Accelerated Multinomial Probit Bayesian Additive Regression Trees

Yizhen Xu
Advisor: Joseph Hogan

Brown University

July 28th, 2019

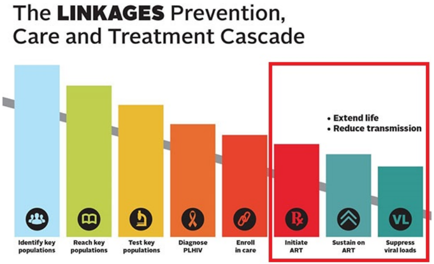
Motivating Work

- Bayesian modeling of state transitions over time under different dynamic regimes
- Causal inference using G computation algorithm (GCA)
  - “ What would have happened if the target population followed a certain regime over time?”
  - Requires correct specification of predictive models
  - Incorporate Bayesian additive regression trees (BART) as predictive models
- Challenge: fitting multinomial probit BART (MPBART) for outcome models

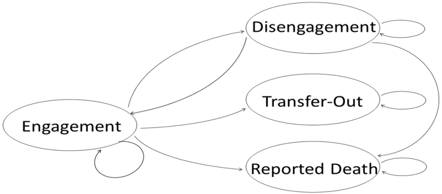
Motivating Work

From

`http://health2615.rssing.com/chan-17973612/all$_$p5.html`



Operationalized **outcome** progression through the HIV care cascade:



- Data: EHRs from AMPATH
- S: Outcome
  $S \in \{0$ Disengaged, 1 Engaged, 2 Transferred, 3 Died$\}$
- A: Treatment status
- X: Time varying confounders
- V: Baseline covariates

Data Excerpt

		S	A	X		V									
myID	Time	Outcome	onARV	CD4 Update	Log CD4+1	Age	Male	Year Enrol	Travel Time	WHO Stage	Married	Height	Log Weight	Log VL+1	VL0
34	0	1	0	1	6.293	33.421	0	2008	3	2	0	163	3.738	NA	0
34	200	1	0	0	6.293	33.421	0	2008	3	2	0	163	3.738	NA	0
34	400	2	0	0	6.293	33.421	0	2008	3	2	0	163	3.738	NA	0
50001	0	1	0	1	2.833	33.927	0	2011	2	4	0	NA	NA	NA	0
50001	200	1	1	0	2.833	33.927	0	2011	2	4	0	NA	NA	NA	0
50001	400	3	1	0	2.833	33.927	0	2011	2	4	0	NA	NA	NA	0
60050	0	1	0	1	3.611	22.828	0	2012	2	NA	0	NA	3.871	NA	0
60050	200	0	0	0	3.611	22.828	0	2012	2	NA	0	NA	3.871	NA	0
60050	400	1	1	0	3.611	22.828	0	2012	2	NA	0	NA	3.871	NA	0
60050	600	1	1	1	3.829	22.828	0	2012	2	NA	0	NA	3.871	NA	0
60050	800	0	1	0	3.829	22.828	0	2012	2	NA	0	NA	3.871	NA	0
60050	1000	0	1	0	3.829	22.828	0	2012	2	NA	0	NA	3.871	NA	0
60050	1200	0	1	0	3.829	22.828	0	2012	2	NA	0	NA	3.871	NA	0

Application goal: Evaluate the causal effectiveness of different HIV treatment initiation policies on the progression of **patients retention and survival** through the HIV care cascade.

