Horseshoe Forests for High-Dimensional Causal Survival Analysis

ISBA World Meeting

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

Vrije Universiteit Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

This talk in one slide

Goal. Estimate heterogeneous causal effects of a binary treatment on a censored survival outcome when the number of covariates is large ($p \gg n$).

Idea. Take the BART prior and *move the regularisation from the tree structure to the leaf parameters* via a horseshoe prior on the step heights.

Cost. Conjugacy is lost $\Rightarrow$ reversible-jump-within-Gibbs sampler with a tailored proposal.

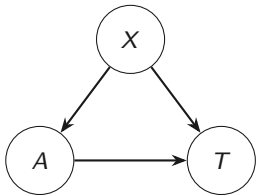
Pay-off.

- Stable RMSE and near-nominal coverage as p grows.
- Robust to regularisation-induced confounding.
- Application: adjuvant radiotherapy in pancreatic cancer (TCGA-PAAD).

The setting

Observational survival data $(Y_i, \delta_i, A_i, X_i)$, $i = 1, \ldots, n$:

- $Y_i = \min(T_i, C_i)$, $\delta_i = \mathbf{1}\{T_i \leq C_i\}$ (right-censored)
- $A_i \in \{0, 1\}$ (binary, non-randomised)
- $X_i \in \mathbb{R}^p$ with $p \gg n$



Question. What is the causal effect of A on T given X, allowing for heterogeneity across patients?

Estimand. The conditional average treatment effect on the log-survival scale,

$$\tau(x) := \mathbb{E}\left[\log T(1) - \log T(0) \,\middle|\, X = x\right].$$

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

An accelerated failure time decomposition

We model the log-event times:

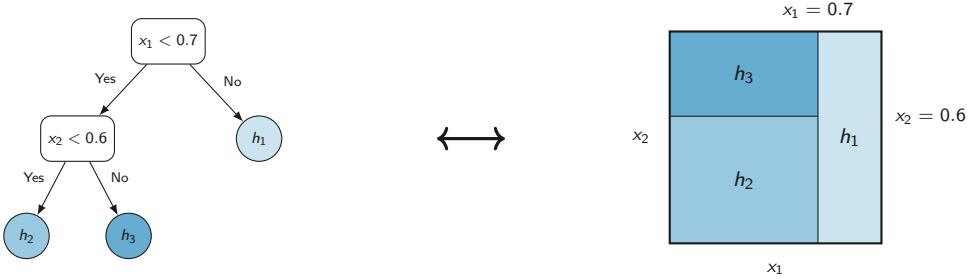
$$\log T(a) \;=\; f\big(x, \hat{e}(x)\big) \;+\; a \cdot \tau(x) \;+\; \varepsilon, \qquad \varepsilon \;\sim\; \mathcal{N}(0, \sigma^2).$$

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

Why AFT here? AFT effects are collapsible: the marginal effect is the average of conditional effects.

Plan. Place a flexible Bayesian tree-ensemble prior on both 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 does regularisation come from in BART?

Two places to regularise a tree ensemble:

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 controlling tree structure alone is not enough in high dimensions, or induces bias by dropping relevant confounders.

Novel contribution. Regularise through (2): a global–local shrinkage prior on the step heights.

Horseshoe prior on the step heights

For a tree with step heights $\ell = 1, \ldots, L$:

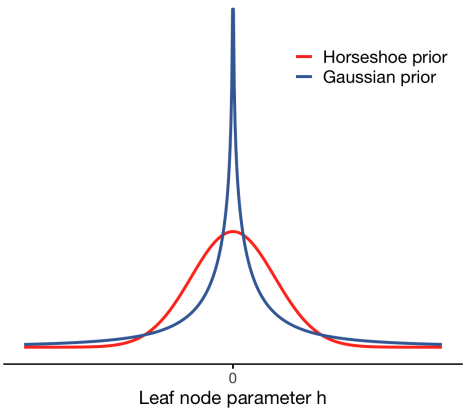
$$h_\ell \mid \lambda_\ell^2, \gamma^2 \;\sim\; \mathcal{N}\big(0,\, \lambda_\ell^2 \gamma^2\big),$$

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

Global–local shrinkage:

- γ pulls *everything* to zero.
- λ_ℓ lets individual step heights *escape* locally.

