Subgroup Detection and Policy Estimation

Antonio Linero

Subgroup Detection

Motivation: Canagliflozin

- **Study focus**: Clinical trial investigating canagliflozin efficacy for type 2 diabetes mellitus (T2DM)
- **Proven benefits**: Canagliflozin reduces cardiovascular and renal event risk in T2DM patients
- **Key challenge**: Preliminary trials revealed heterogeneous treatment effects across patient subpopulations
- **Clinical objective**: Identify patient subgroups with different treatment responses and understand underlying causes
- **Study design advantages**:
  - Prespecified subgroup analyses included for comparison with data-adaptive approaches
  - Randomized design ensures ignorable assignment mechanism
- **Broader applicability**: Methodology extends beyond randomized trials to observational studies

Formal Framework: Subgroup Detection

Let $\tau(x)$ denote the treatment effect function and define $\tau(G) = \frac{1}{N_G} \sum_{i:X_i \in G} \tau(X_i)$ to be the *subgroup treatment effect*.

We are interested in identifying *subgroups* $G_1, \ldots, G_K$ of the X_i's such that the the treatment effects $\tau(G_1), \ldots, \tau(G_K)$ are "different."

A potential strategy is to maximize the *utility*

$$U(G) = \sum_{i=1}^{N} \{\tau(G_{(i)}) - \tau\}^2$$

Bayes: Maximize Posterior Expectation

where $G_{(i)}$ is the class that X_i belongs to.

This Utility Leads to *Risk Seeking*

Proposition 1 (Posterior Expectation of RS Utility). *Under the utility function* (1), *the expected utility* $R(G) = \mathbb{E}\{U(G, \theta) \mid \mathcal{D}\}$ *is given by*

$$R(G) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i:X_i \in G_k} \{\widehat{\tau}(G_k) - \widehat{\tau}(\mathcal{X})\}^2 + \mathrm{Var}\{\tau(X_i) - \tau(G_k) \mid \mathcal{D}\} \tag{2}$$

$$= \mathrm{const}(\mathcal{D}) + \frac{1}{N} \sum_{k=1}^{K} \sum_{i:X_i \in G_k} \mathrm{Var}\{\tau(G_k) \mid \mathcal{D}\} - \{\widehat{\tau}(X_i) - \widehat{\tau}(G_k)\}^2 \tag{3}$$

where $\widehat{\tau}(G) = \mathbb{E}\{\tau(G) \mid \mathcal{D}\}$ *and* $\mathrm{const}(\mathcal{D})$ *is a constant independent of* G*.*

Multi-Stage Utility

We can construct a utility function that is instead *risk averse* by accounting for inaccuracy in our estimates in a subsequent study. Consider the following workflow:

1. We are conduct an initial experiment to assess an overall treatment effect $\tau(\mathcal{X}) = \frac{1}{N}\sum_i \tau(X_i)$ and, as a secondary analysis, we will produce the subgroups G_k as well as predictions t_k for the average effect within each of these subgroups.

2. Based on the recommended subgroups, a follow-up study will be performed to verify the treatment effect estimates within each group

Multi-Stage Utility

A natural utility that captures this scenario, which now requires both selecting subgroups and estimating their treatment effects, is

$$U(G,t,\theta)=\frac{1}{N}\sum_{k=1}^{K}\sum_{i:X_i\in G_k}\{\tau(G_k)-\tau(\mathcal{X})\}^2-\lambda\frac{1}{N}\sum_{k=1}^{K}\sum_{i:X_i\in G_k}\{\tau(G_k)-t_k\}^2 \tag{4}$$

where λ is a tuning parameter used to balance the importance of finding heterogeneous subgroups on the one hand and being able to estimate the parameters on the other. We refer to (4) as the *multi-stage utility*, as its construction is motivated by considering both the current study and a hypothetical follow-up study. Below, we give the expected utility associated with this utility function.