Causal structural model to compare treatment policies

- **Structural model**

$$\begin{aligned} \boldsymbol{S}_1 &= \text{state membership at time 1} \\ A_0 &= \text{treatment assigned at time 0} \\ a_0^q &= q(X_0, V) \text{ where } q \text{ is a regime function} \\ P(\boldsymbol{S}_1^q) &= \text{distribution of } \boldsymbol{S}_1 \text{ under regime } q \end{aligned}$$

- For two different regimes q_1 and q_2 at time 1, we want to compare

$$P(\boldsymbol{S}_1^{q_1}) \qquad \text{and} \qquad P(\boldsymbol{S}_1^{q_2})$$

- Example: ‘treat immediately’ is the regime

$$q \equiv 1 \quad \Rightarrow \quad \overline{a}_K^q = (1, 1, 1, \ldots, 1)$$

GCA: Use Observed-data Models as Plug-ins

Target: $P(\boldsymbol{S_1^q})$

$$P(S_1^q) = \int P(S_1|A_0 = a_0^q, X_1, X_0, V)$$
$$P(X_1|A_0 = a_0^q, X_0, V)$$
$$P(X_0, V)$$
$$d(X_1, X_0, V)$$

With certain assumptions (causal network, GCA assumptions, predictive models),

1. Plug in fitted models for (X_1, S_1): $P(X_1|A_0, X_0, V; \gamma)$, $P(S_1|A_0, X_1, X_0, V; \theta)$
2. Fix treatment a_0^q under regime q
3. Average over the empirical baseline distribution of specific population of interest

Focus: BART for Multinomial Models

The GCA can be extended to longitudinal data with discrete time (Young et al. 2011); here we focus on outcome models at each time k:

$$P(S_k | \overline{A}_{k-1}, \overline{X}_k, \overline{S}_{k-1}, V; \theta)$$

Two predominant ways for fitting multinomial outcomes:

- Multinomial probit (MNP) (Imai and van Dyk 2005)
- Multinomial logistic (MNL)

Focus: BART for Multinomial Models

Under the framework of latent variable model for outcome $S \in \{0, 1, 2, 3\}$, when 0 is the reference level,

$$S = \begin{cases} k & \text{if } \max(W_1, W_2, W_3) = W_k > 0 \\ 0 & \text{if } \max(W_1, W_2, W_3) < 0, \end{cases}$$

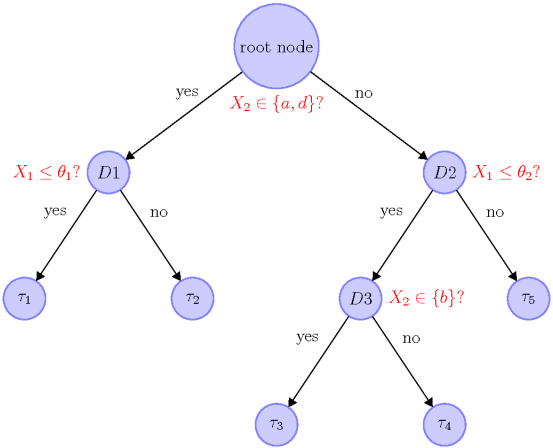
latent utilities $(W_1, W_2, W_3) = (G_1, G_2, G_3) + \epsilon$, where $G_j(X; \theta) = X\theta_j$,

- MNP: $\epsilon \sim MVN(\mathbf{0}, \Sigma)$
- MNL: $\epsilon_k \sim Logistic(0, 1)$ for $k = 1, 2, 3$

Focus: BART for Multinomial Models

- MPBART (Kindo et al 2016): $G_j(X; \theta) = \sum_k g(X; \theta_{jk})$ sum of binary trees
- Binary trees $g(\cdot; \theta_{jk})$

7. FIGURES



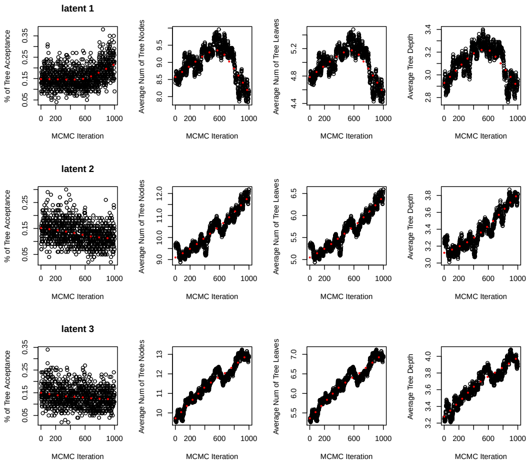
Challenges

- Sensitive to choice of reference level
- Fail to achieve MCMC convergence under unbalanced categories

