Bayesian Causal Inference for Uplift Modeling

Jeong In Lee [1] Seong Il Jo [1]

[1]Dept. of Statistics, Inha University

I. Abstract

- **Goal** — estimate individual uplift under strong, feature-driven ad selection.
- **Issue** — most flexible regularized models suffer from regularization-induced confounding (RIC) bias.
- **Approach** — Include the estimated propensity score or re-parameterize the model(Bayesian Causal Forest).
- **Practical Studies: Criteo data** — BCF allows reliable measurement of conversion uplift as a function of the targeting fraction.

II. Introduction

2.1 Motivation

- In the digital advertising environment, accurate target selection drives real advertising impact and maximizes ROI.
- Causal uplift refers to the expected difference in outcomes depending on whether an individual receives a treatment. Accurate target selection should be guided by this insight.
- Click attribution or simply estimating the Average Treatment Effect (ATE) is not sufficient.

  ATE $:= \mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0), \quad Z =$ treatment
- Estimating the Conditional Average Treatment Effect (CATE) is necessary to uncover true causal effects for each subgroup.

  CATE $:= \mathbb{E}(Y_i \mid X_i = x,\, Z_i = 1) - \mathbb{E}(Y_i \mid X_i = x,\, Z_i = 0)$
- Flexible models to estimate effect require regularization, but under strong confounding that regularization can itself introduce large and uncontrolled bias, so specialized methods are needed.

2.2 BART: Bayesian additive regression trees

- BART, **widely used in causal inference**, has proven practical strength[1, 2, 3]:
  - detecting complex interactions and sharp breaks
  - being invariant to monotone transformations of covariates
  - requiring minimal tuning, and repeatedly outperforming in causal-effect benchmarks
- BART expresses an unknown function $f(x)$ as a sum of piecewise constant binary regression trees. each tree T_l cuts the covariate space into axis-aligned cells $A_b^{(l)}$; the induced step function is $g_l(x) = m_{lb}$ if $x \in A_b^{(l)}$.

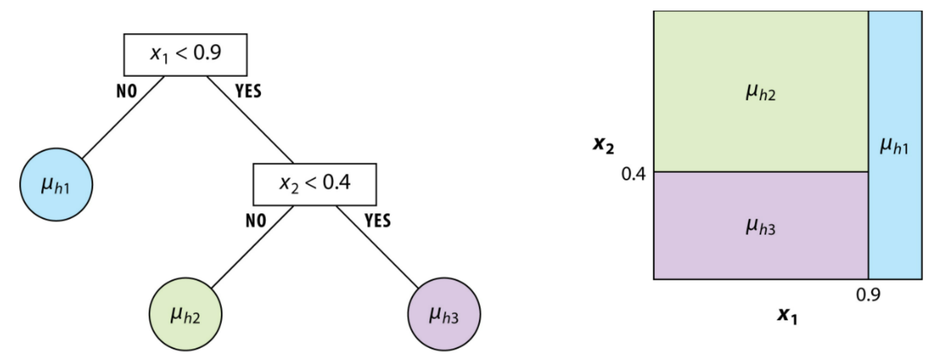

Figure 1. An example binary tree

- **Additive ensemble**: unknown response surface as L weak trees

$$f(x) = \sum_{l=1}^{L} g_l(x), \qquad g_l(x) = m_{lb} \ \text{ if } x \in A_b^{(l)}.$$

- **Tree-structure prior**: node at depth h splits with $\Pr(\text{split}|h) = \eta(1+h)^{-\beta}$ (default $\eta = 0.95,\ \beta = 2$) $\Rightarrow$ small, shallow trees preferred.
- **Leaf-parameter prior**: $m_{lb} \sim \mathcal{N}(0, \sigma_m^2),\ \sigma_m = \sigma_0/\sqrt{L}$ $\Rightarrow$ 95 % of prior mass for $f(x)$ lies in $\pm 2\sigma_0$ (pointwise).
- **Causal target**: individual treatment effect

$$\tau(x) = f(x, z=1) - f(x, z=0).$$

III. Methodology

3.1 Regularization Induced Confounding (RIC)

- under strong confounding that regularization can itself introduce large bias so specialized methods are needed.
- **Setting** : linear ridge example

$$Y_i = \tau Z_i + \beta^\top X_i + \varepsilon_i, \qquad Z_i = \gamma^\top X_i + \nu_i.$$

