Horseshoe Forests for High-Dimensional Causal Survival Analysis

ISBA Meeting

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

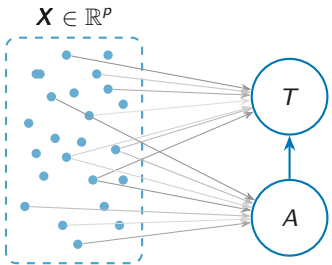
Vrije Universiteit Amsterdam

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?

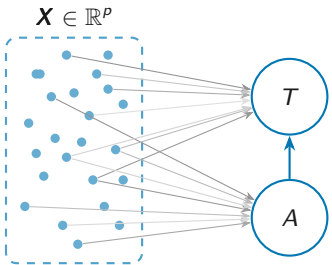
Modern applications, e.g. genomics:

- 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?



Novelties

- **Framework.** We bring the Bayesian Causal Forest (Hahn et al., 2020) to survival data.
- **A new place to regularise a tree ensemble:** a horseshoe prior on the leaf step heights, not the tree structure.
- Non-conjugate $\Rightarrow$ tailored reversible-jump-within-Gibbs sampler.

The setting

Potential outcomes. For treatment $a \in \{0, 1\}$, each patient has an event time $T_i(a)$ and a censoring time $C_i(a)$.

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

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

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

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

Identifiable under standard causal assumptions.

An accelerated failure time decomposition

We model the log-event times with an **accelerated failure time (AFT)** model:

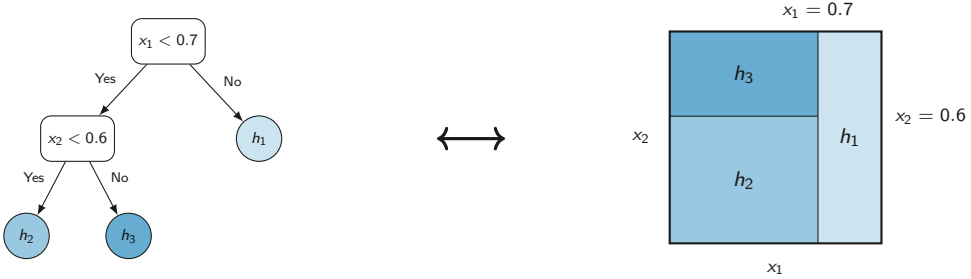
$$\log T(a) \;=\; f\big(x, \hat{e}(x)\big) \;+\; a \cdot \tau(x) \;+\; \varepsilon, \qquad \varepsilon \;\sim\; \mathcal{N}(0, \sigma^2).$$

- 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 [Linero, 2018], 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 $\ell = 1, \ldots, L$:

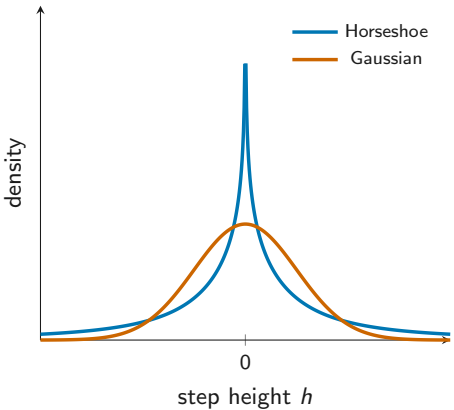
$$h_\ell \mid \lambda_\ell^2, \gamma^2 \;\sim\; \mathcal{N}\big(0,\, \lambda_\ell^2 \gamma^2\big),$$

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

Global–local shrinkage:

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

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

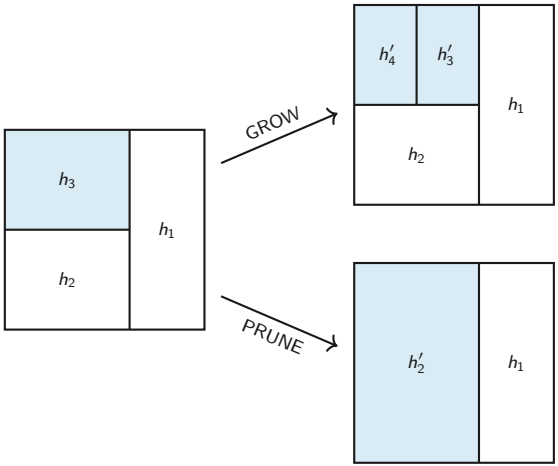


Inference: reversible-jump within Gibbs

Non-conjugate $\Rightarrow$ step heights can't be marginalised out.

GROW / PRUNE change the dimension of $\mathcal{H}_j \Rightarrow$ reversible-jump proposal for $(\mathcal{T}_j, \mathcal{H}_j)$.

