Generalized Bayesian Additive Regression Trees for Dynamic Inference on the Restricted Residual Life

Nicholas C. Henderson

Department of Biostatistics
University of Michigan

WNAR 2025

Introduction

- Treatment decisions are better informed by incorporating **all patient information**, including:
  - risk at baseline
  - time elapsed since diagnosis
  - longitudinal biomarker trajectories
  - treatment history and treatment response
- Recent growth in electronic medical records has led to growing interest in **"dynamic" prediction methods**.
- Dynamic models can identify time-evolving characteristics which explain resistance to certain treatment strategies.

Mean Residual Life (MRL)

- T_i - **time-to-failure** of an event of interest for individual i
- **Aim:** Characterize remaining life expectancy at each of a collection of pre-specified **"landmark times"**.
- We have J landmark times:

$$0 = t_1 < t_2 < \ldots < t_J$$

- The **mean residual life** (MRL) at landmark time t_j is

$$E\Big\{T_i - t_j \mid T_i > t_j, \text{ history of individual } i \text{ up to } t_j\Big\}$$

Data Structure

- T_i - time-to-failure of an event of interest for individual i
- C_i - time of right-censoring for individual i.
- We **observe**:
  - Follow-up times
  $$U_i = \min\{T_i, C_i\}.$$
  - Event indicator:
  $$\delta_i = I(T_i \leq C_i).$$

Summarized Longitudinal Covariates

- For individual i, observe longitudinal information $X_i(t)$ at times $t_{i1}, t_{i2}, t_{i3}, \ldots$.
- Assume relevant longitudinal information up to time t_j can be **summarized** by a $p_j \times 1$ vector $\boldsymbol{x}_{i,j}$.
- We observe

$$\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots$$

  for landmark times while individual i is still at risk.
- The dimensions of $\boldsymbol{x}_{i,j}$ and $\boldsymbol{x}_{i,k}$ **do not** need to be the same.

Drawbacks of using Summarized Longitudinal Covariates

- Cannot **directly handle** many observed data patterns.
  - Requires many user choices about how to summarize complex longitudinal patterns (e.g., wearables, irregularly measured biomarkers, etc.).

- Can be difficult to make the covariate dimension **match** across individuals (i.e., $\boldsymbol{x}_{i,j}$ and $\boldsymbol{x}_{h,j}$ both have length p_j).

- **Currently developing**: extensions of our method to allow for the use of automatically constructed context vectors.

- Bayesian methods can directly **incorporate uncertainty** in such context vectors.

Conditional RMRL functions

- Due to limited observations in the tail of the survival distributions, we focus on the **restricted mean residual life** (RMRL).
- At landmark time t_j, the RMRL function is defined as

$$\mu_{j,\tau_j}(\boldsymbol{x}, \mathcal{H}_{j-1}) = \\ E\Big\{ \min\{T_i, \tau_j\} - t_j \mid T_i > t_j, \boldsymbol{x}_{i,j} = \boldsymbol{x}, \mathcal{H}_{i,j-1} = \mathcal{H}_{j-1} \Big\}.$$

- **Goal:** Perform joint inference on the entire collection of RMRL functions:

$$\mu_{1,\tau_1}(\boldsymbol{x}, \mathcal{H}_{j-1}), \mu_{2,\tau_2}(\boldsymbol{x}, \mathcal{H}_{j-1}), \ldots, \mu_{J,\tau_J}(\boldsymbol{x}, \mathcal{H}_{j-1})$$

Conditional RMRL functions

- We propose a model that directly generates **joint inference** on the collection of RMRL functions.
- Our approach allows for **sharing of information** across landmark times
  - "smoothing" across time
- Number of landmark times is at least moderate (i.e., $J \geq 10$)
- The truncation points τ_j **do not** need to be the same across landmark times j:
  - Should carefully consider comparability and interpretability of these truncation points across time.

