Causal inference in high-dimensional survival analysis

A Bayesian regression trees approach

Tijn Jacobs
joint work with Wessel N. van Wieringen and Stéphanie L. van der Pas

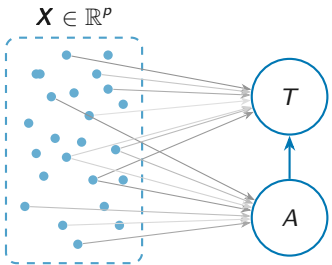
Vrije Universiteit Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

Why causal inference in high dimensions?

Modern applications, e.g. genomics:

- Thousands of covariates per subject.
- Confounding: covariates can influence both treatment and outcome.



Why causal inference in high dimensions?

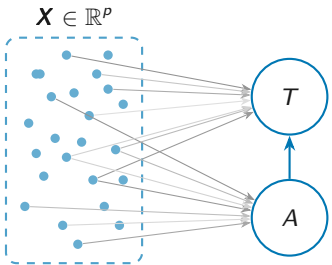
Modern applications, e.g. genomics:

- Thousands of covariates per subject.
- Confounding: covariates can influence both treatment and outcome.

Pancreatic cancer (PDAC):

- Among the deadliest cancers; median survival < 6 months.
- Molecularly heterogeneous across patients.

Does adjuvant radiotherapy improve survival, and for whom?

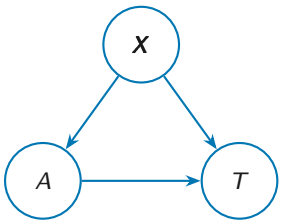


Problem setup

Outcome T is right (or interval) censored.

Covariates $\boldsymbol{X} \in \mathbb{R}^p$ are high-dimensional ($p > n$).

Treatment $A \in \{0, 1\}$ is not randomised, i.e. observational.



Research question:

What is the effect of the treatment on the censored outcome given the covariates (possible confounders)?

Potential outcomes and causal estimands

We adopt the potential outcome framework to define causal effects.

For each individual, consider:

- $T(1)$: the survival time under treatment.
- $T(0)$: the survival time under control.

Only one can be observed.

Potential outcomes and causal estimands

We adopt the potential outcome framework to define causal effects.

For each individual, consider:

- $T(1)$: the survival time under treatment.
- $T(0)$: the survival time under control.

Only one can be observed.

We focus on the *conditional average treatment effect* (CATE):

$$\tau(x) \;:=\; \mathbb{E}\left[\log T(1) - \log T(0) \mid \boldsymbol{X} = x\right],$$

and the corresponding *acceleration factor* (AF):

$$\mathrm{AF}(x) \;:=\; \exp\big(\tau(x)\big).$$

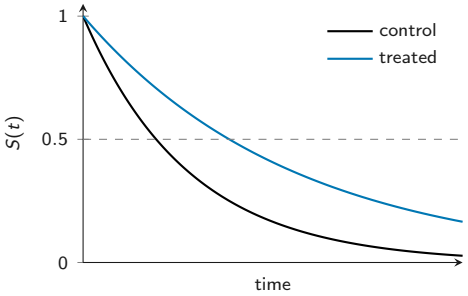
Acceleration factor: an interpretable estimand

On the log scale, $\tau(x)$ is a difference. Exponentiated, it becomes a *multiplier on survival time*:

$$T(1) \,\big|\, \boldsymbol{X} = x \;\overset{d}{=}\; e^{\tau(x)} \cdot T(0) \,\big|\, \boldsymbol{X} = x.$$

Acceleration factor $e^{\tau(x)}$:

- Treatment stretches the patient's clock.
- $e^{\tau(x)} = 2 \;\Rightarrow$ median survival doubles.
- Same shape, different time scale.



Why AFT?

1. Interpretability.
The acceleration factor $e^{\tau(x)}$ acts on the *time scale*, not on hazards.

2. Collapsibility.
The marginal effect is the average of the conditional effects.

3. Robustness.
AFT is robust to omitted independent covariates (Hougaard, 1999).

See Brathovde, Putter, Valberg, Post (2024).

