335)	1.697 (0.174)	0.214 (0.211)	0.012 (0.337)	0.368 (0.199)	0.953 (0.149)	1.840 (0.141)
50	0	0.205 (0.244)	-0.003 (0.223)	0.273 (0.145)	0.953 (0.178)	1.399 (0.244)	0.214 (0.151)	-0.003 (0.222)	0.358 (0.140)	0.960 (0.111)	1.769 (0.143)
	1	0.785 (0.472)	0.012 (0.241)	0.744 (0.288)	0.454 (0.404)	1.495 (0.232)	0.213 (0.232)	-0.003 (0.309)	0.362 (0.203)	0.949 (0.151)	1.830 (0.140)
	2	2.088 (1.446)	-0.008 (0.299)	1.291 (0.508)	0.202 (0.256)	1.593 (0.255)	0.204 (0.203)	0.002 (0.320)	0.359 (0.192)	0.957 (0.145)	1.836 (0.138)
100	0	0.196 (0.234)	-0.020 (0.226)	0.253 (0.143)	0.957 (0.180)	1.325 (0.215)	0.217 (0.162)	-0.021 (0.226)	0.360 (0.143)	0.951 (0.111)	1.785 (0.136)
	1	0.954 (0.418)	0.019 (0.220)	0.872 (0.251)	0.273 (0.351)	1.417 (0.243)	0.230 (0.215)	0.005 (0.319)	0.387 (0.199)	0.950 (0.149)	1.826 (0.139)
	2	2.551 (1.522)	0.013 (0.257)	1.454 (0.526)	0.157 (0.222)	1.535 (0.235)	0.230 (0.238)	0.022 (0.340)	0.376 (0.212)	0.942 (0.164)	1.835 (0.146)

5 Empirical application

We revisit two empirical applications to apply SBCF-IV to real-world data. The Oregon Health Insurance Experiment (OHIE) was a landmark randomized controlled trial (RCT) conducted in 2008

to assess the effects of expanding Medicaid coverage on health outcomes, financial security, and healthcare utilization. The US state Oregon used a lottery system to allocate a limited number of Medicaid spots to uninsured, low-income adults (Finkelstein et al. 2012, Johnson et al. 2022). This created a natural experiment, allowing researchers to compare those who received Medicaid to those who did not. Following Johnson et al. (2022), we apply SBCF-IV to the OHIE data. SBCF-IV identifies a single complier subgroup with a positive, statistically significant Medicaid effect: English-preferring individuals aged 38–59 (effect: 2.2634; adjusted p-value: 0.0945), in contrast to the two subgroups reported by Johnson et al. (2022). Estimated subgroup compliance rates range from 58% to 67%, exhibiting only modest variation relative to the heterogeneity in conditional ITT effects. Both analyses thus locate the dominant source of complier-effect heterogeneity in the variation of the cCACE numerator rather than in compliance probabilities. Appendix F.1 re-investigates the empirical analysis presented in Johnson et al. (2022) in detail.

On the 401(k) data, SBCF-IV partitions eligible households into seven subgroups, of which two possess adjusted p-values at the 10% level: the bulk subgroup with income below \$68,800 (89% of the inference sample, $\widehat{\tau}^{\text{CACE}}(x) = \$17{,}818$) and a small upper-middle-income subgroup with household income between \$92,700 and \$110,400 with $\widehat{\tau}^{\text{CACE}}(x) = \$53{,}422$. The remaining five leaves show adjusted p-values without statistical significance, carry inference-sample shares close to 1%, and are best read as small-sample artifacts. The discovered splits run almost entirely along income, aligning with the lifecycle and earnings-gradient emphasis of the classical 401(k) saving literature (Poterba et al. 1992, Engen & Gale 2000). As in the OHIE application, complier shares are nearly uniform across leaves, so the heterogeneity is driven by variation in the conditional ITT numerator rather than in the complier denominator, mirroring the simulation evidence in Section 4 that SBCF-IV recovers ITT-driven partitions reliably. Both statistically significant leaf estimates exceed the overall cross-fitted CACE of $\approx \$12{,}000$ reported by Chernozhukov et al. (2018), the bulk leaf modestly (by \$5,000–\$9,000) and the upper-middle-income leaf substantially. We present this as a descriptive comparison only: the gap is consistent with several explanations (genuine income-based heterogeneity in the CACE, post-selection effects from reporting leaves after screening, and the upward bias from unobserved saver heterogeneity flagged by Engen et al. (1996)) which the present design cannot separate. We provide an extended discussion of this second empirical application in Appendix F.2.

6 Conclusion

This paper introduces SBCF-IV, an extension of BCF-IV (Bargagli-Stoffi et al. 2022) that combines Bayesian shrinkage with subgroup-level estimation of the conditional CACE under imperfect compliance in high-dimensional covariate settings. SBCF-IV preserves the interpretive structure of BCF-IV with an explicit partition over compliers and subgroup-level CACE while introducing regularization needed to make that structure stable in high dimensions (Caron et al. 2022). The simulation study in Section 4 demonstrates for varying effect sizes k that, as the covariate dimension P grows, SBCF-IV preserves subgroup discovery rates, precision, and nominal coverage, while BCF-IV exhibits a sharp degradation along most criteria leading to a corresponding loss of inferential reliability. Applied to two empirical applications, SBCF-IV recovers interpretable partitions of compliers whose cCACE estimates differ across subgroups, illustrating its practical value beyond the simulated regime. Several limitations warrant emphasis. The current formulation is restricted to binary instruments and binary treatments, and the honest sample split between $\mathcal{I}_{\text{disc}}$ and $\mathcal{I}_{\text{inf}}$ reduces the effective sample size that may hinder subgroup discovery in smaller empirical studies. Extensions to continuous instruments and multi-valued treatments combined with cross-fitting are natural directions for future work.

Acknowledgments and Disclosure of Funding

We thank Christoph Hanck for valuable comments that improved the paper. The paper has been presented at the Statistical Week 2025 and the ICSDS 2025. We thank all participants for valuable discussions that improved the focus of the paper.

References

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of Causal Effects Using Instrumental Variables', *Journal of the American Statistical Association* **91**(434), 444–455. _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476902.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902*

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360.
URL: *https://www.pnas.org/content/113/27/7353*

Athey, S., Tibshirani, J. & Wager, S. (2019), 'Generalized random forests', *Annals of Statistics* **47**(2), 1148–1178. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/euclid.aos/1547197251*

Bach, P., Kurz, M. S., Chernozhukov, V., Spindler, M. & Klaassen, S. (2024), 'DoubleML: An object-oriented implementation of double machine learning in R', *Journal of Statistical Software* **108**(3), 1–56.

Bargagli-Stoffi, F. J., Cadei, R., Lee, K. & Dominici, F. (2024), 'Causal Rule Ensemble: Interpretable Discovery and Inference of Heterogeneous Treatment Effects'. arXiv:2009.09036 [stat].
URL: *http://arxiv.org/abs/2009.09036*

Bargagli-Stoffi, F. J. & Gnecco, G. (2020), 'Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms', *International Journal of Data Science and Analytics* **9**(3), 315–337.
URL: *https://doi.org/10.1007/s41060-019-00187-z*

Bargagli-Stoffi, F. J., Witte, K. D. & Gnecco, G. (2022), 'Heterogeneous causal effects with imperfect compliance: A Bayesian machine learning approach', *The Annals of Applied Statistics* **16**(3), 1986–2009. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/journals/annals-of-applied-statistics/volume-16/issue-3/Heterogeneous-causal-effects-with-imperfect-compliance–A-Bayesian-machine/10.1214/21-AOAS1579.full*

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.
URL: *https://doi.org/10.1023/A:1010933404324*

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall/CRC, New York.

Caron, A., Baio, G. & Manolopoulou, I. (2022), 'Shrinkage Bayesian Causal Forests for Heterogeneous Treatment Effects Estimation', *Journal of Computational and Graphical Statistics* **31**(4), 1202–1214. _eprint: https://doi.org/10.1080/10618600.2022.2067549.
URL: *https://doi.org/10.1080/10618600.2022.2067549*

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.
URL: *https://academic.oup.com/ectj/article/21/1/C1/5056401*

Chetty, R., Friedman, J. N., Leth-Petersen, S., Nielsen, T. H. & Olsen, T. (2014), 'Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark *', *The Quarterly Journal of Economics* **129**(3), 1141–1219.
URL: *https://doi.org/10.1093/qje/qju013*

Chipman, H. A., George, E. I. & McCulloch, R. E. (2010), 'BART: Bayesian additive regression trees', *The Annals of Applied Statistics* **4**(1), 266–298. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-1/BART-Bayesian-additive-regression-trees/10.1214/09-AOAS285.full*

Deaton, A. (2010), 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature* **48**(2), 424–455.
URL: *https://www.aeaweb.org/articles?id=10.1257/jel.48.2.424*

Ding, P., Feller, A. & Miratrix, L. (2019), 'Decomposing Treatment Effect Variation', *Journal of the American Statistical Association* **114**(525), 304–317. _eprint: https://doi.org/10.1080/01621459.2017.1407322.
URL: *https://doi.org/10.1080/01621459.2017.1407322*

Dwivedi, R., Tan, Y. S., Park, B., Wei, M., Horgan, K., Madigan, D. & Yu, B. (2020), 'Stable Discovery of Interpretable Subgroups via Calibration in Causal Studies', *International Statistical Review* **88**(S1), S135–S178. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12427.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12427*

Engen, E. M. & Gale, W. G. (2000), 'The Effects of 401(K) Plans on Household Wealth: Differences Across Earnings Groups'.
URL: *https://papers.ssrn.com/abstract=252207*

Engen, E. M., Gale, W. G. & Scholz, J. K. (1996), 'The Illusory Effects of Saving Incentives on Saving', *Journal of Economic Perspectives* **10**(4), 113–138.
URL: *https://pubs.aeaweb.org/doi/10.1257/jep.10.4.113*