“Generalized” Bayesian inference for the RMRL

- Many standard approaches for residual life estimation in regression essentially involve solving a **weighted least-squares** problem.
  - No **“full probability” modeling** of joint distribution of survival and censoring distributions.
- Avoiding full probability modeling can be useful if the main goal is RMRL estimation and inference.
- When handling **informative censoring**: often easier to **separate modeling** of censoring and survival mechanisms.
- How to do **Bayesian** inference with **weighting-based** approaches?
- How can you incorporate flexible Bayesian **machine learning/nonparametric** methods?

RMRL Regression with Inverse Probability of Censoring Weights (IPCW)

- Consider estimating the RMRL function μ_{j,τ_j} at time zero, $t_1 = 0$.
- An IPCW-based approach for estimating a **regression model** for the RMRL is:

$$\hat{\mu}_{1,\tau_1}(\mathrm{x}_i) = g(\mathrm{x}_i^T \hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ minimizes the loss

$$L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{\delta_i}{\hat{G}(\min\{U_i, \tau_1\} | \boldsymbol{x}_i)}}_{IPCW} \Big(\min\{U_i, \tau_1\} - g(\mathrm{x}_i^T \boldsymbol{\beta}) \Big)^2$$

- Note that loss $L(\boldsymbol{\beta})$ **only depends** on the observations with events (i.e., those with $\delta_i = 1$).

IPCW assumptions

- The IPCW procedure works properly only under certain assumptions about the dependence of T_i and C_i.
- **Noninformative censoring:** T_i and C_i are independent.
- **Informative censoring:** T_i and C_i are independent conditional on covariate history.
  - Use estimates of covariate history-dependent censoring probabilities as inverse weights.

An IPCW-based Loss function for landmark time t_j

- Assume the **censoring distribution** G_j conditional on
  - surviving up to landmark time t_j
  - covariate history up to landmark time t_j

  is known:

$$G_j(t|\boldsymbol{x}_{i,j}, \mathcal{H}_{i,j-1}) = P(C_i > t | T_i > t_j, \boldsymbol{x}_{i,j}, \mathcal{H}_{i,j-1}).$$

- Consider a **candidate estimator** f_j of the RMRL function μ_{j,τ_j}
- An IPCW-based loss function for evaluating f_j is:

$$\mathcal{L}_\eta(f_j|G_j; \mathrm{U}^\tau, \boldsymbol{\delta}) = \eta \sum_{i=1}^{n} \frac{V_{ij}\delta_i}{G_j(U_i^{\tau_j}|\boldsymbol{x}_{i,j}, \mathcal{H}_{i,j-1})} \Big(U_i^{\tau_j} - t_j - f_j(\boldsymbol{x}_{i,j}) \Big)^2.$$

- V_{ij} - **at risk** indicator for landmark time t_j.

An IPCW-based Loss function

- A **joint** IPCW-based loss function is

$$\mathcal{L}_{\eta}(f_1, \ldots, f_J | G_1, \ldots, G_J, \mathrm{data}) = \sum_{j=1}^{J} \eta_j \sum_{i=1}^{n} \frac{V_{ij}\delta_i}{G_j(U_i^{\tau_j} | \boldsymbol{x}_{i,j}, \mathcal{H}_{i,j-1})} \Big\{ U_i^{\tau_j} - t_j - f_j(\boldsymbol{x}_{i,j}) \Big\}^2.$$

- In the above, $\eta_1, \ldots, \eta_J$ can be thought of as **tuning parameters** (or "learning rates") that need to be chosen.

- The η_j control the **relative impact** of the loss function versus the prior distribution on the final inference.

Gibbs Posterior conditional on Censoring Distributions

- Adapt the **"generalized Bayes" framework** (Bissiri et al. (2016)) for loss function-based updating of a prior distribution.
- Assume that we have already specified a prior over the functions $f_1(\cdot), \ldots, f_J(\cdot)$ (which we denote with π_B).
- The **joint "Gibbs posterior"** for the RMRL functions is $f_1, \ldots, f_J$ is

$$p(f_1, \ldots, f_J | G_1, \ldots, G_J, \mathrm{data}) = \frac{\exp\big\{ -\mathcal{L}_\eta(f_1, \ldots, f_J | G_1, \ldots, G_J, \mathrm{data})\big\}\pi_B(f_1, \ldots, f_J)}{\int \exp\big\{ -\mathcal{L}_\eta(f_1, \ldots, f_J | G_1, \ldots, G_J, \mathrm{data})\big\} d\pi_B(f_1, \ldots, f_J)}.$$