Identification assumptions

Potential-outcomes framework $\big(T(a), C(a)\big)$. Identification proceeds in two steps.

1. Causal assumptions — link potential outcomes to observables:

- *SUTVA*: no interference, well-defined potential outcomes.
- *Unconfoundedness*: $T(a) \perp\!\!\!\perp A \mid X$.
- *Positivity*: $0 < e(x) := P(A{=}1 \mid X{=}x) < 1$.

2. Censoring assumption — makes the distribution of $T \mid X, A$ identifiable from the censored data (Y, δ):

- *Independent censoring*: $C(a) \perp\!\!\!\perp T(a) \mid X, A$.

Unconfoundedness, intuitively

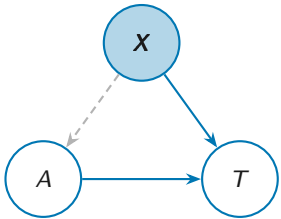
The worry: sicker patients may be more (or less) likely to be treated $\Rightarrow$ naive comparison mixes treatment effect with baseline differences.

Unconfoundedness:

$$T(a) \perp\!\!\!\perp A \mid \boldsymbol{X}, \qquad a \in \{0, 1\}.$$

Within each $\boldsymbol{X}$, treatment looks as if randomised.

Match like-for-like: same $\boldsymbol{X}$ $\Rightarrow$ treatment is the only systematic difference.



Conditioning on $\boldsymbol{X}$ cuts the dashed back-door arrow.

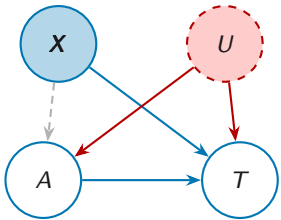
What if we miss a confounder?

Adjusting for $\boldsymbol{X}$ but missing an unmeasured confounder U:

- Back-door path $A \leftarrow U \rightarrow T$ stays open.
- Comparison still picks up the U effect.
- Bias remains — even with infinite data.

High-dimensional consequence: dropping covariates is risky — a dropped variable may be a confounder.

Strategy: keep every covariate in the model.



Unmeasured U leaves an open back-door path $A \leftarrow U \rightarrow T$.

The model

An accelerated failure time decomposition

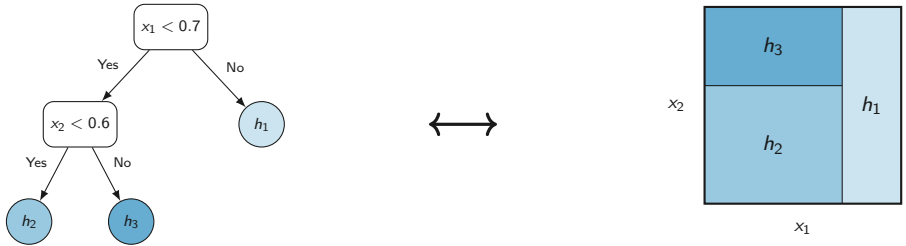
We model the log event-times:

$$\log T(a) \;=\; f\big(x, \hat{e}(x)\big) \;+\; a \cdot \tau(x) \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- f: prognostic function.
- $\hat{e}(x)$: estimated propensity score.
- $\tau(x)$: heterogeneous treatment effect (CATE).

Plan. Place a flexible Bayesian tree-ensemble prior on f and τ, separately.

BART in one slide



Model. $g(x) = \sum_{j=1}^{m} g(x;\, \mathcal{T}_j, \mathcal{H}_j)$ — a sum of many shallow trees, each piecewise constant on a partition of the covariate space.

Prior.

- Tree shape $\mathcal{T}_j$: depth-penalised Galton–Watson.
- Step heights $\mathcal{H}_j$: i.i.d. Gaussian, variance $\propto 1/m$.

Where could you regularise a tree ensemble?