Multi-Stage Utility Allows Risk-Neutral/Averse

Proposition 2 (Posterior Expectation of the Multi-Stage Utility). *Under the utility function (4), the expected utility is given by*

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i:X_i\in G_k}(1-\lambda)\operatorname{Var}\{\tau(G_k)\mid\mathcal{D}\}-\lambda\{\widehat{\tau}(G_k)-t_k\}^2-\{\widehat{\tau}(X_i)-\widehat{\tau}(G_k)\}^2, \tag{5}$$

up-to a constant. This is maximized in $(t_1,\dots,t_K)$ *when* $t_k=\widehat{\tau}(G_k)$ *at*

$$R(G)=const(\mathcal{D})+\frac{1}{N}\sum_{k=1}^{K}\sum_{i:X_i\in G_k}(1-\lambda)\operatorname{Var}\{\tau(G_k)\mid\mathcal{D}\}-\{\widehat{\tau}(X_i)-\widehat{\tau}(G_k)\}^2. \tag{6}$$

Results for Risk Seeking/Averse/Neutral

Variable	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$
Age	-2.56 (10)	-3.26 (10)	-3.96 (9)
Sex	-2.25 (8)	-3.05 (8)	-3.84 (7)
Race	-2.34 (9)	-3.20 (9)	-4.06 (10)
Baseline HBA_{1c}	-1.31 (5)	-2.38 (4)	-3.45 (4)
Age $\times$ Sex	-1.53 (6)	-2.55 (6)	-3.56 (5)
Age $\times$ Race	-1.78 (7)	-2.85 (7)	-3.91 (8)
Age $\times$ Baseline HBA_{1c}	-0.73 (3)	-2.05 (3)	-3.37 (3)
Sex $\times$ Race	-1.27 (4)	-2.47 (5)	-3.66 (6)
Sex $\times$ Baseline HBA_{1c}	-0.42 (2)	-1.85 (2)	-3.27 (1)
Race $\times$ Baseline HBA_{1c}	-0.31 (1)	-1.81 (1)	-3.32 (2)

Note: These were prespecified groups

Learned Subgroups For Risk-Neutral Setting

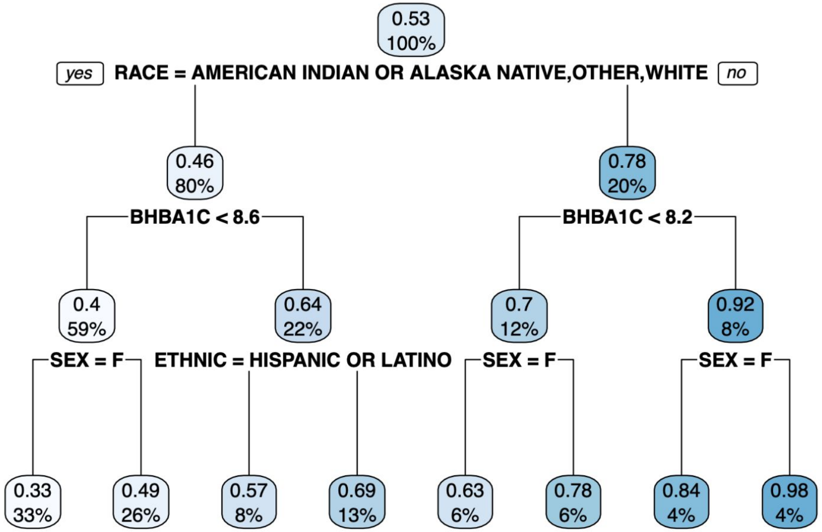


Figure 5: Posterior summarization of the treatment effect heterogeneity using the Bayesian causal forest model (10) with $\lambda = 1$.

Motivation: Policy Estimation

Motivation (Via Athey and Wager)

Often we want to estimate a *coarse* policy for assigning treatments:

- Budgetary/Implementation constraints
- Fairness/equity
- Interpretability

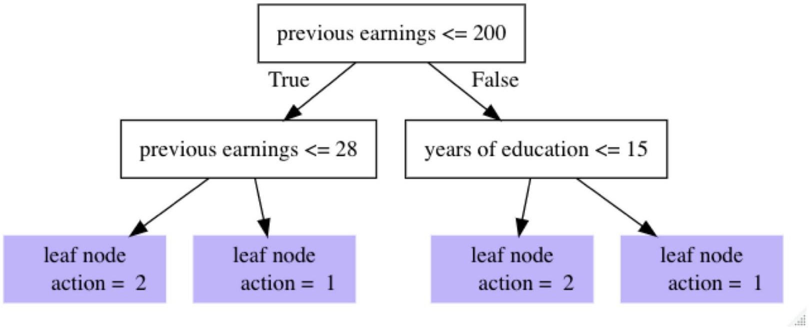


Figure 1: A depth 2 tree fit on data from the National Job Training Partnership Act Study (Bloom et al., 1997). The reward matrix contains two outcomes: not assigning treatment (action 1), and assigning treatment, a job training program (action 2). The covariate matrix contains two variables: a candidate's previous annual earnings in \$1,000 and years of education. Note: the optional package `DiagrammeR` is needed to plot trees.

Figure from Sverdrup et al.

Formal Framework

Let $\rho : \mathcal{X} \to \{0, 1\}$ be constrained to a class $\mathcal{P}$ of shallow binary decision trees. Let η denote a "cost" of applying the treatment. Define an expected utility of the policy as

$$\begin{aligned} U(\rho, \mathbb{P}) &= \mathbb{E}[\rho(X)\{Y(1) - \eta\} + \{1 - \rho(X)\}Y(0)] \\ &= C + \mathbb{E}[\rho(X)\{\tau(X) - \eta\}] \end{aligned}$$

Goal: select the optimal policy

$$\rho^\star = \arg\max_{\rho \in \mathcal{P}} \mathbb{E}[\rho(X)\{\tau(X) - \eta\}]$$

Doubly Robust Policy Estimation (Athey and Wager)

Define pseudo-outcomes

$$Z_i = \widehat{\mu}_1^{(-i)}(X_i) - \widehat{\mu}_0^{(-i)}(X_i) + \{Y_i - \widehat{\mu}_{A_i}^{(-i)}(X_i)\} \times \frac{\{A_i - \widehat{e}^{(-i)}(X_i)\}}{\widehat{e}^{(-i)}(X_i)\{1 - \widehat{e}^{(-i)}(X_i)\})}$$

where $(-i)$ means "use subset of data not including i". Then, set

$$\widehat{\rho} = \arg\max_{\rho \in \mathcal{P}} \sum_i \rho(X_i)(Z_i - \eta)$$

doable in $\approx O(N^d P^d)$ time

Ostensibly has better properties than the naive plug-in estimator

$$\widehat{\rho}_{\text{plug-in}} = \arg\max_{\rho \in \mathcal{P}} \sum_i \rho(X_i)\{\widehat{\tau}(X_i) - \eta\}$$

What Should a Bayesian Do?

Maximize expected utility wrt prior: Under Bayesian bootstrap prior for $\mathbb{P}_X$