Marginalized Gibbs Posterior

- The final posterior **should not** depend on assuming $G_1, \ldots, G_J$ are known.
- We propose using the following **marginalized** Gibbs posterior:

$$p(f_1, \ldots, f_J | \text{data}) = \int p(f_1, \ldots, f_J | G_1, \ldots, G_J, \text{data}) d\tilde{\pi}_{\mathsf{U},\boldsymbol{\delta}}(G_1, \ldots, G_J)$$

- $\tilde{\pi}_{\mathsf{U}^\tau,\boldsymbol{\delta}}(\cdot)$: posterior distributions for $G_1, \ldots, G_J$ based on observed data $(\mathsf{U}^\tau, \boldsymbol{\delta})$.
- **Interpretation:** The marginal posterior $p(f_1, \ldots, f_J | \mathsf{U}, \boldsymbol{\delta})$ is obtained by averaging over the G_j-conditional Gibbs posteriors $p(f_1, \ldots, f_J | G_1, \ldots, G_J, \text{data})$.

Modeling the RMRL Functions

- Main goal is to model the RMRL functions $\mu_{1,\tau_1}, \mu_{2,\tau_2}, \ldots$.
- The RMRL functions are **weighted averages** of a large collection ($J \times H$) of **local, BART models**.
- Landmark j-specific, local BART function with H trees:

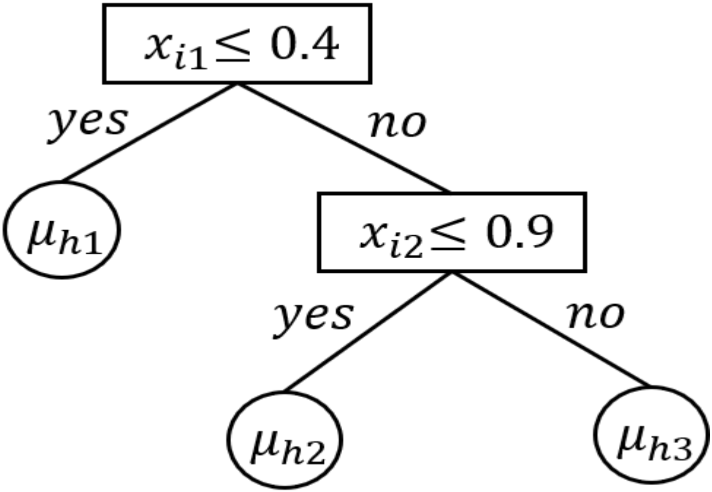
$$\theta_j(\boldsymbol{x}_{i,j}) = \sum_{h=1}^{H} \sum_{k=1}^{n_{jh}} \mu_{jhk} I_{jhk}(\mathrm{x} | \mathcal{T}_{jh})$$

- The RMRL functions are assumed to have the form

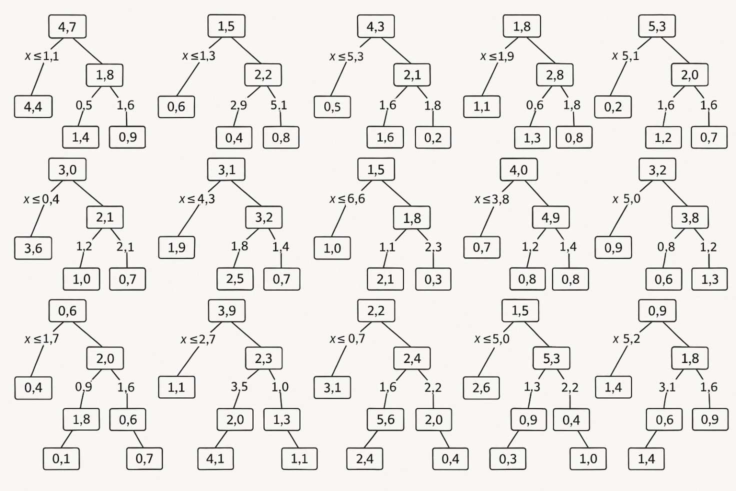
$$\mu_{j,\tau}(\boldsymbol{x}_{i,j}, \mathcal{H}_{i,j-1}) = \sum_{l=1}^{J} A_{jl}(\boldsymbol{\alpha}) \theta_l(\boldsymbol{x}_{i,l}).$$