Two places to regularise:

1. **Via the tree structure** $\mathcal{T}_j$ — depth penalty, splitting rules, covariate selection (DART, Soft-BART, ...).

2. **Via the step heights** $\mathcal{H}_j$ — shrink the *magnitudes* of step heights.

Existing work focuses on (1), but in high dimensions controlling the tree structure either fails to shrink enough or drops relevant confounders.

Our proposal: regularise through (2) — a global–local shrinkage prior on the step heights.

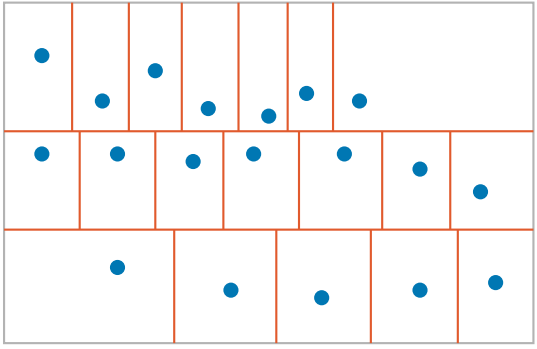
Connection to the normal means model

The normal means model:

$$Y_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

When the partition becomes fine, each step height effectively fits a single observation.

Natural place to plug in a global–local shrinkage prior.



Horseshoe prior on the step heights

For a tree $\mathcal{T}$ with leaves $\ell = 1, \ldots, L$:

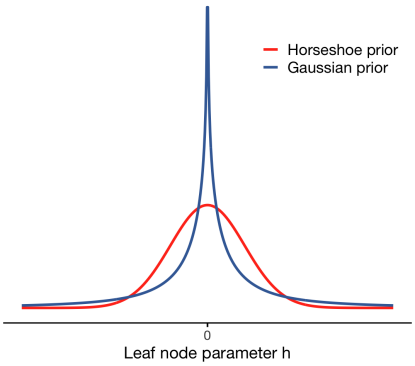
$$h_\ell \mid \lambda_\ell^2, \gamma^2 \;\sim\; \mathcal{N}\big(0,\, \lambda_\ell^2\, \gamma^2\big),$$

$$\lambda_\ell \sim \mathcal{C}^+(0,1), \quad \gamma \sim \mathcal{C}^+(0,1).$$

Two defining features:

- Pole at zero — strong shrinkage of noise.
- Heavy tails — large signals are not shrunk.

Global–local: γ pulls everything to zero; λ_ℓ lets individual leaves escape.



Horseshoe prior on the step heights

Reparametrise: $\kappa_\ell = 1/(1+\lambda_\ell^2\gamma^2) \in [0,1]$. Then the posterior mean is

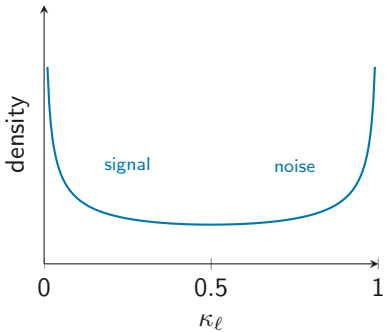
$$\widehat{\theta} \;=\; (1-\kappa_\ell)\,\bar{y}.$$

Under the horseshoe prior:

$$\kappa_\ell \;\sim\; \mathrm{Beta}\big(\tfrac{1}{2},\tfrac{1}{2}\big)\,.$$

This density is U-shaped. It puts mass at the two extremes:

- $\kappa_\ell \approx 1$: noise.
- $\kappa_\ell \approx 0$: signal.