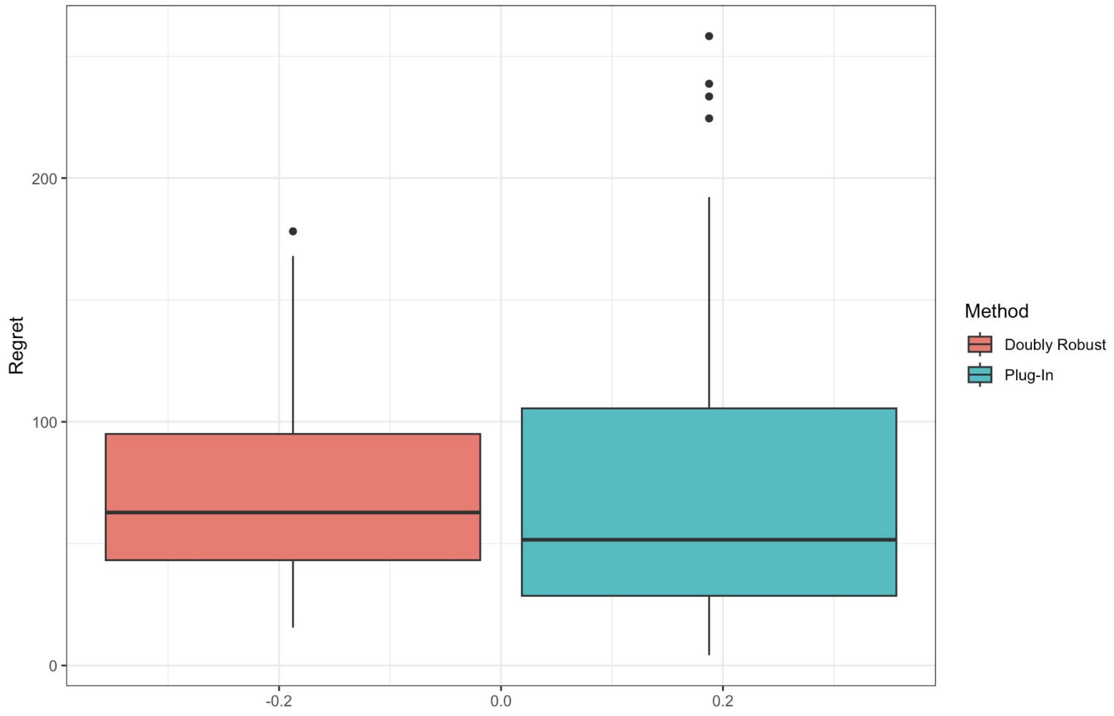
$$U(\rho) = \mathbb{E}\{U(\rho, \mathbb{P}) \mid \mathcal{D}\} = \frac{1}{N} \sum_i \rho(X_i)\{\widehat{\tau}(X_i) - \eta\}$$

where $\widehat{\tau}(x)$ is the Bayes estimator of $\tau(x)$. So

$$\arg\max_\rho U(\rho) = \widehat{\rho}_{\text{plug-in}}$$

What I believe: if we are competent in our prior choice, the plug-in estimator will work just as well.

Somewhat Lower Regret, Somewhat Higher Spread



Based on Causal RF

Post-Selection Inference and the Winner’s Curse

A Funny Thing About Bayesian Inference…

What if we want uncertainty quantification for the subgroup means?

- **Frequentist: "**This is a post-selection inference problem. Inference in treated group is optimistic. You should probably data split."
- **Bayesian:** "The posterior is the posterior…"