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. & Oregon Health Study Group (2012), 'The Oregon Health Insurance Experiment: Evidence from the First Year*', *The Quarterly Journal of Economics* **127**(3), 1057–1106.
URL: *https://doi.org/10.1093/qje/qjs020*

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2020), 'Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)', *Bayesian Analysis* **15**(3), 965–1056. Publisher: International Society for Bayesian Analysis.
URL: *https://projecteuclid.org/journals/bayesian-analysis/volume-15/issue-3/Bayesian-Regression-Tree-Models-for-Causal-Inference–Regularization-Confounding/10.1214/19-BA1195.full*

Hartford, J., Lewis, G., Leyton-Brown, K. & Taddy, M. (2017), Deep IV: A Flexible Approach for Counterfactual Prediction, *in* 'Proceedings of the 34th International Conference on Machine Learning', PMLR, pp. 1414–1423. ISSN: 2640-3498.
URL: *https://proceedings.mlr.press/v70/hartford17a.html*

Hill, J. L. (2011), 'Bayesian Nonparametric Modeling for Causal Inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240. Publisher: ASA Website _eprint: https://doi.org/10.1198/jcgs.2010.08162.
URL: *https://doi.org/10.1198/jcgs.2010.08162*

Holm, S. (1979), 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics* **6**(2), 65–70.
URL: *https://www.jstor.org/stable/4615733*

Huang, M., Tang, T. M. & Kenney, A. M. (2025), 'Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference'. arXiv:2502.07275 [stat].
URL: *http://arxiv.org/abs/2502.07275*

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and Estimation of Local Average Treatment Effects', *Econometrica* **62**(2), 467–475.
URL: *https://www.jstor.org/stable/2951620*

Imbens, G. W. & Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, Cambridge.
URL: *https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB*

Johnson, M., Cao, J. & Kang, H. (2022), 'Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment', *The Annals of Applied Statistics* **16**(2), 1111–1129. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/journals/annals-of-applied-statistics/volume-16/issue-2/Detecting-heterogeneous-treatment-effects-with-instrumental-variables-and-application-to/10.1214/21-AOAS1535.full*

Kennedy, E. H. (2023), 'Towards optimal doubly robust estimation of heterogeneous causal effects', *Electronic Journal of Statistics* **17**(2), 3008–3049.
URL: *https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-17/issue-2/Towards-optimal-doubly-robust-estimation-of-heterogeneous-causal-effects/10.1214/23-EJS2157.full*

Kennedy, E. H., Balakrishnan, S. & G'Sell, M. (2020), 'Sharp instruments for classifying compliers and generalizing causal effects', *The Annals of Statistics* **48**(4).
URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-4/Sharp-instruments-for-classifying-compliers-and-generalizing-causal-effects/10.1214/19-AOS1874.full*

Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. (2019), 'Metalearners for estimating heterogeneous treatment effects using machine learning', *Proceedings of the National Academy of Sciences* **116**(10), 4156–4165.
URL: *https://www.pnas.org/doi/10.1073/pnas.1804597116*

Lee, K., Small, D. S. & Dominici, F. (2021), 'Discovering Heterogeneous Exposure Effects Using Randomization Inference in Air Pollution Studies', *Journal of the American Statistical Association* **116**(534), 569–580. _eprint: https://doi.org/10.1080/01621459.2020.1870476.
URL: *https://doi.org/10.1080/01621459.2020.1870476*

Linero, A. R. (2018), 'Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection', *Journal of the American Statistical Association* **113**(522), 626–636. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1264957.
URL: *https://doi.org/10.1080/01621459.2016.1264957*

Linero, A. R. & Yang, Y. (2018), 'Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80**(5), 1087–1110.
URL: *https://academic.oup.com/jrsssb/article/80/5/1087/7048381*

Lipkovich, I., Dmitrienko, A. & B. D'Agostino Sr., R. (2017), 'Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials', *Statistics in Medicine* **36**(1), 136–196. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7064.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7064*

Poterba, J. M., Venti, S. F. & Wise, D. A. (1992), '401(K) Plans and Tax-Deferred Saving'.
URL: *https://papers.ssrn.com/abstract=1720805*

Poterba, J. M., Venti, S. F. & Wise, D. A. (1995), 'Do 401(k) contributions crowd out other personal saving?', *Journal of Public Economics* **58**(1), 1–32.
URL: *https://linkinghub.elsevier.com/retrieve/pii/004727279401462W*

R Core Team (2024), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Semenova, V. & Chernozhukov, V. (2021), 'Debiased machine learning of conditional average treatment effects and other causal functions', *The Econometrics Journal* **24**(2), 264–289.
URL: *https://doi.org/10.1093/ectj/utaa027*

Swanson, S. A. & Hernán, M. A. (2014), 'Think Globally, Act Globally: An Epidemiologist's Perspective on Instrumental Variable Estimation', *Statistical Science* **29**(3), 371–374.
URL: *https://projecteuclid.org/journals/statistical-science/volume-29/issue-3/Think-Globally-Act-Globally–An-Epidemiologists-Perspective-on-Instrumental/10.1214/14-STS491.full*

Therneau, T. & Atkinson, B. (2026), *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.26.
URL: *https://github.com/bethatkinson/rpart*

Wang, G., Li, J. & Hopp, W. J. (2022), 'An Instrumental Variable Forest Approach for Detecting Heterogeneous Treatment Effects in Observational Studies', *Management Science* **68**(5), 3399–3418.
URL: *https://ideas.repec.org//a/inm/ormnsc/v68y2022i5p3399-3418.html*

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, MA.

A Identification of the conditional CACE

This appendix proves Proposition 2.1: under Assumption 2.1, the complier estimand $\tau^{\text{CACE}}(x)$ of Definition 2.2 equals the Wald ratio $\text{ITT}_Y(x)/\pi_C(x)$ and is therefore identified from the observed distribution of (Y_i, W_i, Z_i, X_i). The argument adapts Angrist et al. (1996) to the conditional target as in Bargagli-Stoffi et al. (2022).[4]

For brevity, contrast this irregular assignment setting with the regular assignment mechanism, where unconfoundedness $W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid X_i$ and overlap $0 < \Pr(W_i = 1 \mid X_i = x) < 1$ identify the conditional average treatment effect $\tau(x) = \mathbb{E}[Y_i \mid W_i = 1, X_i = x] - \mathbb{E}[Y_i \mid W_i = 0, X_i = x]$ directly from observed conditional means (Imbens & Rubin 2015). Under irregular assignment, W_i is potentially confounded and direct identification of $\tau(x)$ fails. Identification of the complier-restricted target $\tau^{\text{CACE}}(x)$ uses the IV approach below.

Fix x in the support of X and define the observable nuisance functions

$$\begin{aligned} \mu_z(x) &\coloneqq \mathbb{E}[Y_i \mid Z_i = z, X_i = x], \\ \delta_z(x) &\coloneqq \mathbb{E}[W_i \mid Z_i = z, X_i = x], \\ & z \in \{0, 1\}. \end{aligned}$$

Both are identified from the observed distribution. By Definition 2.3, $\text{ITT}_Y(x) = \mu_1(x) - \mu_0(x)$, and we will show in Step 4 that $\pi_C(x) = \delta_1(x) - \delta_0(x)$ under monotonicity. The remaining task is to establish

$$\mu_1(x) - \mu_0(x) \;=\; \tau^{\text{CACE}}(x)\,\big(\delta_1(x) - \delta_0(x)\big). \tag{A.1}$$

Step 1: Express $\mu_z(x)$ in potential outcomes. By Assumption 2.1(a) (consistency) and (c) (unconfounded instrument),

$$\mu_z(x) \;=\; \mathbb{E}[Y_i(z, W_i(z)) \mid Z_i = z, X_i = x] \;=\; \mathbb{E}[Y_i(z, W_i(z)) \mid X_i = x].$$

Hence

$$\begin{aligned} \mu_1(x) - \mu_0(x) &= \mathbb{E}\big[Y_i(1, W_i(1)) - Y_i(0, W_i(0)) \;\big|\; X_i = x\big] \\ &\overset{\text{(d)}}{=} \mathbb{E}\big[Y_i(W_i(1)) - Y_i(W_i(0)) \;\big|\; X_i = x\big], \end{aligned} \tag{A.2}$$

where the second equality invokes the exclusion restriction.

Step 2: Apply the binary-W algebraic identity. Since $W_i(z) \in \{0, 1\}$, a case analysis over the four values of $(W_i(0), W_i(1))$ yields

$$Y_i(W_i(1)) - Y_i(W_i(0)) \;=\; \big[Y_i(1) - Y_i(0)\big]\big[W_i(1) - W_i(0)\big], \tag{A.3}$$

which requires no identifying assumption beyond binarity of W_i. Substituting (A.3) into (A.2),

$$\mu_1(x) - \mu_0(x) \;=\; \mathbb{E}\big[\{Y_i(1) - Y_i(0)\}\{W_i(1) - W_i(0)\} \;\big|\; X_i = x\big]. \tag{A.4}$$

[4]Within this appendix, $\mu_z(x)$ denotes an observable conditional mean and is distinct from the BART control function $\mu(\cdot)$ of the main text.

Step 3: Apply monotonicity. By Assumption 2.1(e), $W_i(1)-W_i(0) \in \{0,1\}$, so $W_i(1)-W_i(0) = \mathbf{1}\{G_i = C\}$. Combining with (A.4) and iterating the expectation,

$$\mu_1(x) - \mu_0(x) \;=\; \mathbb{E}\big[Y_i(1) - Y_i(0) \;\big|\; G_i = C,\, X_i = x\big] \cdot \pi_C(x) \;=\; \tau^{\text{CACE}}(x) \cdot \pi_C(x), \quad \text{(A.5)}$$

where the second equality uses Definition 2.2 together with (d) (so that $Y_i(w)$ is unambiguous on the complier subpopulation).

Step 4: Identify $\pi_C(x)$ from observables. By consistency (a) and unconfoundedness (c), $\delta_z(x) = \mathbb{E}[W_i(z) \mid X_i = x]$. Under monotonicity (e),

$$\delta_1(x) - \delta_0(x) \;=\; \mathbb{E}[W_i(1) - W_i(0) \mid X_i = x] \;=\; \Pr(G_i = C \mid X_i = x) \;=\; \pi_C(x), \quad \text{(A.6)}$$

and Assumption 2.1(b) ensures $\pi_C(x) > 0$ almost surely, so the ratio in (A.1) is well-defined.