  Gaussian ridge prior: $(\tau, \beta) \sim \mathcal{N}(0, M^{-1})$.
- **Bias of ridge/Bayes estimator**

$$\text{bias}(\hat{\tau}_{rr}) = -\left[(Z^\top Z)^{-1} Z^\top X\right] \left(I + X^\top (X - \hat{X}_Z)\right)^{-1} \beta,$$
$$\hat{X}_Z = Z(Z^\top Z)^{-1} Z^\top X.$$

- **Implications**
  - Term $\left[(Z^\top Z)^{-1} Z^\top X\right] \neq 0$ if $Z \not\perp X$ (*confounding*).
  - Posterior variance of τ also shrinks $\Rightarrow$**Credible-interval coverage** $< 95\%$
- **Extension to BART** Trees prefer a single split on Z (cheap) over many splits on $X \to$ variability of $\mu(X)$ falsely attributed to $Z \to$ *RIC in nonlinear models*.

3.2 PS-BART & BCF

- **Propensity-Score BART (PS-BART)**

$$Y_i \sim \sum_{j=1}^{m} g_j(X_i, \hat{\pi}(X_i), T_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  Adding the estimated propensity score $\hat{\pi}(X)$ decorrelates Z and X; in linear models $\text{bias}(\hat{\tau}) = 0$, sharply reducing RIC.
- **Bayesian Causal Forest (BCF)**

$$Y_i = \mu(X_i, \hat{\pi}_i) + T_i\, \tau(X_i) + \varepsilon_i, \qquad (\mu, \tau) \sim \text{BART}$$

  - $\mu(X, \hat{\pi})$: *prognostic surface* – expected outcome under control.
  - $\tau(X)$: *treatment-effect function* – incremental uplift for covariates X.
  - Independent priors let us shrink $\tau(X)$ strongly; when $\hat{\pi}$ is extreme the prior pulls $\tau \to 0$, eliminating RIC bias and restoring correct 95% coverage.

3.3 PS-BART & BCF can mitigate problems

- **Add** $\hat{\pi}(X)$ **as a covariate** ($\hat{e}(X)$:propensity score)

$$\tilde{X} = \left[\, Z \ \ \hat{\pi}(X) \ X \,\right],$$

  then $Z \perp X \mid \hat{\pi}(X) \Rightarrow \left[(\tilde{Z}^\top \tilde{Z})^{-1} \tilde{Z}^\top X\right]_{(Z)} = 0$
  $\Rightarrow$ **Bias vanishes**; tree-based models penalize splits on Z and $\hat{\pi}(X)$ equally.
- **BCF** re-parameterizes $Y = \mu(X, \hat{\pi}(X)) + Z\,\tau(X)$ with *separate* priors:
  - Stronger shrinkage on $\tau(X)$ (fewer trees, deeper penalty). And μ captures prognostic part; miss?allocation to Z discouraged.

III. Simulation Studies

3.1 Data-Generating Process

- Covariates: $x_1, x_2, x_3 \overset{i.i.d}{\sim} \mathcal{N}(0,1),\ x_4 \sim \text{Ber}(0.5),\ x_5 \sim \text{Unif}\{1,2,3\}$.
- Treatment effect(homoheneous) : $\tau(x) = 3$.
- Treatment effect(heterogeneous) : $\tau(x) = 1 + 2x_2 x_5$.
- Prognostic surface (non-linear): $\mu(x) = -6 + g(x_5) + 6|x_3 - 1|$.
  Category map g: $g(1) = 2,\ g(2) = -1,\ g(3) = -4$.
- Propensity: $\pi(x) = 0.8\,\Phi\!\left(\frac{3\mu(x)}{s} - 0.5x_1\right) + 0.05 + \frac{u}{10},\ u \sim \text{Unif}(0,1),\ s = \text{sd}\{\mu(x_i)\}_{i=1}^n$.
- Sample size: $n = 250$ (single scenario, non-linear μ).

3.2 Simulation results

- Under strong confounding, feeding the estimated propensity score into the outcome model improves performance.
- In particular, BCF (Bayesian Causal Forest) is better when treatment effects are heterogeneous.