- $A_{jl}(\boldsymbol{\alpha})$ - **smoothing weights** whose form can depend on a vector of hyperparameters $\boldsymbol{\alpha}$.

Single Decision Tree from BART



RMRL BART - $J \times H$ trees



Independent BARTs model

- Simplest approach.
- Assume that **all** the local BART models are **independent**.
- RMRL functions:

$$\mu_{j,\tau_j}(\boldsymbol{x}_{i,j}, \mathcal{H}_{j-1}) = \theta_j(\boldsymbol{x}_{i,j}),$$

- Smoothing weights:

$$A_{jl}(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{if } l = j \\ 0 & \text{otherwise} \end{cases} .$$

Autoregressive Model of Order 1

- An autoregressive **"super landmark"** model that connects the local BART functions is

$$\begin{aligned} \mu_{1,\tau}(\boldsymbol{x}_{i,1}, \mathcal{H}_{j-1}) &= \theta_1(\boldsymbol{x}_{i,1}) \\ \mu_{j,\tau}(\boldsymbol{x}_{i,j}, \mathcal{H}_{j-1}) &= \alpha\mu_{j-1,\tau}(\boldsymbol{x}_{i,j-1}, \mathcal{H}_{j-1}) + \theta_j(\boldsymbol{x}_{i,j}), \quad \text{for } j \geq 2 \end{aligned}$$

- One can show that this is equivalent to having

$$\mu_{j,\tau}(\boldsymbol{x}_{i,j}, \mathcal{H}_{j-1}) = \sum_{l=1}^{J} A_{jl}(\boldsymbol{\alpha})\theta_l(\boldsymbol{x}_{i,l}).$$

with smoothing weights of the form

$$A_{jl}(\boldsymbol{\alpha}) = \begin{cases} \alpha^{j-l} & \text{if } l \leq j \\ 0 & \text{if } l > j. \end{cases}$$

Modeling the Censoring Distributions

- Many options available.
- **Noninformative censoring:** Nonparametric gamma process priors for cumulative hazard of censoring distribution.
- **Informative censoring:** A single, BART-based AFT model for C_i given baseline covariates.
- **Informative censoring with time-varying covariates:** Sequential BART-based AFT models for

$$C_i | T_i > t_j, \text{covariate history}.$$

- In simulations, using BART seems to work quite well, **even when** censoring is noninformative.

Posterior Computation

- **Similar** to posterior computation strategy for "classic BART" in a standard regression context.
- Update trees and terminal node parameters **one at a time**, using residuals from fit **of all other** terms.
- In our case, we have **more** trees and terminal node parameters to update ($J \times H$) versus H in standard BART.
- In our approach, the residuals **depend on the hyperparameters $\boldsymbol{\alpha}$**.
- Need an extra step in each MCMC iteration for drawing **IPCW weights**.
- Also, need an extra step for **updating smoothing hyperparameters $\boldsymbol{\alpha}$**.

Friedman Function Simulation Study

- Generate $Z_{ij} \sim \mathrm{Bernoulli}(\theta_i)$ and $W_{ij} \sim \mathrm{Bernoulli}(\lambda)$
- **Landmark times:** $t_j = (j-1)$, for $j = 1, \ldots, 30$.
- The success probability of Z_{ij} is

  $$\theta_i = 1/\{1 + \exp(-f(\boldsymbol{x}_i))\}.$$

  $f(\boldsymbol{x})$ - the **"Friedman" function** and $\boldsymbol{x}$ has length 10.
-

  $$\begin{aligned} T_i &= \min\{t_j : Z_{ij} = 1\} + v_{it} \\ C_i &= \min\{t_j : W_{ij} = 1\} + v_{ic}, \end{aligned}$$

  where $v_{it}, v_{ic} \sim \mathrm{Uniform}(0, 1)$.