$\Rightarrow$ *regularise globally, pick up signals locally.*

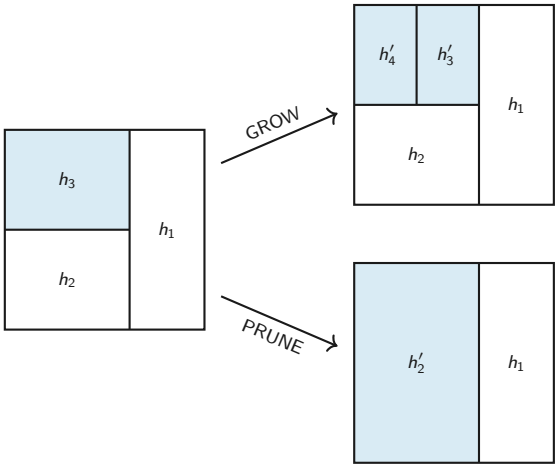


Inference: reversible-jump on trees

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

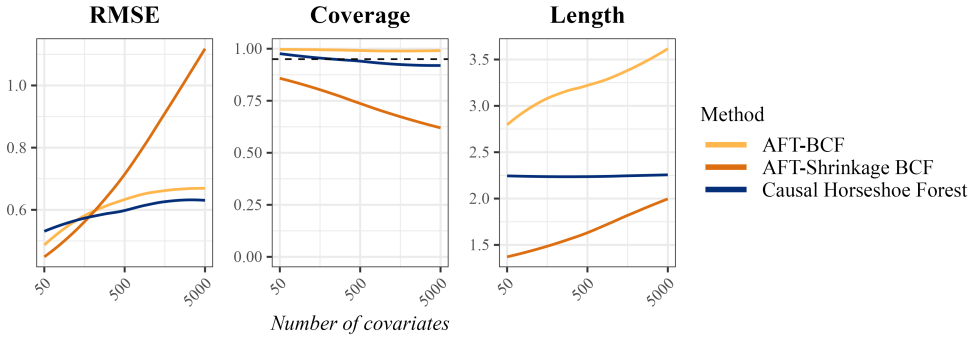
GROW and PRUNE change the dimension of $\mathcal{H}_j$. We use a **reversible-jump** step with a tailored proposal for $(\mathcal{T}_j, \mathcal{H}_j)$.

This sits inside a Gibbs sampler that updates the forests f and τ, the error variance σ^2, and the augmented censored event times.



Simulation 1: CATE estimation as p grows

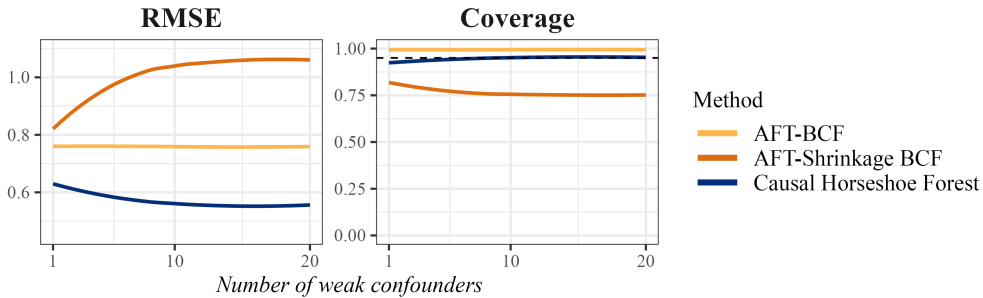
Non-linear treatment effect with sparse main (5%) + interaction (1%) structure; $n = 200$, $p \in [50, 5000]$, censoring $\approx 35\%$.



Simulation 2: regularisation-induced confounding

$n = 100$, $p = 500$, censoring $\approx 35\%$. Two groups of confounders:

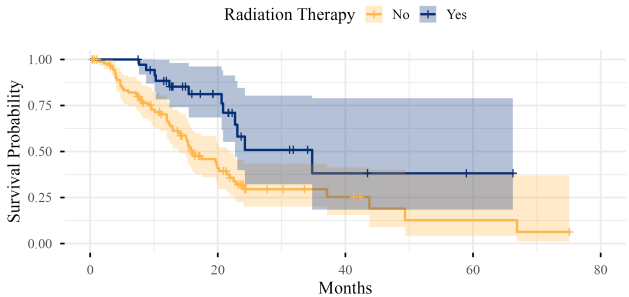
- $|S| = 5$ **strong** confounders — contribute substantially to both treatment and outcome.
- $|W| \in \{1, \ldots, 20\}$ **weak** confounders — predict treatment, but contribute only weakly to the outcome.