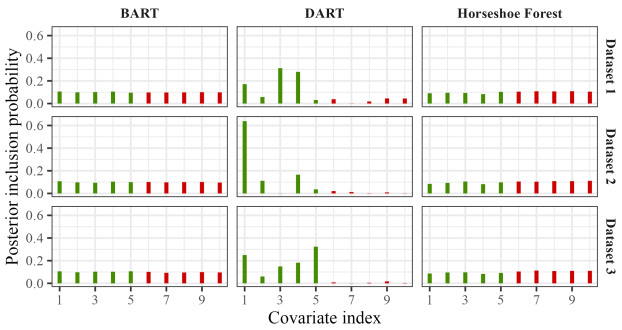


Why horseshoe beats DART

DART (Linero, 2018): places a Dirichlet prior on the *splitting probabilities* of the covariates.

Many covariates receive near-zero posterior inclusion probability.

Dropping a (weak) confounder ⇒ regularisation-induced confounding.



Posterior inclusion probabilities across covariates.

Horseshoe forest: keep every covariate split-eligible; shrink the *magnitudes* instead.

Posterior inference*

Sampling from the posterior

Recall the causal model:

$$\log T(a) \;=\; f(x, \hat{e}) \;+\; a \cdot \tau(x) \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

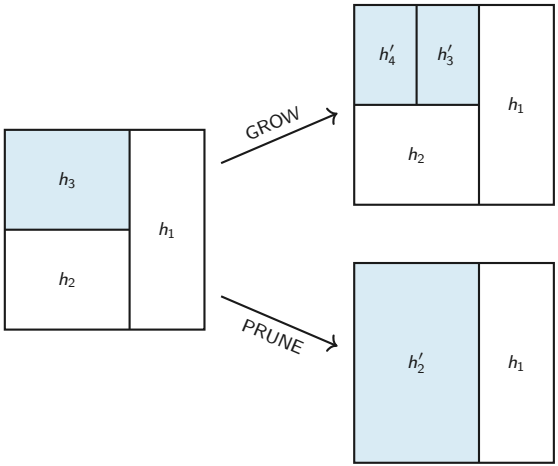
The sampling procedure is at the outer level a Gibbs sampler:

1. update f
2. update τ
3. augment censored observations
4. update σ^2

Reversible-jump within Gibbs

The horseshoe prior is *not* Gaussian-conjugate $\Rightarrow$ step heights cannot be marginalised out.

GROW and PRUNE moves change the dimension of $\mathcal{H}_j$. We use a reversible-jump step with a tailored proposal for $(\mathcal{T}_j, \mathcal{H}_j)$.



Augmentation of the censored data

Tanner & Wong (1987): treat censored event times as latent variables in the Gibbs sampler.

At each iteration t, draw the unobserved $\log T_i^{(t)}$ from

$$\mathcal{N}\big(\mu_i^{(t)},\, \sigma^{2,(t)}\big)$$

truncated to $(\log Y_i, \infty)$, with

$$\mu_i^{(t)} = f^{(t)}\big(x_i, \hat{e}(x_i)\big) + a_i\, \tau^{(t)}(x_i).$$

Augmentation of the censored data

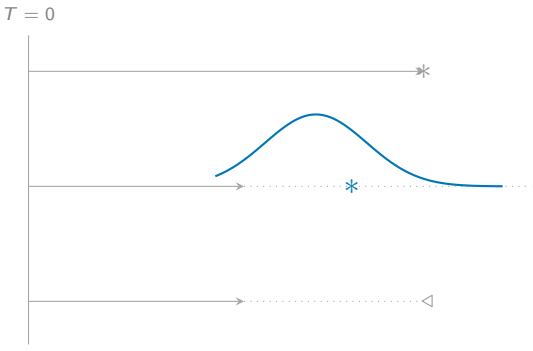
Tanner & Wong (1987): treat censored event times as latent variables in the Gibbs sampler.