RMSE from Friedman Function Simulation Study (Informative Censoring)

# of observations	$n = 5000$			
# of predictors	$p = 10$		$p = 50$	
Method	LC	HC	LC	HC
RMRL-BART-DEP (smooth)	0.63	**0.74**	**0.77**	**0.99**
RMRL-BART-DEP	0.65	0.79	0.81	1.03
RMRL-BART	**0.60**	1.09	0.78	1.54
CoxPH	2.55	2.51	2.57	2.55
Regularized CoxPH	2.57	2.52	2.59	2.54
IPCW tree boost	2.61	3.09	2.62	3.08
AFT-BART	2.00	7.56	2.09	7.78

Table: Friedman function simulation study with informative censoring. Average RMSE across landmark times for RMRL-BART and related approaches.

Empirical Coverage Proportions

# of observations	$n = 1000$			
# of predictors	$p = 10$		$p = 100$	
Method	$r = 0.1$	$r = 0.2$	$r = 0.1$	$r = 0.2$
RRML-BART (Autoreg)	**0.96**	**0.94**	**0.95**	**0.91**
RRML-BART-**fixedG** (smooth)	0.93	0.93	0.91	0.84
RRML-BART (Indep)	0.94	0.93	0.94	0.88
RRML-BART-**fixedG** (Indep)	0.93	0.91	0.92	0.80
RMRL-BART-default (Indep)	0.92	0.89	0.91	0.87
RMRL-BART-default-**fixedG** (Indep)	0.88	0.80	0.86	0.79

Table: Coverage results for independent and autoregressive versions of RRML-BART. Also, comparisons of coverage where a different set of IPCW weights are drawn in each MCMC iteration versus a version of RRML-BART (fixedG) where the same sets of IPCW weights are used in each MCMC iteration.

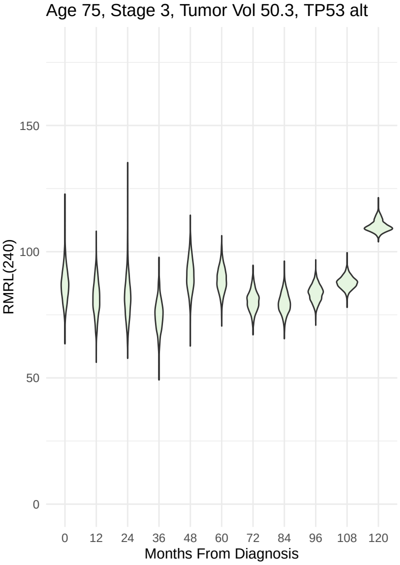
RMRL-BART in action: The METABRIC Study

- 1839 clinically annotated **breast cancer specimens** from tumor banks across multiple institutions in the UK and Canada.
- Data taken from study first reporting on this: Curtis et al. (2012).
- Substantial follow-up time for many patients
  - Median follow up: 115 months (9.6 years)
  - 9% of patients with more than 20 years of follow up
  - Time-to-event endpoint of interest: time from diagnosis until death (OS)

RMRL-BART in action: The METABRIC Study

- We utilized 30 of the available **clinicogenomic covariates** made available.
  - Clinical covariates: (e.g., age, treatment received, menopausal status, ...)
  - Tumor related measures: (e.g., tumor size, node positive status, ...)
  - Gene expression measures: (top 10 genes with the most variation)

OS in the METABRIC Study: Independent BARTs model



Conclusion

- Results from RMRL-BART seem promising so far.
- Borrowing information through an autoregressive model does seem to improve prediction.
  - However, not necessarily better than applying to loess to landmark-specific posterior means.
  - Better coverage than post-processing independent BART fits with loess.
- Interested in expanding RMRL-BART to handle constructed covariates with associated uncertainty measures.

Thanks