Adjuvant radiotherapy in PDAC

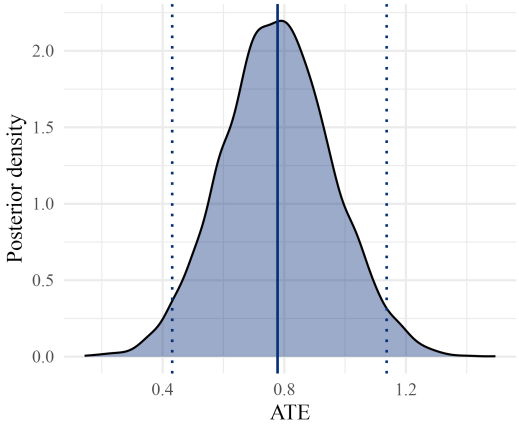
Data. TCGA-PAAD, $n = 130$ resected PDAC patients.

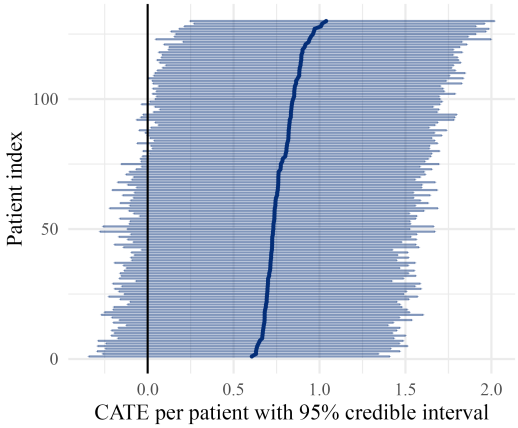
- *Outcome*: overall survival; censoring $\approx 47\%$, median follow-up 23.3 mo.
- *Treatment*: adjuvant radiotherapy.
- *Covariates*: 11 clinical + driver genes + top 3,000 genes by MAD ($p \approx 3{,}029$).



Question. Does adjuvant radiotherapy improve survival, and is the effect heterogeneous?

PDAC: results





$\widehat{\text{ATE}} \approx 0.75$, 95% CrI (0.40, 1.11).

$\approx$ 96.9% of patients have $P(\tau(X_i) > 0 \mid \text{data}) > 0.95$.

Take-aways

- **Where you regularise matters.** Shifting the regularisation from the tree structure onto the *step heights* of a tree ensemble gives a flexible, adaptive prior for high-dimensional CATE estimation.
- **The price is a reversible-jump sampler.** A tailored tree proposal keeps the chain mixing well.
- **Strong empirical performance.** Stable RMSE and near-nominal coverage as p grows; robust to regularisation-induced confounding where competing methods break down.

Thank you

- Preprint on arXiv: 2507.22004.
- R package **ShrinkageTrees** on CRAN (and GitHub).
- Questions? t.jacobs@vu.nl.

tijn-jacobs.github.io

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe program under Grant agreement 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

Simulation 1: data-generating process

$n = 200$, censoring $\approx 35\%$, p swept from 50 to 5000:

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

$e(x_i) = \Phi(-0.5 + 0.4x_{i1} - 0.1x_{i3} + 0.3x_{i5})$; $f(x_i) = \beta_f^\top x_i$ with β_f spike-and-slab.

Treatment effect. Friedman non-linear core + sparse main effects + sparse pairwise interactions,

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

with β_τ spike-and-slab ($s_{\mathrm{main}} = 0.05$) and sparse symmetric Γ ($s_{\mathrm{int}} = 0.01$).

Simulation 2: data-generating process

$p = 500$, $n = 100$, $|S| = 5$ strong confounders, $|W| \in \{1, \ldots, 20\}$ weak confounders:

$$
\begin{aligned}
e(x_i) &\propto \tfrac{1}{\sqrt{5+|W|}} \left(\textstyle\sum_{j \in S} x_{ij} + 2 \sum_{j \in W} x_{ij} \right), \\
f(x_i) &= 4 \sum_{j \in S} x_{ij} + \sum_{j \in W} x_{ij}, \\
\tau(x_i) &= 1 + \sum_{j \in S} \beta_j \, x_{ij}.
\end{aligned}
$$

Censoring via data augmentation

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

At iteration t, for each censored i with Y_i the censoring time:

$$\log T_i^{(t)} \;\sim\; \mathcal{N}\big(\hat{\mu}_i^{(t)},\, \sigma^{2,(t)}\big) \quad \text{truncated to } (\log Y_i, \infty),$$

where $\hat{\mu}_i^{(t)} = f^{(t)}(x_i, \hat{e}(x_i)) + a_i \tau^{(t)}(x_i)$.

Propensity score: estimated separately with a (binary-outcome) Horseshoe Forest via probit data augmentation; $\hat{e}(x)$ is then *plugged into* the prognostic forest. Avoids feedback (Zigler, 2016) and improves finite-sample bias (Hahn et al., 2020).

Sparsity on splits can drop real signal

Setup. $n = 500$, $p = 10$; first 5 covariates carry strong signal, rest are noise. Three independent replications shown.