Conclusion. Combining (A.5) and (A.6) yields (A.1), hence

$$\tau^{\text{CACE}}(x) \;=\; \frac{\mu_1(x) - \mu_0(x)}{\delta_1(x) - \delta_0(x)} \;=\; \frac{\text{ITT}_Y(x)}{\pi_C(x)}, \quad \text{(A.7)}$$

which is identified from the observed distribution of (Y_i, W_i, Z_i, X_i). □

B Conditional 2SLS: reduced form and asymptotic properties

This appendix collects the reduced form of the subgroup-restricted simultaneous system (2.2) and the asymptotic properties of the 2SLS estimator $\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j}$ defined in Definition 2.4. The results restate the unconditional 2SLS theory of Wooldridge (2010) with identification of the unconditional CACE following Angrist et al. (1996) and Imbens & Rubin (2015), transferred to the conditional target $\tau^{\text{CACE}}_{\mathbb{X}_j}$ (Bargagli-Stoffi et al. 2022).

Reduced form. Substituting the first-stage equation of (2.2) into the outcome equation gives

$$Y_{i,\mathbb{X}_j} = \big(\alpha_{\mathbb{X}_j} + \tau^{\text{CACE}}_{\mathbb{X}_j}\,\pi_{0,\mathbb{X}_j}\big) + \underbrace{\big(\tau^{\text{CACE}}_{\mathbb{X}_j}\,\pi_{C,\mathbb{X}_j}\big)}_{\gamma_{\mathbb{X}_j}} Z_{i,\mathbb{X}_j} + \big(\varepsilon_{i,\mathbb{X}_j} + \tau^{\text{CACE}}_{\mathbb{X}_j}\,\eta_{i,\mathbb{X}_j}\big).$$

Under $\mathbb{E}[\varepsilon_{i,\mathbb{X}_j}] = \mathbb{E}[\eta_{i,\mathbb{X}_j}] = 0$ and $\mathbb{E}[Z_{i,\mathbb{X}_j}\,\eta_{i,\mathbb{X}_j}] = 0$, OLS consistently estimates $\pi_{C,\mathbb{X}_j}$ and $\gamma_{\mathbb{X}_j}$, so

$$\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j} \;=\; \frac{\widehat{\gamma}_{\mathbb{X}_j}}{\widehat{\pi}_{C,\mathbb{X}_j}} \quad \text{(B.1)}$$

coincides with the moment-based ratio of Definition 2.4 (Wooldridge 2010).

Asymptotic properties. Let $N_{\mathbb{X}_j}$ denote the number of observations in $\mathbb{X}_j$, and consider the subgroup-restricted moment conditions

(A1) $\mathbb{E}[Z^2_{i,\mathbb{X}_j}] \neq 0$;

(A2) $\mathbb{E}[Z_{i,\mathbb{X}_j}\,\varepsilon_{i,\mathbb{X}_j}] = 0$;

(A3) $\pi_{C,\mathbb{X}_j} \neq 0$;

(A4) $\mathbb{E}[Z^2_{i,\mathbb{X}_j}\,\varepsilon^2_{i,\mathbb{X}_j}] < \infty$.

Conditions (A1)–(A3) are the subgroup analogues of the standard IV moment restrictions and are implied within $\mathbb{X}_j$ by Assumption 2.1 together with the exogeneity of Z_i.

Theorem B.1 (Consistency)**.** *Under (A1)–(A3), $\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j} - \tau^{\text{CACE}}_{\mathbb{X}_j} \xrightarrow{p} 0$ as $N_{\mathbb{X}_j} \to \infty$.*

Theorem B.2 (Asymptotic normality)**.** *Under (A1)–(A4),*

$$\sqrt{N_{\mathbb{X}_j}}\,\big(\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j} - \tau^{\text{CACE}}_{\mathbb{X}_j}\big) \;\xrightarrow{d}\; \mathcal{N}\big(0,\; N_{\mathbb{X}_j} \cdot \text{avar}(\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j})\big) \qquad \text{as } N_{\mathbb{X}_j} \to \infty,$$

with $\text{avar}(\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j})$ *the asymptotic variance of the 2SLS estimator, approximated as in Wooldridge (2010).*

Proof sketch. Both results follow from the unconditional 2SLS case applied to the i.i.d. subsample $\{(Y_l, W_l, Z_l) : X_l \in \mathbb{X}_j\}$ (Wooldridge 2010, Bargagli-Stoffi et al. 2022). Consistency uses the continuous mapping theorem applied to (B.1) together with the LLN for the numerator and denominator; normality follows from the delta method combined with a CLT for $(\widehat{\gamma}_{\mathbb{X}_j}, \widehat{\pi}_{C,\mathbb{X}_j})$, where (A4) ensures a finite asymptotic variance. □

Remark. Assumption (A3) is the subgroup-level relevance condition and need not follow from overall instrument strength; as noted in Bargagli-Stoffi et al. (2022) and Ding et al. (2019), the approximations in Theorems B.1–B.2 require a sufficient number of observations within each $\mathbb{X}_j$. Shallower trees produce subgroups with both higher interpretability and higher statistical power (Athey & Imbens 2016, Lee et al. 2021), motivating the trimming of small or weak-instrument nodes in SBCF-IV.

C Background: BART, BCF, SBCF, and CART

This appendix recalls the ingredients on which SBCF-IV is built. We keep the exposition minimal; full treatments are in the cited references.

Classification and Regression Tree (CART). The CART algorithm (Breiman et al. 1984) recursively partitions the covariate space along axis-aligned splits, choosing at each node the split that maximizes within-node homogeneity of the response. The terminal-node piecewise-constant predictor is interpretable but a high-variance single learner.

Bayesian Additive Regression Trees (BART). BART (Chipman et al. 2010) models the conditional mean as a sum of regression trees, and place three regularizing priors: one penalizing tree depth, one shrinking leaf-level predictions toward the response center, and one bounding the error variance away from zero. Jointly, these priors prevent any single tree from dominating the fit and yield coherent posterior uncertainty. At each split, BART draws the candidate splitting variable according to a vector of selection probabilities $s = (s_1, \ldots, s_P)$, which by default is uniform, $s_j = 1/P$.

Bayesian Causal Forest (BCF). The BCF of Hahn et al. (2020) adapts BART to causal inference under a regular assignment mechanism. Two features carry over to our setting: (i) the conditional mean of the outcome is decomposed into a control function and a treatment effect function, each assigned an independent BART prior; (ii) the propensity score is included as a covariate in the control function to mitigate regularization-induced confounding and targeted selection. Different depth-penalty parameters across the two components encourage shallower, more interpretable trees for the treatment effect component. Equation (3.1) is the IV analogue of this decomposition, with Z_i and $\mathrm{ITT}_Y(x)$ replacing the treatment indicator and treatment effect, respectively.

Shrinkage Bayesian Causal Forest (SBCF). SBCF (Caron et al. 2022) augments BCF with a sparsity-inducing Dirichlet prior on the split-variable selection probabilities, replacing BART's default uniform $s_j = 1/P$ with $s \sim \mathrm{Dir}(\alpha/P, \ldots, \alpha/P)$ and a hyperprior $\alpha/(\alpha+\rho) \sim \mathrm{Beta}(a, b)$ on the concentration parameter. Small α concentrates posterior mass on a few covariates; the hyperprior lets the data determine how aggressive this concentration should be, with the default $(a, b, \rho) = (0.5, 1, P)$ strengthening the sparsity preference as P grows. The prior acts at each split independently, so covariates with low posterior inclusion probability are rarely chosen throughout the ensemble. Caron et al. (2022) show that SBCF improves CATE estimation over BCF when many covariates are irrelevant and in confounded-data settings, and that the posterior split frequencies $\widehat{s}$ provide a natural variable-importance summary. SBCF-IV inherits both the estimation benefit and the interpretability of $\widehat{s}$: the latter is exploited in (3.4) to steer the subgroup-finding CART toward the covariates the ensemble deemed most relevant.

D Prior specifications for SBCF-IV

SBCF-IV combines the BCF decomposition in (3.1)–(3.2) with the sparsity-inducing Dirichlet prior on split-variable selection probabilities proposed by Caron et al. (2022). The main text states the prior generically in (3.3); here we record the component-specific hyperparameter choices.

Let $s_\bullet = (s_{\bullet,1}, \ldots, s_{\bullet,P_\bullet})$ denote the split-probability vector for each of the three functions $\mu(e(x), x)$, $\text{ITT}_Y(x)$, and $\delta(z, x)$ in (3.1)–(3.2). We place independent priors

$$s_\mu \sim \text{Dir}\Big(\tfrac{\alpha_\mu}{P+1}, \ldots, \tfrac{\alpha_\mu}{P+1}\Big), \qquad \tfrac{\alpha_\mu}{\alpha_\mu+\rho_\mu} \sim \text{Beta}(a, b), \qquad \rho_\mu = P + 1, \tag{D.1}$$

$$s_{\text{ITT}_Y} \sim \text{Dir}\big(\tfrac{\alpha_{\text{ITT}_Y}}{P}, \ldots, \tfrac{\alpha_{\text{ITT}_Y}}{P}\big), \qquad \tfrac{\alpha_{\text{ITT}_Y}}{\alpha_{\text{ITT}_Y}+\rho_{\text{ITT}_Y}} \sim \text{Beta}(a, b), \qquad \rho_{\text{ITT}_Y} = P/2, \tag{D.2}$$

$$s_\delta \sim \text{Dir}\Big(\tfrac{\alpha_\delta}{P+1}, \ldots, \tfrac{\alpha_\delta}{P+1}\Big), \qquad \tfrac{\alpha_\delta}{\alpha_\delta+\rho_\delta} \sim \text{Beta}(a, b), \qquad \rho_\delta = P + 1, \tag{D.3}$$

with shared defaults $(a, b) = (0.5, 1)$ throughout. Three comments are in order.