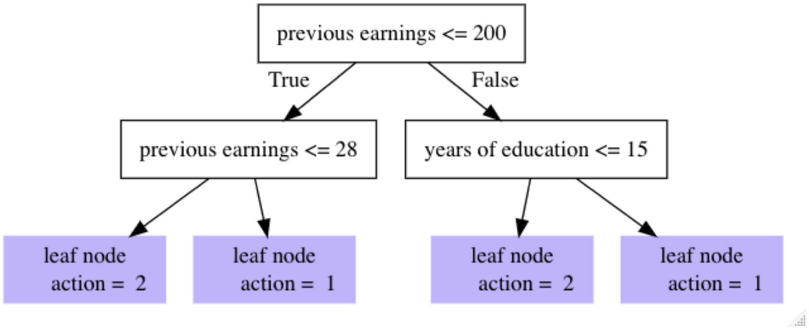


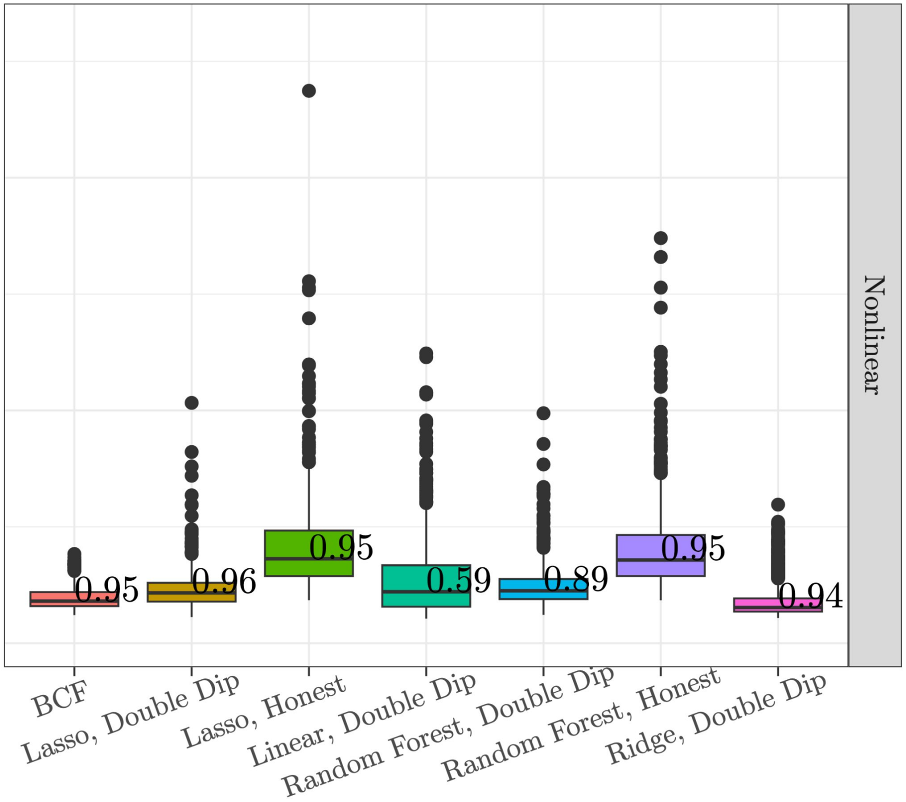
Figure 1: A depth 2 tree fit on data from the National Job Training Partnership Act Study (Bloom et al., 1997). The reward matrix contains two outcomes: not assigning treatment (action 1), and assigning treatment, a job training program (action 2). The covariate matrix contains two variables: a candidate's previous annual earnings in $1,000 and years of education. Note: the optional package `DiagrammeR` is needed to plot trees.

What Should I Do?

- *You can't just ignore selection!*
- *Your coverage will be all wrong!*
- *No reason Bayes should cover!*
- *Confidence and credible intervals represent different types of uncertainty!*

- *Don't you trust your prior?*
- *The procedure **must** be accurate/conservative for some ground truths!*
    - *Otherwise coverage wouldn't average to the nominal rate over the prior!*
- *Just choose the prior to make the actual ground truth one of the good ones!*
- ***You get to keep all the data for inference!***

What If We Try It? (Blessings of Dishonesty)



Inference on Selected Groups: Canagliflozin

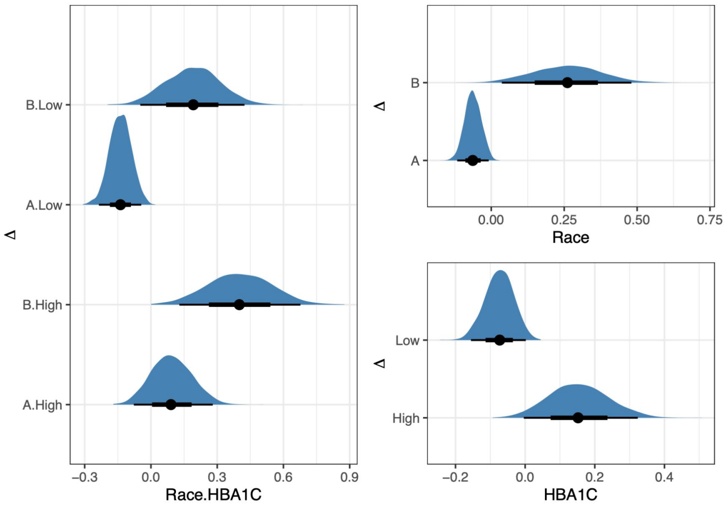


Figure 7: Posterior density for the deviation Δ from the overall/average treatment effect within the data-adaptive subgroups from first and second-level child nodes in Figure 5, where the subgroup splits are based on only Race (top-right), only HBA1C (bottom-right), or both (left). Race is divided into group A (White, Other and Native American) and group B (Black, Asian, Pacific Islander, and Multi-Racial), while HBA_{1c} is divded according to < 8.4 or > 8.4.

Go Over Code