Solution: Sample the sum-of-trees based on latent utilities W under a constraint on the covariance matrix Σ

Challenges

Diagnostic plots of MPBART (Kindo et al 2016) for $P(S_3 | X_3, \mathcal{F}_2, \theta)$



MPBART

- Correlation among alternatives is captured by Σ
- Identifiability issue: for a constant $\alpha > 0$, **unconstrained** latent utilities

$$\tilde{W} = \alpha W \sim MVN(G(X;\tilde{\theta}), \tilde{\Sigma}), \quad \text{where}$$
$$G(X;\tilde{\theta}) = \alpha G(X;\theta) \quad \Rightarrow \quad \tilde{\theta} = \alpha\theta \text{ for MNP}$$
$$\tilde{\Sigma} = \alpha^2\Sigma$$

$\Rightarrow S(W) = S(\tilde{W})$.

- **Constraint** on latent utilities W: trace$(\Sigma) = C - 1$, where C is the number of categories
- Sample α jointly as a **working parameter** (marginal augmentation)

MPBART

For any variable θ:

- $\tilde{\theta}$ - unconstrained counterpart;
- θ^* - intermediate draw.

Gibbs sampling of (W, θ, Σ)

Algorithm 1 (Kindo et al 2016):

1. Sample $W, \alpha^* | S, \mu, \Sigma \quad \Rightarrow \quad \tilde{W} = \alpha^* W$, $\tilde{\Sigma} = (\alpha^*)^2 \Sigma$
2. Sample $\tilde{\theta} | \tilde{W}, \tilde{\Sigma}, X \quad \Rightarrow \quad \tilde{\mu} = G(X; \tilde{\theta}), \mu^* = \tilde{\mu}/\alpha^*$
3. Sample $\tilde{\Sigma}, \alpha | \tilde{W} - \tilde{\mu} \quad \Rightarrow \quad \mu = \tilde{\mu}/\alpha$, $\Sigma = \tilde{\Sigma}/\alpha^2$, and $W = \mu^* + \frac{\tilde{W} - \tilde{\mu}}{\alpha}$.

MPBART

Algorithm 2 (Accelerated MPBART):

Change Step 2 of Algorithm 1

$$\tilde{\theta}|\tilde{W}, \tilde{\Sigma}, X \quad \Rightarrow \quad \tilde{\mu} = G(X; \tilde{\theta}), \mu^* = \tilde{\mu}/\alpha^*$$

into

$$\theta|W, \Sigma, X \quad \Rightarrow \quad \mu^* = G(X; \theta), \tilde{\mu} = \alpha^* \mu^*$$

R package available at `https://github.com/yizhenxu/GcompBART`

MPBART

Intuition: Algorithm 1 fits θ to **unconstrained** latent utilities $\tilde{W}$
– this may cause trouble to model convergence

1. $\tilde{W}$ is unstable
2. sum-of-trees parameters θ are fitted by stochastic search $\Rightarrow \tilde{\theta} \neq \alpha^* \theta$

Constrained latent utilities W are more stable $\Rightarrow$ Algorithm 2

Simulation

$$(X_1, \ldots, X_5) \sim \text{Uniform}(0, 1)$$
$$X_6 \sim \text{Uniform}(0, 2)$$
$$G_1 = 15\sin(\pi X_1 X_2) + (X_3 - 0.5)^2 - 10X_4 - 5X_5$$
$$G_2 = (X_3 - 0.5)^2 - X_4 X_5 + 4X_6$$
$$G^T = (G_1, G_2), \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$
$$\tilde{W} = (\tilde{W}_1, \tilde{W}_2)^T \sim \text{MVN}(G, \Sigma)$$
$$S = \begin{cases} 1 & \text{if } \tilde{W}_1 > \tilde{W}_2, \tilde{W}_1 \geq 0 \\ 2 & \text{if } \tilde{W}_2 \geq \max\{0, \tilde{W}_1\} \\ 3 & \text{if } \tilde{W}_1 < 0 \text{ and } \tilde{W}_2 < 0 \end{cases}$$

The proportion of $S = 3$ is less than 4%, presenting an extremely imbalanced outcome distribution.