(i) Dimension of s_μ. The propensity score $e(x)$ enters μ as an additional covariate, so $P_\mu = P + 1$; the extra dimension is reflected both in the Dirichlet parameter and in the scale ρ_μ.

(ii) Targeted shrinkage via ρ. Under $(a, b) = (0.5, 1)$, the hyperprior on $\alpha/(\alpha + \rho)$ concentrates near zero, so $\alpha \approx \rho \cdot t/(1 - t)$ for small t; smaller ρ thus pulls α toward zero and strengthens the Dirichlet's concentration on few covariates. Setting $\rho_{\text{ITT}_Y} = P/2 < P + 1 = \rho_\mu$ imposes a sharper sparsity preference on the treatment effect component than on the control function, consistent with the premise that heterogeneity is driven by a smaller subset of covariates than the outcome's main effects. The case $(a, b) = (1, 1)$ with $\alpha \to \infty$ recovers the uniform $s_{\bullet,j} = 1/P_\bullet$ used by BCF-IV.

(iii) Unified treatment of Y_i and W_i. The instrument z enters δ as an additional splittable covariate, so $P_\delta = P + 1$; the extra dimension is reflected in the Dirichlet parameter and in the scale ρ_δ, paralleling the treatment of $e(x)$ in μ. Specification (D.3) thus applies the same Dirichlet-based sparsity mechanism to the compliance model, implemented as a SoftBART probit in the sense of Hill (2011). No separate justification is needed: the sparsity concern (only a subset of covariates drives either outcome heterogeneity or differential compliance) is identical in both components, and the Dirichlet prior is the natural vehicle for it in both cases. Because $\delta(z, x)$ is not further decomposed as in (3.1), a single s_δ suffices rather than a pair of component-specific priors.

E Supplementary materials for the simulation study

E.1 Full simulation design

This appendix collects the full specification of the data-generating process summarized in Section 4. For each unit $i = 1, \ldots, N$ with $N = 1{,}000$, potential outcomes, potential treatments, and covariates are drawn independently as

$$\begin{aligned}
Z_i &\sim \text{Bin}(0.5), \\
W_i(1) &\sim \text{Bin}(\pi_{\text{comp}} = 0.75), \qquad W_i(0) = 0, \\
W_i &= Z_i\, W_i(1) + (1 - Z_i)\, W_i(0), \\
Y_i(0) &= \mu(X_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1), \\
Y_i(1) &= Y_i(0) + W_i(1)\, \tau^{\text{CACE}}(X_i),
\end{aligned}$$

with observed outcome $Y_i = Z_i\, Y_i(1) + (1 - Z_i)\, Y_i(0)$ under SUTVA (Assumption 2.1(a)).

Covariates. The covariate vector $X_i = (X_{i,1}, \ldots, X_{i,P}) \in \mathbb{R}^P$ consists of $P/2$ binary and $P/2$ continuous independent components,

$$X_{i,1}, \ldots, X_{i,P/2} \overset{\text{i.i.d.}}{\sim} \text{Bin}(0.5), \qquad X_{i,P/2+1}, \ldots, X_{i,P} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),$$

with $P \in \{10, 50, 100\}$. The three values of P probe increasing degrees of sparsity: only $X_{i,1}$ and $X_{i,2}$ drive heterogeneity in $\tau^{\text{CACE}}(x)$, so the share of relevant covariates in the treatment effect component is $2/P \in \{0.2, 0.04, 0.02\}$.

Compliance and instrument strength. The compliance rate $\pi_{\text{comp}} = 0.75$ governs the strength of Z_i as an instrument for W_i. One-sided non-compliance is built in via $W_i(0) = 0$: units not assigned to treatment cannot receive it, so the population decomposes into compliers ($W_i(1) = 1$) and never-takers ($W_i(1) = 0$), with $\pi_C = \pi_{\text{comp}}$ and $\pi_{NT} = 1 - \pi_{\text{comp}}$. Defiers and always-takers are ruled out by design, and the monotonicity assumption 2.1(e) holds with equality for the never-takers and strictly for the compliers.

Heterogeneity in the CACE. The conditional CACE is piecewise constant over three regions of the binary covariate space,

$$\tau^{\text{CACE}}(X_i) = \begin{cases} k, & X_i \in l_1 = \{X_{i,1} = 0,\ X_{i,2} = 0\}, \\ -k, & X_i \in l_2 = \{X_{i,1} = 1,\ X_{i,2} = 1\}, \\ 0, & X_i \in l_0 = \{X_i \notin l_1 \cup l_2\}, \end{cases} \tag{E.1}$$

with effect size $k \in \{0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2\}$ to distinguish no, moderate, and large heterogeneity regimes. The two non-null subgroups l_1 and l_2 each have population mass 0.25 under the Bernoulli distribution of $(X_{i,1}, X_{i,2})$, and carry opposite-signed effects, so the marginal CACE averages to zero. This design makes a naive marginal estimator uninformative and forces any algorithm targeting $\tau^{\text{CACE}}(x)$ to recover the heterogeneity structure. The subgroups l_1, l_2, l_0 are ground-truth objects defined by the DGP; they are to be distinguished from the data-driven partition $\{\mathbb{X}_j\}_j$ produced by SBCF-IV on $\mathcal{I}_{\text{disc}}$, whose leaves approximate l_1 and l_2 when discovery succeeds.

Control function. The baseline outcome function $\mu(X_i)$, adopted from Caron et al. (2022), depends only on the first three continuous covariates and takes the explicit form

$$\mu(X_i) \ = \ 3 + 1.5 \sin\big(\pi\, X_{i,P/2+1}\big) + 0.5\, \big(X_{i,P/2+2} - 0.5\big)^2 + 1.5\, \big(2 - |X_{i,P/2+3}|\big). \tag{E.2}$$

The three terms introduce nonlinearity, curvature, and non-differentiability into the control component, providing a realistic stress test for the control-function BART prior on $\mu(e(x), x)$ in (3.1). The relevant covariates for $\mu(X_i)$ in Equation (E.2) and for $\tau^{\text{CACE}}(\cdot)$ in Equation (E.1) are disjoint (continuous covariates for the former, binary for the latter) which lets us assess the discovery ability of SBCF-IV cleanly, without confounding signal in the treatment effect component with signal in the control function.

Design grid. The full Monte Carlo design varies (P, k) on the grid $\{10, 50, 100\} \times \{0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2\}$, yielding 33 scenarios. Each scenario is replicated $M = 500$ times, and performance is summarized via the classification metrics (Recall, Precision, F-score, FPR) for correctly identifying l_1 and l_2, together with MSE, bias, and coverage for the estimated conditional CACE.

E.2 Performance measures for the simulation study

Our evaluation framework follows the spirit of Bargagli-Stoffi et al. (2022), but refines it by separating structural recovery from inferential significance: tree-level metrics benchmark structural discovery in simulation, while unit-level metrics provide the evaluation criteria relevant for applications where the true partition is unobserved. Thus, we evaluate SBCF-IV and BCF-IV along three dimensions: (i) tree-level subgroup detection, measuring whether the true heterogeneity regions l_1, l_2 appear as leaves of the discovered tree; (ii) unit-level individual classification, measuring whether each observation in $\mathcal{I}_{\text{inf}}$ is assigned to a correctly signed leaf; and (iii) unit-level estimation precision of the resulting CACE estimates. All metrics are computed on $\mathcal{I}_{\text{inf}}$ in each simulation run $m = 1, \ldots, M$ and averaged across runs. Throughout, $\mathcal{T}_m$ denotes the tree with corresponding subgroups discovered in run m, with leaves $\{\mathbb{X}_j(\mathcal{T}_m)\}_{j=1}^{J}$, and $p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}$ denotes the Holm-adjusted p-value of the estimated CACE within the discovered leaf $\mathbb{X}_j(\mathcal{T}_m)$. We set $\alpha = 0.05$.

E.2.1 Subgroup detection: DR and FDR

At the tree level, we assess whether the true heterogeneous subgroups l_1, l_2 from (4.2) are recovered as leaves of $\mathcal{T}_m$. The Detection Rate (DR) is the average share of true subgroups recovered, computed without a significance requirement to isolate the structural recovery ability of the algorithm from inferential uncertainty. We report DR both overall and separately for l_1 and l_2. The False Discovery Rate (FDR) is the share of replications in which at least one spurious leaf (a leaf not corresponding to a truly heterogeneous subgroup) is flagged as significant. FDR is evaluated at the tree level rather than normalized by the number of non-effect subgroups, since the latter is a random quantity that

varies across replications and therefore cannot serve as a stable denominator.[5] Formally,

$$\text{DR} = \frac{1}{M}\sum_{m=1}^{M}\frac{\sum_{j=1}^{J}\mathbf{1}\{\mathbb{X}_j(\mathcal{T}_m)\in\{l_1,l_2\}\}}{\#\{l_1,l_2\}}, \tag{E.3}$$

$$\text{FDR} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left\{\sum_{j=1}^{J}\mathbf{1}\big\{\mathbb{X}_j(\mathcal{T}_m)\notin\{l_1,l_2\},\ p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}\le\alpha\big\}\ge 1\right\}. \tag{E.4}$$

Because l_1, l_2 are only observable in simulation, DR and FDR serve as benchmarking tools; for empirical applications the unit-level classification and estimation metrics below are the relevant ones.

