Non-parametric Estimation of Causal Effects on Spatially Clustered Survival Data

Your Name

April 15, 2025

Outline

Research Motivation and Context

- **Objective**: Estimate causal effects on survival outcomes using a non-parametric framework.
- **Challenge**: The survival data is *spatially clustered* meaning observations within the same geographical area tend to share unmeasured features.
- **Why Account for Clustering?**
  - Ignoring spatial associations can result in biased treatment effect estimates.
  - Proper clustering improves efficiency by accounting for correlated errors.
- **Non-parametric Rationale**: Methods such as BART and its spatial extension (SBART) naturally capture non-linearities and interactions without a rigid model specification.

References: *Chipman et al. (2010)*; *Hill (2011)*

Why is Clustering Important in Causal Inference?

- **Within-Cluster Homogeneity:**
  - Subjects in the same cluster share unobserved factors (e.g., environmental, institutional).
  - Ignoring this can lead to confounding and biased estimates.
- **Accurate Standard Error Estimation:**
  - Independence assumptions fail within clusters.
  - Clustering adjustment prevents underestimation of standard errors and overconfident inferences.
- **Mitigation of Omitted Variable Bias:**
  - Unobserved cluster-specific confounders can impact both treatment and outcome.
  - Hierarchical models account for these, reducing bias in causal effect estimates.

References: Imbens & Rubin (2015); Wooldridge (2010); Gelman & Hill (2007)

Importance of Clustering in Causal Analysis

- **Unmeasured Commonalities:** Grouped individuals (e.g., same hospital or neighborhood) share latent traits that impact outcomes.
    - *Reference:* Wooldridge (2010); Imbens and Rubin (2015)
- **Statistical Implications:** Ignoring clusters can lead to biased estimates and underestimated standard errors, which adversely affects inference.

Why is Spatial Clustering Important in Causal Inference?

- **Shared Environmental Exposures:**
  - Nearby units often experience similar environmental, socioeconomic, or healthcare conditions.
  - Ignoring spatial proximity may lead to unmeasured confounding.
- **Spatial Autocorrelation:**
  - Outcomes for spatially close observations are correlated.
  - Correctly modeling spatial autocorrelation prevents biased estimates and improves uncertainty quantification.
- **Enhanced Precision:**
  - Accounting for spatial effects increases model efficiency and robustness.
  - Hierarchical spatial models capture latent geographical factors impacting treatment effects.

References: Anselin (1995); Tobler's First Law of Geography; Linero (2020)

Importance of Spatial Dependencies in Causal Inference

- **Environmental Socioeconomic Influences:** Geographic proximity implies similar exposures (e.g., air quality, access to care) that can confound treatment effects.
    - *Reference:* Anselin (1995); Tobler's First Law of Geography.
- **Spatial Random Effects:** Incorporating a spatial component (e.g., via a Gaussian Process) helps account for unobserved spatial heterogeneity, thus improving the validity of causal estimates.
    - *Reference:* Linero (2020)

Advantages of Non-parametric Methods for Causal Inference

- **Flexibility**: Sum-of-trees models adapt to complex relationships between treatment, covariates, and outcomes.
- **Bayesian Framework**: Direct uncertainty quantification and regularization help mitigate overfitting.
- **Spatial Extensions**: SBART augments the standard BART model by incorporating spatial random effects, crucial when data exhibit geographical clustering.
- **Empirical Justification**: Studies show that non-parametric methods outperform classical parametric approaches when dealing with heterogeneous effects and complex interactions.

References: *Chipman et al. (2010)*, *Linero (2020)*

Why Use BART/SBART for Causal Inference?

- **Flexibility:**
  - BART models the regression function as a sum-of-trees, capturing nonlinearities and complex interactions without a fixed parametric form.
  - Uncertainty is quantified via posterior draws, which is essential for reliable causal inference.
- **Handling Spatial Clustering:**
  - SBART extends BART by incorporating spatial random effects (e.g., using a CAR prior) to adjust for correlated outcomes within clusters.
  - This reduces bias and improves the precision of treatment effect estimates in clustered data settings.
- **Empirical Evidence:**
  - *Chipman et al. (2010)* and *Hill (2011)* demonstrate the benefits of BART in causal contexts.
  - *Linero (2020)* shows that accounting for spatial correlations enhances causal effect estimation.

Model for Propensity Score

$$P[Z_{ij} = 1 \mid \mathbf{X}_{ij}, \mathbf{V_i}] = e_{ij}, \quad logit(e_{ij}) = g(\mathbf{X}_{ij}), \text{ where } g(\cdot) \sim BART$$

Two-stage implementation: (i) estimate PS $\hat{e}_{ij}$, (ii) plug in $\hat{e}_{ij}$ into the survival model. (Doubly Robust)

The AFT-BART Model with Spatial CAR Prior

- **Model:**

$$\log T_{ij} = f\Big(Z_{ij}, \mathbf{X}_{ij}, \mathbf{V}_i, \hat{e}(\mathbf{X}_{ij}, \mathbf{V}_i)\Big) + W_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2).$$

- **Spatial Random Effects:**

$$p(W \mid \tau^2, \rho) \propto \exp\Big\{-\frac{1}{2\tau^2}\, W^\top (D - \rho A) W\Big\},$$

where A is the spatial adjacency matrix, D is diagonal with $d_{ii} = \sum_{i'} A_{ii'}$, and ρ is the spatial parameter.