Accuracy Measures

- J posterior samples, N subjects
- Posterior mean accuracy: the average accuracy across all posterior predictions,

$$\frac{1}{NJ} 1\{\hat{S}_i^{(j)} = S_i\}, \tag{1}$$

Simulation

Algorithm	Train	Test
1	0.632	0.595
2	0.896	0.877

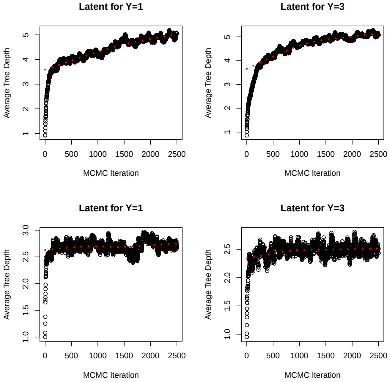
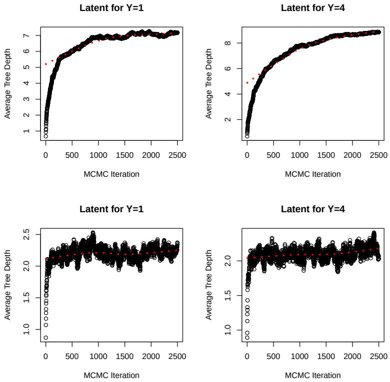


Figure: Plot of average tree depth for each latent utility as time series.

Application - AMPATH Data

Engagement in care problem at $t = 1$

Algorithm	Train	Test
1	0.616	0.608
2	0.786	0.781



Method

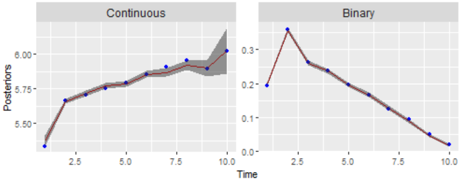
EHRs

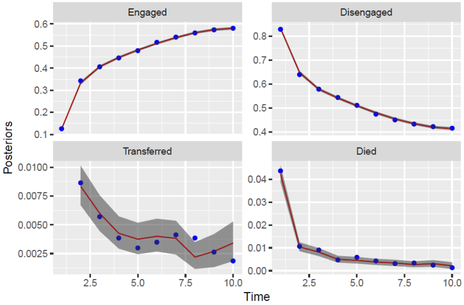
Step 1: Model estimations on 50,000 subjects