E.2.2 Individual classification: Recall, Precision, FPR, F-score

At the unit level, we ask whether each observation in $\mathcal{I}_{\text{inf}}$ is assigned to a leaf whose significance verdict matches its true heterogeneity status. Let $g_i \in \{l_0, l_1, l_2\}$ denote the true subgroup of unit i under (4.2), and let $p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}$ denote the adjusted p-value of the leaf $\mathbb{X}_j(\mathcal{T}_m)$ containing i. Each unit contributes to one of four classification cells:

$$\begin{aligned}
&\text{TP}(\mathcal{T}_m) = \#\big\{i : g_i\in\{l_1,l_2\},\ p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}\le\alpha\big\}, \quad &&\text{FN}(\mathcal{T}_m) = \#\big\{i : g_i\in\{l_1,l_2\},\ p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}>\alpha\big\},\\
&\text{FP}(\mathcal{T}_m) = \#\big\{i : g_i = l_0,\ p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}\le\alpha\big\}, &&\text{TN}(\mathcal{T}_m) = \#\big\{i : g_i = l_0,\ p^{\text{adj}}_{\mathbb{X}_j(\mathcal{T}_m)}>\alpha\big\}.
\end{aligned}$$

From these we compute the standard classification metrics:

$$\begin{aligned}
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}},\\
\text{FPR} &= \frac{\text{FP}}{\text{FP}+\text{TN}},\\
F\text{-score} &= \frac{\text{TP}}{\text{TP}+\frac{1}{2}(\text{FP}+\text{FN})}.
\end{aligned} \tag{E.5}$$

Recall measures the share of truly heterogeneous units correctly flagged; Precision the share of flagged units that are genuinely heterogeneous; FPR the share of null units spuriously flagged; and the F-score their harmonic mean. All four are averaged across the M simulation runs.

E.2.3 Estimation precision: bias, MSE, and coverage

For each simulation m, we evaluate the per-replication bias, MSE, and 95% coverage on the truly heterogeneous regions $l_1 \cup l_2$ within $\mathcal{I}_{\text{inf}}$, denoted $\mathcal{I}^{\text{het}}_{\text{inf}} = \{i \in \mathcal{I}_{\text{inf}} : X_i \in l_1 \cup l_2\}$ with cardinality $N^{\text{het}}_{\text{inf}}$:

$$\begin{aligned}
|\text{Bias}|_m\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big) &= \frac{1}{N^{\text{het}}_{\text{inf}}}\sum_{i\in\mathcal{I}^{\text{het}}_{\text{inf}}}\big|\tau^{\text{CACE}}(X_i)-\widehat{\tau}^{\text{CACE}}(X_i)\big|,\\
\text{Bias}_m\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big) &= \frac{1}{N^{\text{het}}_{\text{inf}}}\sum_{i\in\mathcal{I}^{\text{het}}_{\text{inf}}}\big(\tau^{\text{CACE}}(X_i)-\widehat{\tau}^{\text{CACE}}(X_i)\big),\\
\text{MSE}_m\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big) &= \frac{1}{N^{\text{het}}_{\text{inf}}}\sum_{i\in\mathcal{I}^{\text{het}}_{\text{inf}}}\big(\tau^{\text{CACE}}(X_i)-\widehat{\tau}^{\text{CACE}}(X_i)\big)^2,\\
\text{Coverage}_m\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big) &= \frac{1}{N^{\text{het}}_{\text{inf}}}\sum_{i\in\mathcal{I}^{\text{het}}_{\text{inf}}}\mathbf{1}\Big\{\tau^{\text{CACE}}(X_i)\in\widehat{\text{CI}}_{95}\big(\widehat{\tau}^{\text{CACE}}(X_i)\big)\Big\},
\end{aligned} \tag{E.6}$$

with Monte Carlo averages $\text{Bias}\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big) = \frac{1}{M}\sum_{m=1}^{M}\text{Bias}_m\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big)$, and analogously for $|\text{Bias}|_{\text{m}}$, MSE_m and Coverage_m. Under correct coverage, $\text{Coverage}\big(\mathcal{I}^{\text{het}}_{\text{inf}}\big) \to 0.95$.

[5] Normalizing by the tree would yield a less conservative metric; the tree-level indicator imposes a stricter penalty.

E.3 Computational details

All simulations were implemented in R (version 4.0.0 or later) (R Core Team 2024) and run on CPU resources only. We note that no GPU acceleration is required. The reference machine was a small university server with 25 cores and 256 GB of RAM, though a workstation with at least four cores and 16 GB of RAM is sufficient to reproduce a single simulation cell. We ran 500 MCMC replications for every combination of effect size and covariate dimension, with effect sizes spanning $k \in \{0,\, 0.2,\, \ldots,\, 2.0\}$ and number of covariates $P \in \{10,\, 50,\, 100\}$, yielding $500 \times 11 \times 3 = 16{,}500$ runs. Parallelizing across 25 cores, one full setting required approximately six days of wall-clock time. As a single-machine benchmark, an unparallelized run averaged 376 seconds (with a standard deviation of 37 seconds) on an Intel i7-1365U laptop.

We note that in few sparse covariate settings, BCF-IV of Bargagli-Stoffi et al. (2022) occasionally returned $\widehat{\pi}_C^{\text{BCF}}(x) = 0$, which renders the second-stage ratio $\widehat{\text{ITT}}_Y^{\text{BCF}}(x) \,/\, \widehat{\pi}_C^{\text{BCF}}(x)$ for BCF-IV undefined when forming $\widehat{\tau}^{\text{BCF}}(x)$. We replaced exact zeros with a small positive constant to let the algorithm proceed. The corresponding SBCF-IV estimates were numerically well behaved across all replications and required no such adjustment.

E.4 Results

Table 2: Tree-level subgroup detection: DR and FDR for BCF-IV and SBCF-IV across covariate dimensions $P \in \{10, 50, 100\}$ and effect sizes $k \in \{0, 0.2, \ldots, 2\}$. DR is reported overall and separately for the two true heterogeneity subgroups l_1 and l_2. Results are averaged over $M = 500$ Monte Carlo replications with $N = 1{,}000$. This table reports the numerical values underlying Figure 4.1.

		BCF-IV				SBCF-IV			
		DR				DR			
P	k	Overall	l_1	l_2	FDR	Overall	l_1	l_2	FDR
	0	0.003	0.004	0.002	0.036	0.013	0.012	0.014	0.030
	0.2	0.006	0.006	0.006	0.036	0.051	0.052	0.050	0.034
	0.4	0.018	0.018	0.018	0.058	0.211	0.210	0.212	0.062
	0.6	0.061	0.070	0.052	0.094	0.391	0.396	0.386	0.058
	0.8	0.103	0.098	0.108	0.122	0.656	0.648	0.664	0.062
10	1	0.178	0.178	0.178	0.236	0.873	0.876	0.870	0.040
	1.2	0.222	0.220	0.224	0.336	0.950	0.948	0.952	0.046
	1.4	0.263	0.286	0.240	0.468	0.984	0.986	0.982	0.028
	1.6	0.289	0.282	0.296	0.552	0.997	1.000	0.994	0.048
	1.8	0.322	0.336	0.308	0.624	1.000	1.000	1.000	0.030
	2	0.328	0.328	0.328	0.728	1.000	1.000	1.000	0.060
	0	0.000	0.000	0.000	0.046	0.004	0.004	0.004	0.038
	0.2	0.000	0.000	0.000	0.032	0.019	0.020	0.018	0.040
	0.4	0.000	0.000	0.000	0.024	0.148	0.146	0.150	0.042
	0.6	0.001	0.000	0.002	0.048	0.485	0.490	0.480	0.040
	0.8	0.013	0.012	0.014	0.124	0.838	0.836	0.840	0.054
50	1	0.021	0.014	0.028	0.186	0.963	0.962	0.964	0.028
	1.2	0.027	0.022	0.032	0.310	0.984	0.986	0.982	0.046
	1.4	0.034	0.034	0.034	0.400	0.997	0.996	0.998	0.036
	1.6	0.037	0.026	0.048	0.516	0.999	0.998	1.000	0.050
	1.8	0.042	0.042	0.042	0.562	0.999	1.000	0.998	0.052
	2	0.048	0.050	0.046	0.634	1.000	1.000	1.000	0.048
	0	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.040
	0.2	0.000	0.000	0.000	0.028	0.004	0.002	0.006	0.028
	0.4	0.000	0.000	0.000	0.036	0.067	0.058	0.076	0.054
	0.6	0.001	0.000	0.002	0.058	0.391	0.390	0.392	0.086
	0.8	0.000	0.000	0.000	0.086	0.817	0.814	0.820	0.048
100	1	0.001	0.000	0.002	0.118	0.954	0.952	0.956	0.046
	1.2	0.006	0.004	0.008	0.208	0.990	0.990	0.990	0.040
	1.4	0.009	0.006	0.012	0.256	0.995	0.996	0.994	0.030
	1.6	0.013	0.016	0.010	0.376	1.000	1.000	1.000	0.030
	1.8	0.014	0.008	0.020	0.482	0.998	0.998	0.998	0.044
	2	0.015	0.020	0.010	0.516	0.999	0.998	1.000	0.052

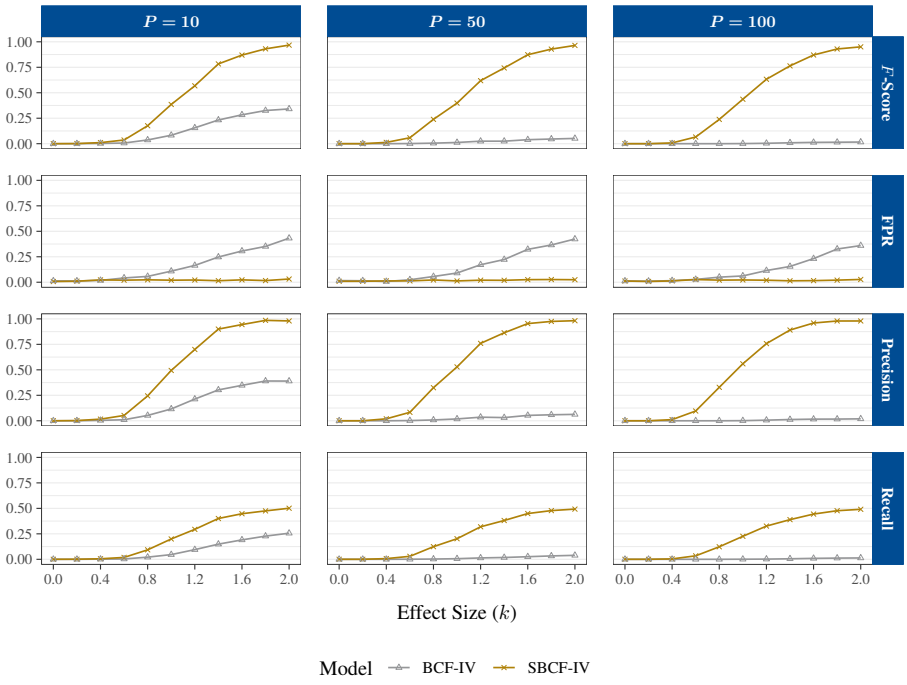

Figure E.1: Unit-level classification performance: Recall, Precision, False Positive Rate (FPR), and F-score as a function of effect size k, across covariate dimensions $P \in \{10, 50, 100\}$. All four metrics are defined in (E.5) and computed over all units in $\mathcal{I}_{\text{inf}}$, with significance evaluated at $\alpha = 0.05$ using Holm-adjusted p-values. Results are averaged over $M = 500$ Monte Carlo replications with $N = 1{,}000$, and compare SBCF-IV (orange) with BCF-IV (grey). This figure complements Figure 4.2, which reports Precision and F-score only, by additionally reporting Recall and FPR across the same (k, P) grid.