- **BART:**

$$f(Z_{ij}, \mathbf{X}_{ij}, \mathbf{V}_i, \hat{e}(\mathbf{X}_{ij}, \mathbf{V}_i)) = \sum_{h=1}^{H} g(Z_{ij}, \mathbf{X}_{ij}, \mathbf{V}_i, \hat{e}(\mathbf{X}_{ij}, \mathbf{V}_i); \mathcal{T}_h, \mathcal{M}_h).$$

Prior Specification

- Error variance: $\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$.
- Spatial variance: $\sigma_W^2 \sim \mathrm{IG}(a_W, b_W)$.
- Spatial correlations: $\rho \sim \mathit{Uniform}(\frac{1}{\alpha_{(1)}}, \frac{1}{\alpha_{(K)}})$. $\alpha_{(1)}, \alpha_{(K)}$ are the minimum and maximum eigenvalues of A, respectively.
- Function $f(\cdot) \sim$ SBART

Observed Data Likelihood

- For each subject j in cluster i, we observe

$$y_{ij} = \min(T_{ij}, C_{ij}), \quad \delta_{ij} = 1(T_{ij} < C_{ij}).$$

- Define the model mean (on the log-scale) as

$$\mu_{ij} = f(Z_{ij}, \mathbf{X}_{ij}, \mathbf{V}_i, \hat{e}(\mathbf{X}_{ij}, \mathbf{V}_i)) + W_i.$$

- Then the individual likelihood contribution is

$$L_{ij}^{\text{obs}}(\theta) = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log y_{ij} - \mu_{ij})^2}{2\sigma^2} \right] \right\}^{\delta_{ij}} \left\{ 1 - \Phi\left(\frac{\log y_{ij} - \mu_{ij}}{\sigma} \right) \right\}^{1}$$

where $\Phi(\cdot)$ is the standard normal CDF.

- The full observed-data likelihood is

$$L_{\text{obs}}(\theta) = \prod_{i=1}^{K} \prod_{j=1}^{n_i} L_{ij}^{\text{obs}}(\theta).$$

Data Augmentation

- Define the latent log survival time $\tilde{y}_{ij}$ as

$$\tilde{y}_{ij} = \begin{cases} \mathrm{TruncNormal}\left(\mu_{ij}, \sigma^2; \log y_{ij}\right), & \text{if } \delta_{ij} = 0, \\ \log y_{ij}, & \text{if } \delta_{ij} = 1. \end{cases}$$

Here, TruncNormal$(\mu, \sigma^2; a)$ denotes a $N(\mu, \sigma^2)$ distribution truncated to the interval (a, ∞). The imputed values are used in the complete-data likelihood.

Complete Data Likelihood

- Introduce the latent (complete) log survival times:

$$\tilde{y}_{ij} = \begin{cases} \log y_{ij}, & \delta_{ij} = 1, \\ \text{draw from } N(\mu_{ij}, \sigma^2) \text{ truncated to } [\log y_{ij}, \infty), & \delta_{ij} = 0. \end{cases}$$

- With μ_{ij} defined as before, the complete-data likelihood is

$$L_{\text{complete}}(\theta) = \prod_{i=1}^{K} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\tilde{y}_{ij} - \mu_{ij})^2}{2\sigma^2} \right].$$

Algorithm 1: A Single Iteration

1. **Update Spatial Random Effects & Variance:** Update W, τ^2, and ρ from their full conditionals based on the CAR prior.
2. **Impute Censored Data:** For subjects ij, sample the latent log survival time as

$$\tilde{y}_{ij} = \begin{cases} \text{TruncNormal}\Big(\mu_{ij}, \sigma^2; \log y_{ij}\Big), & \text{if } \Delta_{ij} = 0, \\ \log y_{ij}, & \text{if } \Delta_{ij} = 1. \end{cases}$$

3. **Update BART:** With responses $\tilde{y}_{ij} - W_i$ and covariates $(z_{ij}, \mathbf{x}_{ij})$, update the BART parameters $\{\mathcal{T}_h, \mathcal{M}_h\}$ and the error variance σ^2 via Bayesian backfitting.

Conclusion and Future Directions

- **Summary**: We presented a non-parametric approach combining BART with a spatial Gaussian Process component to estimate causal effects in spatially clustered survival data.
- **Benefits**:
  - Captures complex, non-linear interactions without requiring pre-specified model forms.
  - Incorporates spatial dependence to reduce bias and improve uncertainty quantification.
- **Future Work**:
  - Extend the approach to incorporate time-varying covariates.
  - Enhance computational efficiency for larger datasets.

Thank You

Questions?

References I

- Chipman, H.A., George, E.I. and McCulloch, R.E., 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), pp.266–298.

- Hill, J., 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), pp.217–240.

- Linero, A.R., 2020. Bayesian additive regression trees for spatial data. *Journal of the American Statistical Association* (reference details as appropriate).