At each iteration t, draw the unobserved $\log T_i^{(t)}$ from

$$\mathcal{N}\big(\mu_i^{(t)},\, \sigma^{2,(t)}\big)$$

truncated to $(\log Y_i, \infty)$, with

$$\mu_i^{(t)} = f^{(t)}\big(x_i, \hat{e}(x_i)\big) + a_i\, \tau^{(t)}(x_i).$$



Augmentation of the censored data

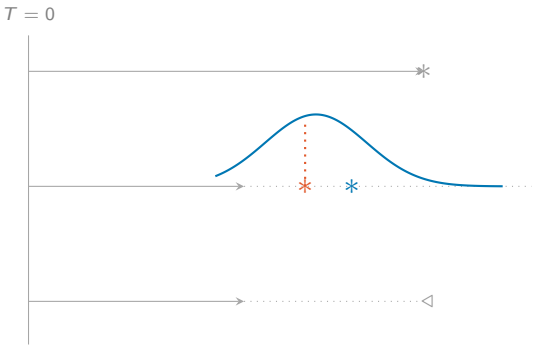
Tanner & Wong (1987): treat censored event times as latent variables in the Gibbs sampler.

At each iteration t, draw the unobserved $\log T_i^{(t)}$ from

$$\mathcal{N}\big(\mu_i^{(t)},\, \sigma^{2,(t)}\big)$$

truncated to $(\log Y_i, \infty)$, with

$$\mu_i^{(t)} = f^{(t)}\big(x_i, \hat{e}(x_i)\big) + a_i\, \tau^{(t)}(x_i).$$



Augmentation of the censored data

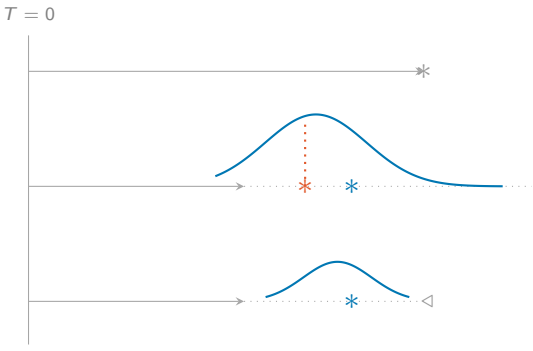
Tanner & Wong (1987): treat censored event times as latent variables in the Gibbs sampler.

At each iteration t, draw the unobserved $\log T_i^{(t)}$ from

$$\mathcal{N}\big(\mu_i^{(t)},\, \sigma^{2,(t)}\big)$$

truncated to $(\log Y_i, \infty)$, with

$$\mu_i^{(t)} = f^{(t)}\big(x_i, \hat{e}(x_i)\big) + a_i\, \tau^{(t)}(x_i).$$



Augmentation of the censored data

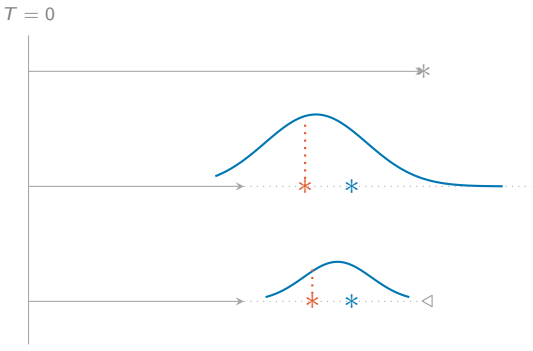
Tanner & Wong (1987): treat censored event times as latent variables in the Gibbs sampler.

At each iteration t, draw the unobserved $\log T_i^{(t)}$ from

$$\mathcal{N}\big(\mu_i^{(t)},\, \sigma^{2,(t)}\big)$$

truncated to $(\log Y_i, \infty)$, with

$$\mu_i^{(t)} = f^{(t)}\big(x_i, \hat{e}(x_i)\big) + a_i\, \tau^{(t)}(x_i).$$



Does it work?

Simulation 1: CATE estimation as p grows

Setting: $n = 200$, $50 \leq p \leq 5000$, censoring $\approx 35\%$.