Table 3: Unit-level classification metrics: Recall, Precision, F-score, and FPR for BCF-IV and SBCF-IV across covariate dimensions $P \in \{10, 50, 100\}$ and effect sizes $k \in \{0, 0.2, \ldots, 2\}$. All four metrics are defined in (E.5) and computed over units in $\mathcal{I}_{\text{inf}}$, with significance evaluated at $\alpha = 0.05$ using Holm-adjusted p-values. Results are averaged over $M = 500$ Monte Carlo replications with $N = 1{,}000$; standard deviations are reported in parentheses. This table reports the numerical values underlying Figure E.1.

		BCF-IV				SBCF-IV			
P	k	Recall	Precision	F-Score	FPR	Recall	Precision	F-Score	FPR
10	0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.013 (0.087)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.007 (0.046)
	0.2	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.012 (0.081)	0.001 (0.015)	0.002 (0.045)	0.001 (0.030)	0.010 (0.061)
	0.4	0.001 (0.021)	0.004 (0.063)	0.003 (0.041)	0.018 (0.095)	0.005 (0.042)	0.016 (0.126)	0.011 (0.084)	0.023 (0.108)
	0.6	0.003 (0.033)	0.010 (0.100)	0.006 (0.065)	0.042 (0.144)	0.018 (0.080)	0.052 (0.222)	0.036 (0.156)	0.020 (0.088)
	0.8	0.021 (0.087)	0.052 (0.216)	0.037 (0.149)	0.057 (0.166)	0.093 (0.166)	0.244 (0.425)	0.176 (0.310)	0.024 (0.099)
	1	0.046 (0.124)	0.116 (0.312)	0.083 (0.221)	0.109 (0.214)	0.200 (0.209)	0.494 (0.495)	0.385 (0.398)	0.020 (0.102)
	1.2	0.095 (0.175)	0.213 (0.390)	0.156 (0.280)	0.165 (0.262)	0.294 (0.206)	0.700 (0.451)	0.568 (0.387)	0.022 (0.104)
	1.4	0.149 (0.204)	0.303 (0.424)	0.232 (0.317)	0.247 (0.300)	0.401 (0.156)	0.900 (0.290)	0.783 (0.293)	0.015 (0.091)
	1.6	0.191 (0.222)	0.348 (0.423)	0.283 (0.336)	0.306 (0.322)	0.447 (0.122)	0.944 (0.204)	0.870 (0.227)	0.023 (0.105)
	1.8	0.228 (0.231)	0.390 (0.418)	0.326 (0.338)	0.351 (0.327)	0.475 (0.084)	0.985 (0.087)	0.931 (0.142)	0.016 (0.095)
	2	0.256 (0.245)	0.390 (0.398)	0.341 (0.339)	0.432 (0.331)	0.501 (0.068)	0.980 (0.081)	0.967 (0.090)	0.031 (0.126)
50	0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.016 (0.107)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.009 (0.049)
	0.2	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.015 (0.112)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.009 (0.047)
	0.4	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.007 (0.069)	0.007 (0.049)	0.019 (0.135)	0.013 (0.090)	0.013 (0.066)
	0.6	0.001 (0.014)	0.002 (0.045)	0.001 (0.029)	0.024 (0.127)	0.029 (0.098)	0.083 (0.275)	0.058 (0.195)	0.013 (0.070)
	0.8	0.003 (0.029)	0.008 (0.089)	0.005 (0.058)	0.056 (0.174)	0.124 (0.181)	0.325 (0.464)	0.239 (0.348)	0.022 (0.099)
	1	0.006 (0.044)	0.018 (0.133)	0.012 (0.088)	0.091 (0.212)	0.202 (0.198)	0.529 (0.498)	0.398 (0.389)	0.013 (0.077)
	1.2	0.014 (0.069)	0.035 (0.181)	0.025 (0.125)	0.172 (0.284)	0.319 (0.191)	0.760 (0.418)	0.619 (0.366)	0.020 (0.095)
	1.4	0.018 (0.086)	0.031 (0.157)	0.025 (0.119)	0.223 (0.314)	0.381 (0.170)	0.864 (0.333)	0.743 (0.320)	0.019 (0.100)
	1.6	0.025 (0.099)	0.053 (0.210)	0.039 (0.151)	0.323 (0.359)	0.449 (0.115)	0.954 (0.181)	0.873 (0.209)	0.025 (0.111)
	1.8	0.033 (0.119)	0.059 (0.213)	0.046 (0.160)	0.366 (0.372)	0.478 (0.086)	0.975 (0.112)	0.928 (0.149)	0.027 (0.114)
	2	0.039 (0.126)	0.063 (0.211)	0.052 (0.168)	0.424 (0.384)	0.493 (0.070)	0.982 (0.084)	0.964 (0.103)	0.025 (0.112)
100	0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.012 (0.103)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.012 (0.060)
	0.2	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.010 (0.094)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.007 (0.047)
	0.4	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.016 (0.109)	0.004 (0.038)	0.011 (0.104)	0.008 (0.071)	0.013 (0.057)
	0.6	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.027 (0.137)	0.034 (0.104)	0.096 (0.293)	0.066 (0.201)	0.026 (0.092)
	0.8	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.050 (0.185)	0.123 (0.178)	0.328 (0.466)	0.238 (0.343)	0.021 (0.099)
	1	0.001 (0.021)	0.001 (0.025)	0.001 (0.023)	0.062 (0.195)	0.226 (0.209)	0.559 (0.491)	0.436 (0.398)	0.022 (0.107)
	1.2	0.002 (0.026)	0.006 (0.077)	0.004 (0.052)	0.115 (0.252)	0.326 (0.197)	0.757 (0.420)	0.632 (0.376)	0.020 (0.098)
	1.4	0.005 (0.046)	0.012 (0.105)	0.009 (0.077)	0.155 (0.300)	0.389 (0.157)	0.890 (0.302)	0.764 (0.299)	0.014 (0.081)
	1.6	0.009 (0.063)	0.016 (0.113)	0.013 (0.087)	0.232 (0.337)	0.444 (0.119)	0.960 (0.179)	0.871 (0.214)	0.016 (0.092)
	1.8	0.012 (0.074)	0.017 (0.107)	0.014 (0.089)	0.326 (0.384)	0.477 (0.084)	0.979 (0.105)	0.930 (0.149)	0.021 (0.098)
	2	0.013 (0.082)	0.019 (0.119)	0.016 (0.099)	0.360 (0.395)	0.491 (0.077)	0.979 (0.092)	0.951 (0.121)	0.027 (0.119)

Table 4: Estimation precision and uncertainty quantification for $\widehat{\tau}^{\text{CACE}}(x)$: MSE, Bias, Absolute Bias, 95% CI Coverage, and CI Length for BCF-IV and SBCF-IV across covariate dimensions $P \in \{10, 50, 100\}$ and effect sizes $k \in \{0, 0.2, \ldots, 2\}$. Results are averaged over $M = 500$ Monte Carlo replications with $N = 1{,}000$; standard deviations are reported in parentheses. This table reports the full numerical results underlying Table 1, which restricts attention to $k \in \{0, 1, 2\}$ and pools l_1 and l_2.