Step 2: Model validation on 10,000 subjects

Step 3: Bayesian GCA simulation on 30,000 subjects

Validation of Predictive Models





Counterfactual Simulation

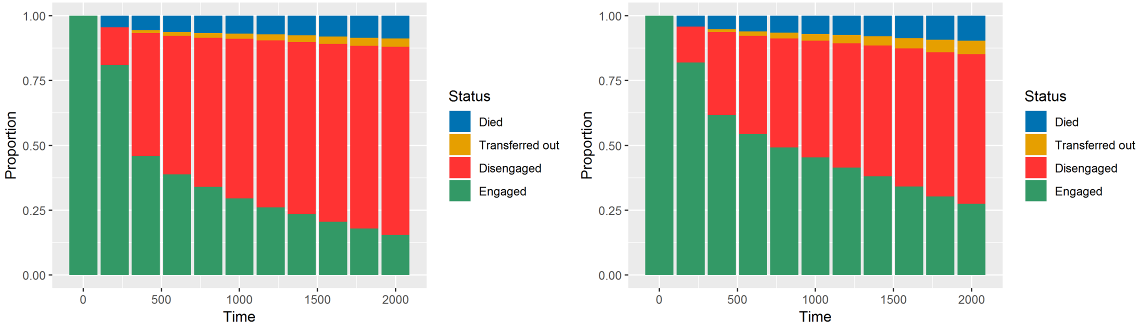
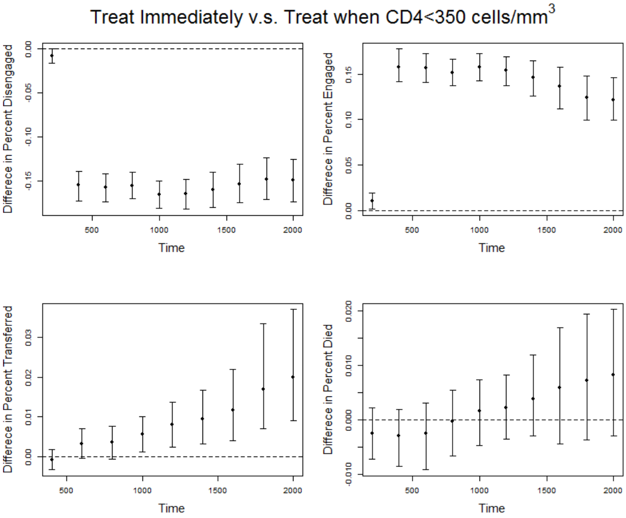


Figure: Predicted marginal state probabilities for an out-of-sample 30,000 individuals engaged in AMPATH-supported HIV care at baseline, under treat when CD4 drops below 350 cells/mm^3 and treat immediately policies (in the order of display, left to right).

Comparison of Causal Effectiveness



End

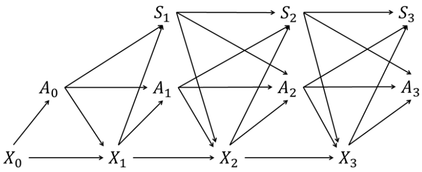
Thank you

Collaborators:

- Liu, Tao - Brown University
- Daniels, Michael - University of Florida
- Marshall, Brandon - Brown University
- Kantor, Rami - Brown University
- Omodi, Victor - Moi University / AMPATH
- Mwangi, Ann - Moi University

This work is supported by NIH grant R01 AI 108441

Model Structure for the Motivating Application



$$[X_1|A_0, X_0, \gamma_1]$$
$$[S_1|A_0, X_1, \theta_1]$$
$$[X_2|A_1, X_1, S_1, \gamma_2]$$
$$[S_2|A_1, X_2, S_1, \theta_2]$$
$$\vdots$$
$$[X_t|A_{t-1}, X_{t-1}, S_{t-1}, \gamma_t]$$
$$[S_t|A_{t-1}, X_t, S_{t-1}, \theta_t]$$

Assumptions:

- No unmeasured confounders
- First-order Markov dependence for S and X

Baseline covariates V is left out for simplicity.

Marginal Augmentation

Imai and van Dyk (2005)

- Data augmentation (DA) algorithm: sample $p(\theta, W|S)$ by iterative posterior sampling of $p(\theta|W, S)$ and $p(W|\theta, S)$
- Marginal augmentation: $L(\theta|S) \propto \int [\int p(S, W|\theta, \alpha)p(\alpha|\theta)d\alpha]dW$; Meng and van Dyk (1999) theoretically proved that this can improve the geometric rate of convergence of the DA algorithm
- "using unidentifiable parameters within a Markov chain is the key to the substantial computational gains offered by marginal augmentation."
- The constraint on Σ is made to be sure the model parameters (θ, Σ) are identified; parameter α is unidentifiable. Even with the constraint, model parameters may be unidentifiable without certain conditions on X and S.