Outer Gibbs sampler: f, τ, σ^2, augmented censored times (Tanner & Wong, 1987).

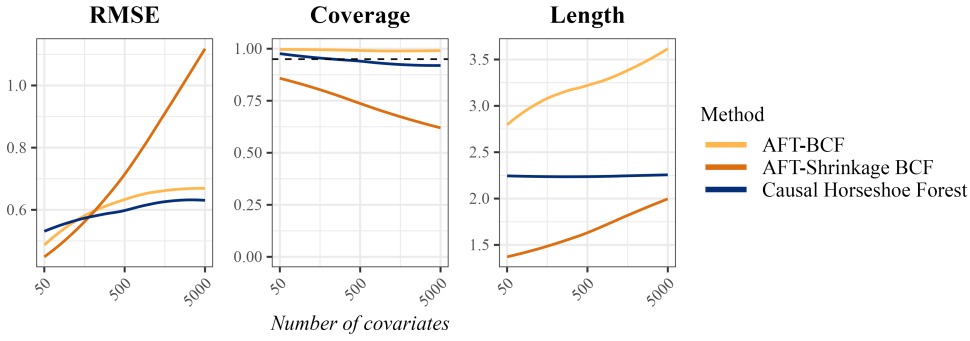


Simulation 1: CATE estimation as p grows

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

Simulation 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

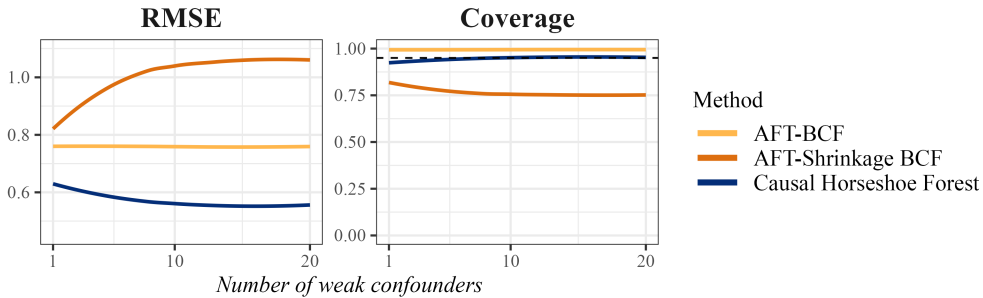
$n = 100$, $p = 500$, censoring $\approx 35\%$. Two groups of confounders:

- $|S| = 5$ **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.

Simulation 2: regularisation-induced confounding

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

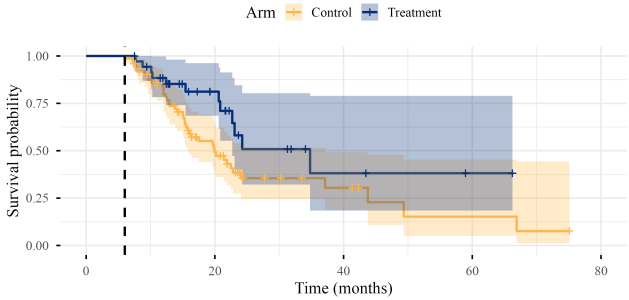
- $|S| = 5$ **strong** confounders — contribute substantially to both treatment and outcome.
- $|W| \in \{1, \ldots, 20\}$ **weak** confounders — predict treatment, but contribute only weakly to the outcome.



Adjuvant radiotherapy in PDAC

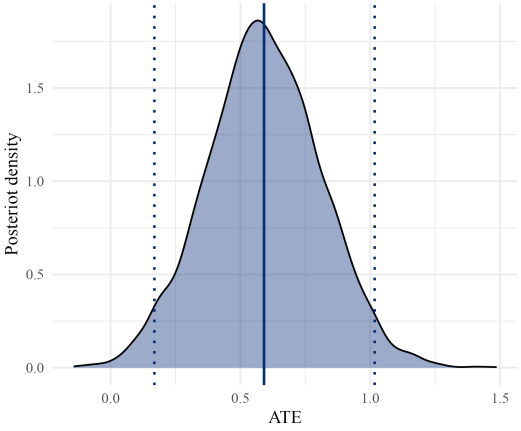
Data. TCGA-PAAD, $n = 110$ resected PDAC patients (6-month landmark).

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

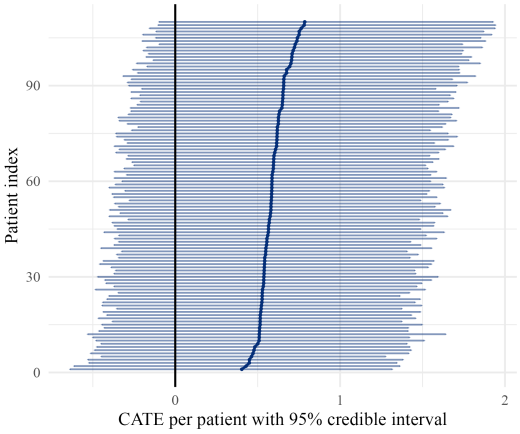


Question. Does adjuvant radiotherapy improve survival, and is the effect heterogeneous?