		BCF-IV					SBCF-IV				
P	k	MSE	Bias	Abs. Bias	Coverage	CI Length	MSE	Bias	Abs. Bias	Coverage	CI Length
10	0	0.188 (0.163)	-0.004 (0.225)	0.286 (0.140)	0.948 (0.176)	1.434 (0.219)	0.191 (0.161)	-0.004 (0.222)	0.314 (0.131)	0.957 (0.131)	1.600 (0.176)
	0.2	0.214 (0.148)	0.011 (0.233)	0.336 (0.114)	0.891 (0.178)	1.448 (0.200)	0.221 (0.160)	0.005 (0.229)	0.357 (0.129)	0.912 (0.143)	1.612 (0.183)
	0.4	0.294 (0.178)	0.014 (0.232)	0.416 (0.122)	0.797 (0.206)	1.474 (0.208)	0.306 (0.193)	0.018 (0.234)	0.438 (0.133)	0.847 (0.176)	1.652 (0.185)
	0.6	0.375 (0.222)	-0.012 (0.237)	0.494 (0.172)	0.708 (0.297)	1.537 (0.205)	0.376 (0.256)	-0.002 (0.271)	0.493 (0.198)	0.784 (0.234)	1.706 (0.193)
	0.8	0.452 (0.326)	-0.004 (0.254)	0.547 (0.230)	0.697 (0.341)	1.591 (0.258)	0.370 (0.328)	-0.013 (0.288)	0.485 (0.262)	0.809 (0.283)	1.759 (0.206)
	1	0.473 (0.333)	0.000 (0.286)	0.569 (0.240)	0.695 (0.337)	1.600 (0.210)	0.298 (0.327)	-0.002 (0.308)	0.426 (0.256)	0.899 (0.235)	1.794 (0.160)
	1.2	0.581 (0.425)	-0.003 (0.325)	0.626 (0.275)	0.631 (0.333)	1.653 (0.205)	0.273 (0.317)	-0.008 (0.337)	0.404 (0.245)	0.911 (0.216)	1.823 (0.151)
	1.4	0.607 (0.483)	0.021 (0.358)	0.641 (0.271)	0.643 (0.318)	1.667 (0.186)	0.232 (0.248)	0.008 (0.331)	0.383 (0.208)	0.947 (0.161)	1.824 (0.147)
	1.6	0.718 (0.574)	-0.006 (0.380)	0.711 (0.314)	0.553 (0.334)	1.660 (0.198)	0.221 (0.230)	0.009 (0.328)	0.373 (0.200)	0.954 (0.148)	1.831 (0.148)
	1.8	0.814 (0.584)	-0.016 (0.406)	0.760 (0.313)	0.536 (0.319)	1.681 (0.174)	0.223 (0.225)	-0.029 (0.341)	0.376 (0.203)	0.945 (0.161)	1.829 (0.142)
	2	0.931 (0.729)	0.014 (0.446)	0.811 (0.354)	0.478 (0.335)	1.697 (0.174)	0.214 (0.211)	0.012 (0.337)	0.368 (0.199)	0.953 (0.149)	1.840 (0.141)
50	0	0.205 (0.244)	-0.003 (0.223)	0.273 (0.145)	0.953 (0.178)	1.399 (0.244)	0.214 (0.151)	-0.003 (0.222)	0.358 (0.140)	0.960 (0.111)	1.769 (0.143)
	0.2	0.229 (0.180)	0.012 (0.226)	0.321 (0.110)	0.894 (0.192)	1.387 (0.206)	0.240 (0.155)	0.014 (0.228)	0.383 (0.129)	0.938 (0.107)	1.780 (0.150)
	0.4	0.332 (0.164)	0.013 (0.235)	0.449 (0.079)	0.712 (0.210)	1.381 (0.204)	0.297 (0.199)	0.003 (0.261)	0.431 (0.144)	0.895 (0.144)	1.786 (0.142)
	0.6	0.471 (0.220)	0.015 (0.232)	0.574 (0.140)	0.515 (0.314)	1.422 (0.258)	0.271 (0.232)	0.020 (0.279)	0.413 (0.193)	0.912 (0.173)	1.807 (0.150)
	0.8	0.640 (0.338)	-0.004 (0.244)	0.671 (0.222)	0.457 (0.384)	1.445 (0.233)	0.269 (0.284)	-0.010 (0.324)	0.410 (0.217)	0.919 (0.193)	1.825 (0.147)
	1	0.785 (0.472)	0.012 (0.241)	0.744 (0.288)	0.454 (0.404)	1.495 (0.232)	0.213 (0.232)	-0.003 (0.309)	0.362 (0.203)	0.949 (0.151)	1.830 (0.140)
	1.2	0.885 (0.574)	-0.004 (0.257)	0.802 (0.333)	0.470 (0.389)	1.531 (0.228)	0.229 (0.253)	-0.030 (0.331)	0.375 (0.206)	0.942 (0.167)	1.819 (0.150)
	1.4	1.130 (0.750)	0.003 (0.253)	0.923 (0.375)	0.373 (0.349)	1.546 (0.274)	0.223 (0.216)	0.001 (0.338)	0.380 (0.204)	0.953 (0.146)	1.831 (0.142)
	1.6	1.401 (0.968)	-0.011 (0.279)	1.031 (0.427)	0.341 (0.335)	1.566 (0.245)	0.228 (0.239)	-0.021 (0.343)	0.377 (0.213)	0.944 (0.161)	1.845 (0.144)
	1.8	1.695 (1.171)	0.001 (0.296)	1.157 (0.457)	0.260 (0.307)	1.587 (0.225)	0.229 (0.242)	-0.011 (0.348)	0.380 (0.206)	0.945 (0.161)	1.839 (0.143)
	2	2.088 (1.446)	-0.008 (0.299)	1.291 (0.508)	0.202 (0.256)	1.593 (0.255)	0.204 (0.203)	0.002 (0.320)	0.359 (0.192)	0.957 (0.145)	1.836 (0.138)
100	0	0.196 (0.234)	-0.020 (0.226)	0.253 (0.143)	0.957 (0.180)	1.325 (0.215)	0.217 (0.162)	-0.021 (0.226)	0.360 (0.143)	0.951 (0.111)	1.785 (0.136)
	0.2	0.244 (0.277)	-0.003 (0.226)	0.314 (0.100)	0.882 (0.198)	1.337 (0.216)	0.247 (0.151)	-0.002 (0.226)	0.391 (0.123)	0.930 (0.109)	1.780 (0.136)
	0.4	0.357 (0.232)	-0.002 (0.237)	0.455 (0.070)	0.667 (0.200)	1.346 (0.211)	0.320 (0.252)	-0.008 (0.257)	0.450 (0.134)	0.887 (0.143)	1.807 (0.134)
	0.6	0.535 (0.304)	-0.006 (0.246)	0.608 (0.114)	0.445 (0.279)	1.351 (0.211)	0.304 (0.220)	0.002 (0.291)	0.441 (0.188)	0.889 (0.180)	1.797 (0.148)
	0.8	0.736 (0.430)	0.006 (0.236)	0.750 (0.175)	0.299 (0.338)	1.379 (0.244)	0.252 (0.260)	-0.001 (0.313)	0.392 (0.210)	0.919 (0.184)	1.834 (0.156)
	1	0.954 (0.418)	0.019 (0.220)	0.872 (0.251)	0.273 (0.351)	1.417 (0.243)	0.230 (0.215)	0.005 (0.319)	0.387 (0.199)	0.950 (0.149)	1.826 (0.139)
	1.2	1.206 (0.633)	-0.013 (0.245)	0.975 (0.326)	0.284 (0.343)	1.460 (0.339)	0.197 (0.196)	-0.007 (0.310)	0.354 (0.186)	0.954 (0.147)	1.820 (0.143)
	1.4	1.544 (0.806)	0.021 (0.249)	1.122 (0.369)	0.237 (0.307)	1.461 (0.233)	0.229 (0.242)	0.030 (0.336)	0.377 (0.208)	0.945 (0.161)	1.829 (0.142)
	1.6	1.803 (0.998)	-0.013 (0.253)	1.210 (0.424)	0.221 (0.277)	1.504 (0.269)	0.205 (0.214)	-0.021 (0.325)	0.360 (0.195)	0.960 (0.136)	1.832 (0.134)
	1.8	2.095 (1.283)	-0.011 (0.262)	1.301 (0.489)	0.211 (0.275)	1.516 (0.255)	0.227 (0.245)	-0.001 (0.349)	0.373 (0.210)	0.942 (0.166)	1.825 (0.139)
	2	2.551 (1.522)	0.013 (0.257)	1.454 (0.526)	0.157 (0.222)	1.535 (0.235)	0.230 (0.238)	0.022 (0.340)	0.376 (0.212)	0.942 (0.164)	1.835 (0.146)

F Supplementary materials for empirical applications

F.1 Oregon Health Insurance Experiment (OHIE)

We evaluate SBCF-IV on the OHIE data, replicating the empirical setup of Johnson et al. (2022). The OHIE assembled administrative records on hospital discharges, credit reports, and mortality; survey data on healthcare utilization, financial strain, and self-reported health; and pre-randomization demographic information. Johnson et al. (2022) match individuals on pre-randomization demographics, including sex, age, language preference at lottery sign-up (English or other), MSA residency, education (less than high school, high school diploma or GED, vocational or two-year degree, four-year degree or higher), and self-identified race. Because Hispanic and Black self-identification as well as education contained missing values, missingness indicators are included in the matching set.

SBCF-IV partitions on age at the root and on English (lottery sign-up language preference) within both age branches. The education covariate appears as a tertiary split within the younger non-English-preferring subtree. We receive seven leaves with conditional CACEs spanning -6.017 to 3.630, of which only one subgroup has a statistically significant p-value at the 10% level: English-preferring compliers aged 38–59, with $\widehat{\tau}^{\mathrm{CACE}}(x) = 2.263$ (adjusted $p = 0.0945$). This leaf carries roughly half of the inference sample and a complier share of 0.663, and its positive sign and middle-age, English-preferring composition closely mirror the substantive Medicaid effects reported by Johnson et al. (2022).

The remaining six leaves carry inference-sample shares between 4% and 14% and produce estimates that should be read as exploratory rather than substantive findings. Two leaves carry negative point estimates ($\widehat{\tau}^{\mathrm{CACE}}(x) = -6.017$ for non-English-preferring compliers aged ≥ 60 (4% share) and $\widehat{\tau}^{\mathrm{CACE}}(x) = -1.403$ for non-English-preferring compliers aged 26–37 with above-median education (9%)) but are not statistically significant. We do not interpret them as evidence of adverse Medicaid effects as they are most plausibly small-sample noise in thin demographic strata. The non-significant positive leaves at $\widehat{\tau}^{\mathrm{CACE}}(x) = 3.630$ (English-preferring under 38, 11%) and $\widehat{\tau}^{\mathrm{CACE}}(x) = 1.563$ (non-English-preferring aged 26–37 with low education, 14%) suggest that positive effects may extend beyond the central English-preferring middle-aged subgroup but are not statistically resolvable at this sample size and tree depth.

Relative to the two complier subgroups identified by Johnson et al. (2022) ((i) non-Asian, English-preferring males over age 36, and (ii) compliers under age 36 who prefer English with at most a high-school education or GED) SBCF-IV recovers a partition close to a coarsened version of their subgroup (i): English preference and middle age are the leading dimensions, but our discovery tree does not split on sex or race and therefore cannot reproduce their male-only refinement. The qualitative location of the dominant positive Medicaid effect (older, English-preferring compliers) is preserved across both methods, while the demographic resolution differs.