Table 1. CATE performance

	Homogeneous effect			Heterogeneous effect		
Method	rmse	cover	len	rmse	cover	len
BART	1.20	0.90	4.1	1.8	0.87	5.2
PS-BART	1.00	0.96	4.3	1.7	0.91	5.4
BCF	0.63	0.94	2.5	1.3	0.93	4.5

IV. Uplift Modeling under Strong Confounding

4.1 Data

- 5,000 users from Criteo incrementality tests (84% treated).
- Variables:
  - **anonymized user features**: $\{f_0, \ldots, f_{11}\}$ (dense floats)
  - **treatment**: random-assignment flag
  - **exposure**: ad shown
  - **conversion**: binary outcome
- Exposure is served mainly to treated users with high conversion propensity $\Rightarrow$ **strong confounding**

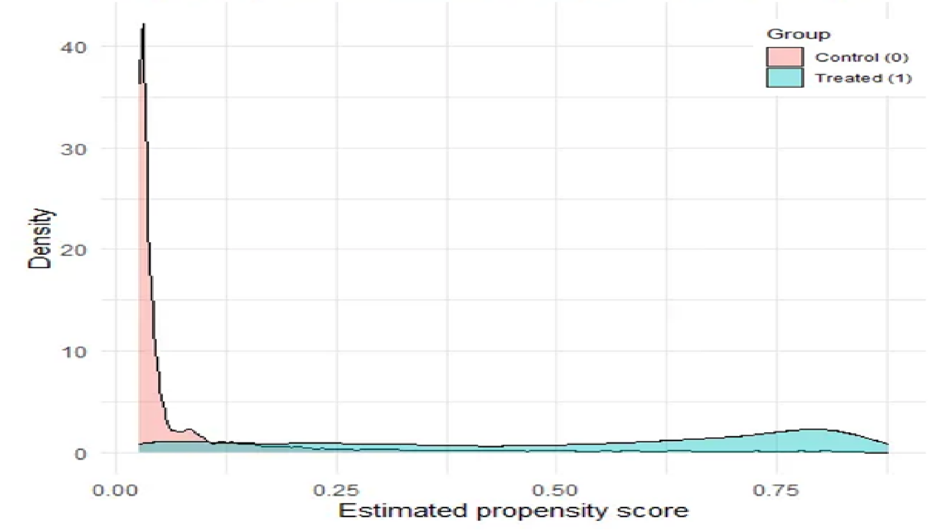

Figure 2. Propensity score distribution

4.2 Results

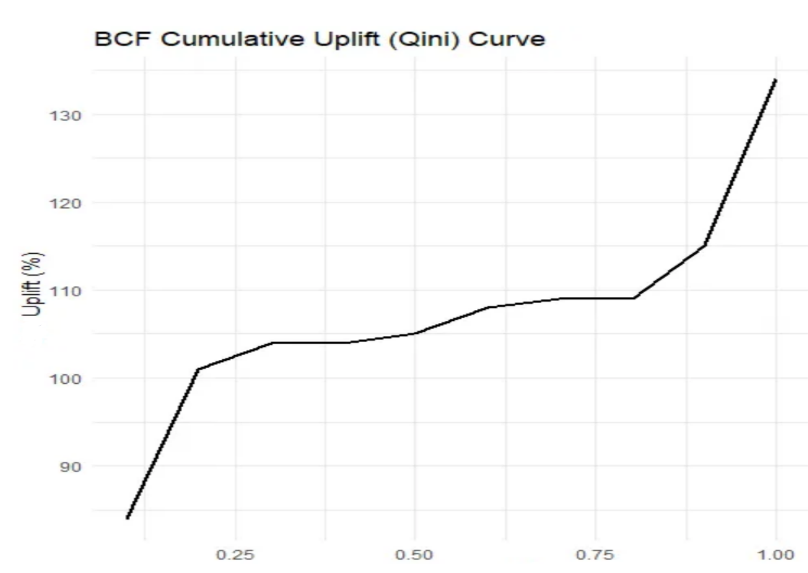

Figure 3. BCF cumulative uplift (Qini) curve

- **Top-decile impact:** Targeting the top 10% yields an observed uplift of ≈8.9%, over three times the overall average (2.7%).
- **Diminishing returns:** As the targeted pool expands (20–80%), additional uplift contributions plateau around 2–4%.
- **Prediction gap:** A noticeable divergence between predicted and observed uplift at high deciles suggests room for model calibration.

VI. References

[1] H. A. Chipman, E. I. George, and R. E. McCulloch, “Bart: Bayesian additive regression trees,” *The Annals of Applied Statistics*, vol. 4, Mar. 2010.

[2] V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone, “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition,” 2018.

[3] P. R. Hahn, J. S. Murray, and C. Carvalho, “Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects,” 2019.