Connection of our Proposal to Imai and van Dyk (2005)

- Imai and van Dyk (2005) provided two algorithms (1' and 2') for implementing MNP, and they expected algorithm 1' to outperform algorithm 2', because algorithm 1' is a complete marginal augmentation procedure while 2' is not.
- In Step 2, algorithm 1' updates α first and then samples θ conditional on the updated α, while algorithm 2' samples θ without conditioning on α
- Kindo et al (2016) employed the algorithm 1' for extending MNP to incorporate BART, skipping the sampling of α in Step 2 and updating θ conditional the α from Step 1; they called this sampling procedure a "semi marginal augmentation"
- Our proposal is somehow similar to the algorithm 2' of Imai and van Dyk (2005), sampling θ from its conditional distribution that does not depend on α, i.e. updating θ conditional on the constrained latent utilities W

Connection of our Proposal to Imai and van Dyk (2005)

Algorithm 1'

- Step 1: $(W, \alpha) \downarrow \tilde{W}$
- Step 2: $(\tilde{\theta}, \alpha) \downarrow \theta$
- Step 3: $\tilde{\Sigma} \downarrow \alpha \downarrow (W, \Sigma)$

Algorithm 2'

- Step 1: $(W, \alpha) \downarrow \tilde{W}$
- Step 2: $\tilde{\Sigma} \downarrow \alpha \downarrow (W, \Sigma)$
- Step 3: θ

Algorithm 1 (Kindo et al)

- Step 1: $(W, \alpha) \downarrow \tilde{W}$
- Step 2: $\tilde{\theta}$
- Step 3: $\tilde{\Sigma} \downarrow \alpha \downarrow (W, \Sigma)$

Algorithm 2 (Proposal)

- Step 1: $(W, \alpha) \downarrow \tilde{W}$
- Step 2: θ
- Step 3: $\tilde{\Sigma} \downarrow \alpha \downarrow (W, \Sigma)$

MNP

Gibbs sampling of (W, θ, Σ)
Linear model specification: $G(X; \theta) = X\theta$

Algorithm 0

1. $(W, \alpha^2)|S, G(X; \theta), \Sigma$, set $\tilde{W} = \alpha W$
2. $(\tilde{\theta}, \alpha^2)|\tilde{W}, \Sigma, \alpha^2, X$, set $\theta = \tilde{\theta}/\alpha$
3. $(\tilde{\Sigma}, \alpha^2)|\tilde{W} - G(X; \tilde{\theta})$, set $W = \tilde{W}/\alpha$ and $\Sigma = \tilde{\Sigma}/\alpha^2$.

$$G(X; \tilde{\theta}) = \alpha G(X; \theta)$$

Bayesian GCA Simulation

Specify predictive models at time $t \in \{1, \ldots, K\}$ using BART,

$$P(X_t | \mathcal{F}_{t-1}, \gamma) \quad (2)$$

$$P(S_t | X_t, \mathcal{F}_{t-1}, \theta) \quad (3)$$

$\mathcal{F}_{t-1}$: observed history up to time $t - 1$.

1. Posterior sampling of parameters (γ^*, θ^*) from (2) and (3)
2. Use the fitted models as generative components.
   Sequentially generate counterfactual paths under certain treatment regime $h(\cdot)$:

$$a^*_{t-1} = h(\mathcal{F}^*_{t-1}) \quad (4)$$

$$x^*_t \sim P(X_t | \mathcal{F}^*_{t-1}, \gamma^*_t) \quad (5)$$

$$s^*_t \sim P(S_t | X_t = x^*_t, \mathcal{F}^*_{t-1}, \theta^*_t), \quad (6)$$

$\mathcal{F}^*_{t-1}$: counterfactual history up to time $t - 1$; $\mathcal{F}^*_0$ represents baseline covariates.

Inclusion Proportions of Covariates

Outcome at $t = 1$