F.2 401(k) eligibility and retirement savings

We apply SBCF-IV to the 401(k) data of Chernozhukov et al. (2018) and Bach et al. (2024), a canonical instrumental-variable benchmark in which the endogenous decision to participate in an employer-sponsored retirement plan is instrumented by eligibility. Conditional on a small set of job-choice covariates, eligibility is plausibly unconfounded and satisfies the exclusion restriction, while participation remains confounded by unobserved preferences for saving. In Subsection F.2, we use SBCF-IV to discover the subpopulations of eligible households for which participation has the largest impact on net financial assets.

The data are drawn from the 1991 Survey of Income and Program Participation, which has become a canonical testbed for instrumental-variable methods in labor and public economics. The outcome Y_i is a household's net financial assets (IRA and 401(k) balances, checking accounts, savings bonds, and related holdings, net of non-mortgage debt); the endogenous treatment W_i is an indicator for participation in a 401(k) plan; and the instrument Z_i is an indicator for eligibility to enroll in such a plan through one's employer. Following Poterba et al. (1992) and Poterba et al. (1995), the identifying argument is that, conditional on a small set of job-choice covariates X_i (notably income, together with age, family size, education, marital status, two-earner and defined-benefit pension status, IRA participation, and home ownership), whether an employer offered a 401(k) plan around the time of the survey can be treated as exogenous to the household's saving behavior, while the decision to actually

participate conditional on eligibility remains endogenous. This places us in the irregular assignment mechanism of Section 2: Z_i is plausibly unconfounded given X_i and satisfies the exclusion restriction (eligibility affects assets only through participation), whereas W_i is confounded by unobserved preferences for saving. The SBCF-IV algorithm is used to discover for which subpopulations of eligible households participation has the largest impact on net financial assets.

SBCF-IV partitions primarily on inc (household income), which appears at the root and at every internal node except one; pira (IRA holdings) enters only as a secondary split within a thin upper-middle-income window. None of the other portfolio or household-structure binaries (db, hown, marr, twoearn) survive into the depth-four discovery tree, in contrast to the OHIE partition, which split on age, education, and language-of-materials preference. The dominance of income aligns SBCF-IV's discovered structure with the lifecycle and earnings-gradient emphasis of the classical 401(k) saving literature (Poterba et al. 1992, 1995, Engen et al. 1996, Engen & Gale 2000): heterogeneity in $\tau^{\mathrm{CACE}}(x)$ is concentrated along the income margin, with effects rising broadly from the lower-middle-income mass into the upper-middle-income range before becoming statistically unidentifiable in the thin high-income tail.

Six of the seven leaves carry inference-sample shares below 7%, and four have shares at or below 1%. These small leaves correspond to a few dozen households each and should be read as small-sample artifacts rather than substantive heterogeneity: none of them display statistically significant Holm-adjusted p-values at the 10% level. Only two leaves remain significant at the 10% level after adjustment: the bulk subgroup with household income lower than $68{,}810$ (89.1% of observations in the inference sample and $\widehat{\tau}^{\mathrm{CACE}}(x) = \$17{,}818$ for this subgroup), and the small upper-middle-income subgroup with household income between $92{,}690$ and $110{,}400$) (2.6%, $\widehat{\tau}^{\mathrm{CACE}}(x) = \$53{,}422$). The qualitative ordering (larger effects in the upper-middle-income subgroup than in the lower-income mass)is consistent with the finding in Engen & Gale (2000) that 401(k) effects on net financial assets are larger for higher-earnings groups, who hold the bulk of 401(k) assets and for whom contributions are most likely to substitute from taxable accounts rather than represent new saving (Chetty et al. 2014).

Two structural parallels with the OHIE application in Section F.1 persist. First, complier shares are nearly flat across leaves (0.82–1.00 here, 0.58–0.67 in OHIE), so heterogeneity is identified almost entirely from variation in the conditional ITT, matching the simulation evidence in Section 4 that SBCF-IV reliably recovers ITT-driven partitions. Second, both statistically significant leaves carry estimates above the cross-fitted LATE on the same trimmed sample ($\approx \$9{,}000$–$\$13{,}000$ in Chernozhukov et al. (2018)): the bulk leaf at $17,818 lies a modest $5,000–$9,000 above that identification benchmark, while the upper-middle-income leaf at $53,422 lies further above and should be read as partly inflated by sample selection. The residual upward bias is consistent with longstanding concerns that 401(k) effects on net financial assets are inflated by selection on saver type (Engen et al. 1996, Engen & Gale 2000) and by mechanical accumulation in tax-favored accounts that does not represent net new saving (Chetty et al. 2014). We therefore read the leaf estimates of the statistically significant subgroups as mildly upward-biased point estimates of the structural participation effect. Further, we treat the high-income tail leaves as exploratory and emphasize the shape of the discovered partition (income-dominated, with effects rising through middle income before destabilizing at the top) over the effect estimate magnitudes of any individual subgroup.

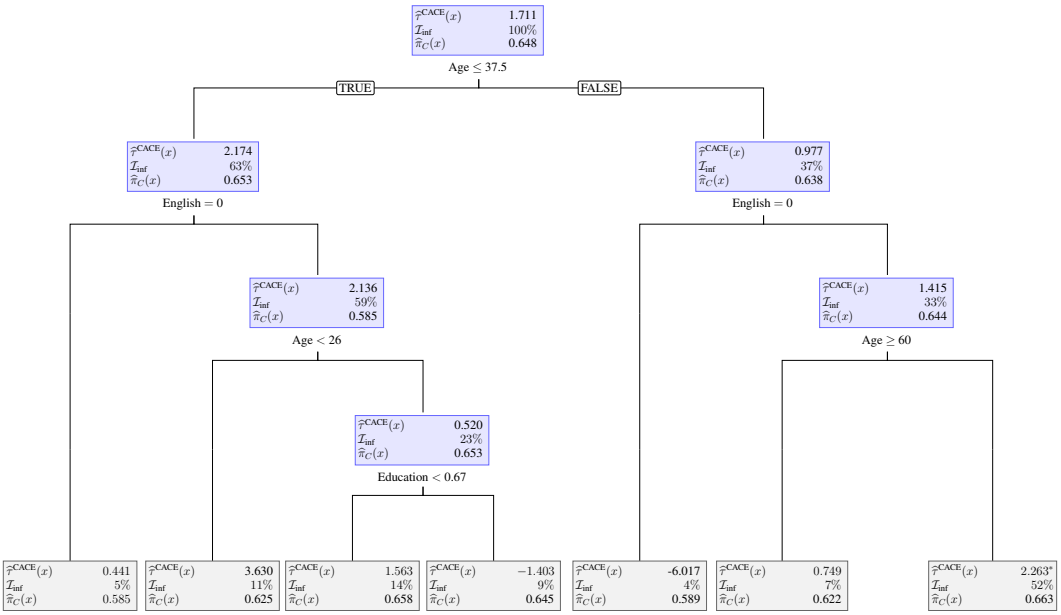

Figure F.1: Discovered partition from applying SBCF-IV with Algorithm 1 to the Oregon Health Insurance Experiment. Each internal and terminal node reports (top) the estimated conditional CACE $\widehat{\tau}^{\text{CACE}}(x)$, (middle) the share of the inference sample $\mathcal{I}_{\text{inf}}$ falling into the node, and (bottom) the estimated complier share $\widehat{\pi}_C(x)$. Splits are made on *Age*, *English*, and *Education*, with the splitting rule indicated at each branch. At terminal nodes, $\widehat{\tau}^{\text{CACE}}(x)$ coincides with the subgroup-level 2SLS estimator $\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j}$ in Definition 2.4; an asterisk marks leaves whose subgroup CACE is statistically significant at the 10% level.

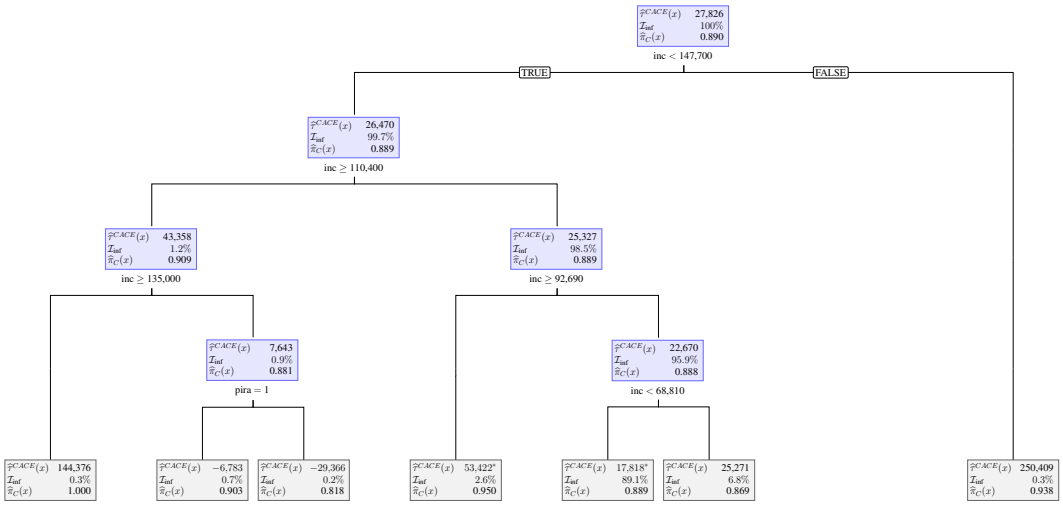

Figure F.2: Discovered partition from applying SBCF-IV with Algorithm 1 to the 401(k) dataset. Each internal and terminal node reports (top) the estimated conditional CACE $\widehat{\tau}^{\text{CACE}}(x)$, (middle) the share of the inference sample $\mathcal{I}_{\text{inf}}$ falling into the node, and (bottom) the estimated complier share $\widehat{\pi}_C(x)$. At terminal nodes, $\widehat{\tau}^{\text{CACE}}(x)$ coincides with the subgroup-level 2SLS estimator $\widehat{\tau}^{\text{2SLS}}_{\mathbb{X}_j}$ in Definition 2.4; an asterisk marks leaves whose subgroup CACE is statistically significant at the 10% level.