Adjuvant radiotherapy in PDAC



$\widehat{\text{ATE}} \approx 0.55$, 95% CrI (0.10, 1.02); acceleration factor $e^{0.55} \approx 1.73$.



Predictive performance: concordance index = 0.75.

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.

Thanks for your attention!

Paper

Accepted at

Bayesian Analysis.

R package `ShrinkageTrees`

on CRAN and GitHub.

Stay up to date

`tijn-jacobs.github.io`

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.
- Tanner, M.A., Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *JASA*, 82(398), 528–540.

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

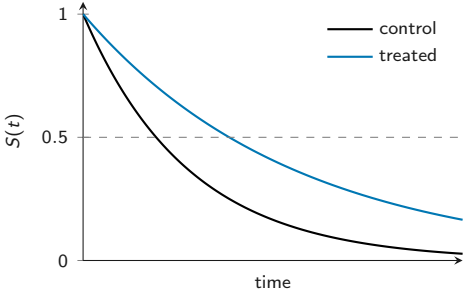
Acceleration factor: an interpretable estimand

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

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

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 "horseshoe"? The shrinkage weight

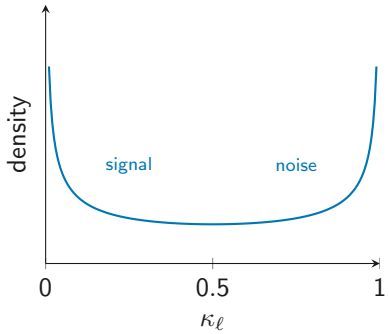
Reparametrise $\kappa_\ell = 1/(1+\lambda_\ell^2\gamma^2) \in [0,1]$. The posterior mean is $\widehat{\theta} = (1-\kappa_\ell)\,\bar{y}$, and under the horseshoe

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

The U-shape puts mass at the extremes:

- $\kappa_\ell \approx 0$: signal (no shrinkage).
- $\kappa_\ell \approx 1$: noise (full shrinkage).

(van der Pas et al., 2014, 2017: near-minimax contraction.)

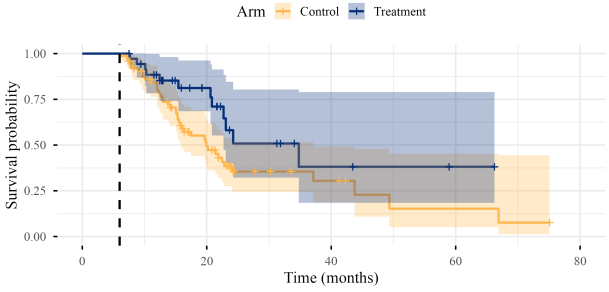


PDAC: the data

TCGA-PAAD cohort:

- $n = 110$ resected patients (6-month landmark).
- 11 clinical + $\sim$3,000 gene expressions.
- Censoring $\approx 47\%$; median follow-up 23.3 mo.

Treatment: adjuvant radiotherapy.



Censoring via data augmentation

We follow Tanner & Wong (1987): treat censored event times as latent variables. At iteration t, for each censored i,

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

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

Propensity score: estimated separately with a binary-outcome Horseshoe Forest (probit augmentation), then plugged into f. Avoids feedback (Zigler, 2016); improves finite-sample bias (Hahn et al., 2020).

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 + sparse interactions,

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

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

Simulation 2: data-generating process

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

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

Full conditional update of step heights

Given the current tree $\mathcal{T}$ with L leaves, encode the partition by $D \in \mathbb{R}^{n \times L}$. The conditional posterior of $h \in \mathbb{R}^L$ is

$$h \sim \mathcal{N}_L\Big(\Omega^{-1} D^\top R,\ \Omega^{-1}\Big), \qquad \Omega = D^\top D + \big(\omega\gamma^2 \operatorname{diag}(\lambda_1^2, \ldots, \lambda_L^2)\big)^{-1}.$$

Coordinate-wise,

$$h_\ell \mid \cdot \ \sim\ \mathcal{N}\Big(\tfrac{\bar{R}_\ell}{n_\ell + 1/(\gamma^2\lambda_\ell^2\omega)},\ \tfrac{\sigma^2}{n_\ell + 1/(\gamma^2\lambda_\ell^2\omega)}\Big).$$

The shrinkage parameters update in closed form via the inverse-gamma auxiliary representation (Makalic & Schmidt, 2016).

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.