Data generating process:

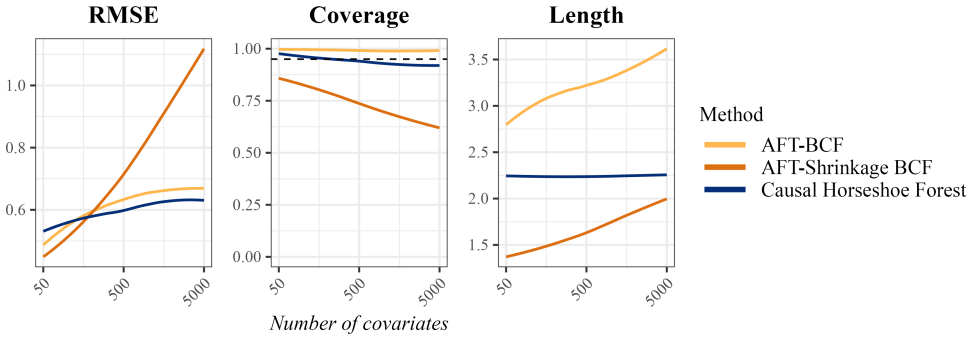
$$X_i \sim \mathcal{U}[0,1]^p, \quad A_i \sim \mathrm{Ber}(e(x_i)), \quad \log T_i \sim \mathcal{N}\big(f(x_i) + A_i\,\tau(x_i),\, \sigma^2\big),$$

where σ is chosen s.t. $\mathrm{Var}(\log T)\,/\,\sigma^2 = 10/9 \approx 1.111$.

Treatment effect (Friedman non-linear core + sparse main effects + sparse interactions):

$$\tau(x_i) = 10\sin(\pi x_{i1} x_{i2}) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5} + \beta_\tau^\top x_i + x_i^\top \Gamma\, x_i.$$

Simulation 1: results



Horseshoe Forest: stable RMSE and near-nominal coverage as p grows. AFT-BCF inflates intervals; AFT-Shrinkage BCF loses coverage.

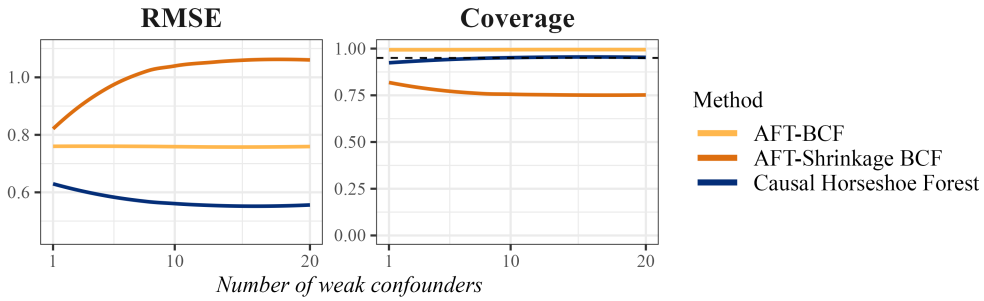
Simulation 2: regularisation-induced confounding

Setting: $n = 100$, $p = 500$, censoring $\approx 35\%$.

Two groups of confounders:

- $|S| = 5$ **strong** confounders: strong on both treatment and outcome.
- $|W| \in \{1, \ldots, 20\}$ **weak** confounders: strong on treatment, weak on outcome.

Simulation 2: results



Horseshoe Forest stays unbiased as weak confounders accumulate. DART-based competitors drop them and suffer regularisation-induced confounding.

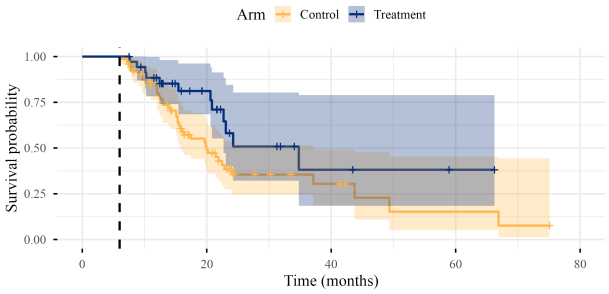
Application: pancreatic cancer

PDAC: the data

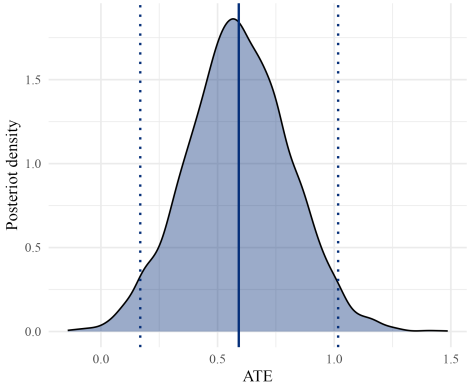
TCGA-PAAD cohort:

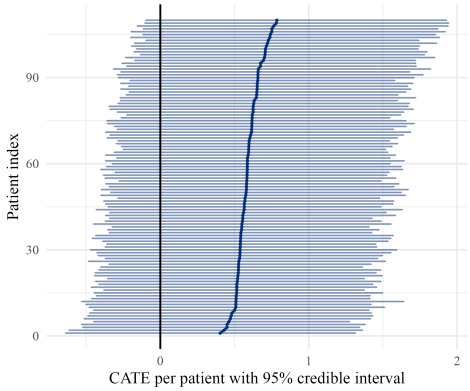
- $n = 110$ patients (6-month landmark).
- 11 clinical covariates + $\sim$ 3,000 gene expressions.
- Censoring $\approx 47\%$.
- Median follow-up: 23.3 months.

Treatment: adjuvant radiotherapy.



PDAC: results





$\widehat{\text{ATE}} \approx 0.55$, 95% CrI (0.10, 0.98).

Acceleration factor $e^{0.55} \approx 1.73$

Predictive performance: concordance index $= 0.75$.

Thanks for your attention!

Paper

Accepted at

Bayesian Analysis.

R package `ShrinkageTrees`

on CRAN (5000+ downloads).

Software paper on arXiv soon.

Stay up to date

`tijn-jacobs.github.io`

This work is supported by ERC Grant BayCause, no. 101074082. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

References

- Brathovde, M., Putter, H., Valberg, M., Post, R.A.J. (2024). The causal interpretation of acceleration factors. *arXiv:2409.01983*.
- Hahn, P.R., Murray, J.S., Carvalho, C.M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3), 965–1056.
- Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55(1), 13–22.
- Linero, A.R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *JASA*, 113(522), 626–636.
- Zigler, C.M. (2016). The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. *The American Statistician*, 70(1), 47–54.

Appendix

Unconfoundedness, intuitively

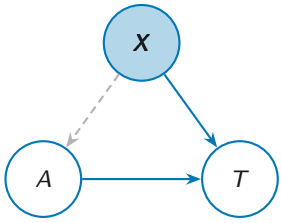
The worry: sicker patients may be more (or less) likely to be treated $\Rightarrow$ naive comparison mixes treatment effect with baseline differences.

Unconfoundedness:

$$T(a) \perp\!\!\!\perp A \mid \boldsymbol{X}, \qquad a \in \{0, 1\}.$$

Within each $\boldsymbol{X}$, treatment looks as if randomised.

Match like-for-like: same $\boldsymbol{X}$ $\Rightarrow$ treatment is the only systematic difference.



Conditioning on $\boldsymbol{X}$ cuts the dashed back-door arrow.

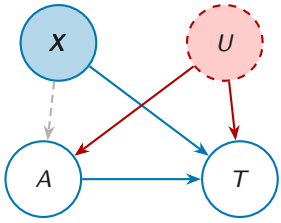
What if we miss a confounder?

Adjusting for $\boldsymbol{X}$ but missing an unmeasured confounder U:

- Back-door path $A \leftarrow U \rightarrow T$ stays open.
- Comparison still picks up the U effect.
- Bias remains — even with infinite data.

High-dimensional consequence: dropping covariates is risky — a dropped variable may be a confounder.

Strategy: keep every covariate in the model.



Unmeasured U leaves an open back-door path $A \leftarrow U \rightarrow T$.