Improve Survey Inference Using Bayesian Machine Learning

Qixuan Chen, PhD

Department of Biostatistics
Columbia University

August 8, 2024
2024 Joint Statistical Meetings

Background

- Inference about a target population based on sample data relies on the assumption
  - the sample is representative
  - the sample can be adjusted to account for nonrepresentativeness
- Random samples are often not available in real data problems
  - probability surveys with low response rates are often nonrepresentative
- There is a need to generalize inference from an available nonrandom sample to the target population

Outline

- **A data-rich setting**: unit-level high-dimensional auxiliary variables are available in the population
- **A common setting**: only population margins of the auxiliary variables are available

A data-rich setting: unit-level high-dimensional auxiliary variables are available in the population

Notation and data setting

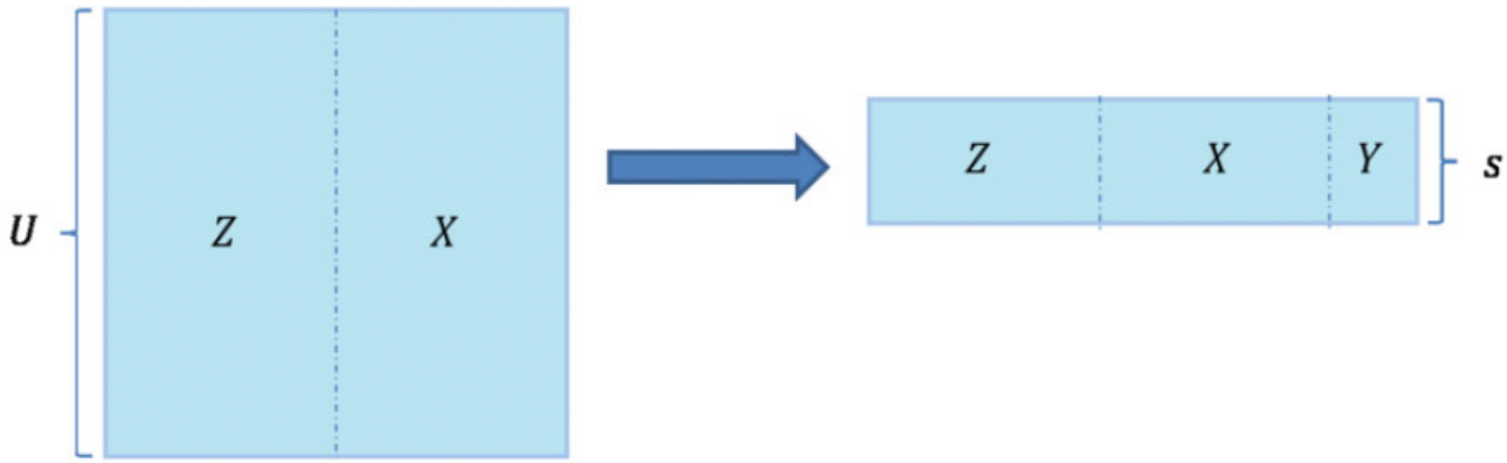


Figure 1. Population U and Nonrandom Sample s with Shared Discrete Auxiliary Variables Z and Continuous Auxiliary Variables X as well as Outcome y Measured Only in s.

We consider a continuous y with the estimand of interest

$$Q(y) = \frac{1}{N} \sum_{i \in U} y_i \tag{1}$$

Challenges with existing methods

- Commonly used methods to adjust for selection bias in nonrandom samples
  - poststratification
  - raking
  - multilevel regression and poststratification (MRP)
- Limitations of these methods
  - all require first discretizing the continuous auxiliary variables $\mathbf{X}$
  - it can be challenging to implement in high-dimensional settings

New approach: regularized prediction

- The population mean can be estimated from a model-based perspective

$$\widehat{Q}(y) = \frac{1}{N}\left(\sum_{i\in s} y_i + \sum_{j\notin s}\hat{y}_j\right) = \frac{1}{N}\left(\sum_{j\in U}\hat{y}_j + \left(\sum_{i\in s} y_i - \sum_{i\in s}\hat{y}_i\right)\right) \tag{2}$$

- To estimate $\hat{y}$, we consider BART (Chipman, George, and McCulloch 2010) or soft BART (Linero and Yang 2018) models. With continuous y,

$$y = G(\mathbf{z},\mathbf{x}) + \epsilon = \sum_{m=1}^{M} g(\mathbf{z},\mathbf{x};T_m,\mu_m) + \epsilon, \quad \epsilon \sim N(0,\sigma^2) \tag{3}$$

M is number of trees, T_m is the mth binary tree with μ_m being parameters associated with the terminal nodes, $g(\cdot)$ assigns μ_m according to $\mathbf{z}$ and $\mathbf{x}$.

The BART and SBART estimators

Step 1 Model $p(y|\boldsymbol{z}, \boldsymbol{x})$ using BART or soft BART: $y = G(\boldsymbol{z}, \boldsymbol{x}) + \epsilon$, $\epsilon \sim N(0, \sigma^2)$.

Step 2 Obtain posterior distributions of $Q(y)$. At iteration t,

- draw $G^{(t)}, \sigma^{(t)} | y_{i\in s}, \boldsymbol{z}_{i\in U}, \boldsymbol{x}_{i\in U}$
- compute $\tilde{\theta}_i^{(t)} = G^{(t)}(\boldsymbol{z}_i, \boldsymbol{x}_i)$ for $i \in U$
- obtain $\widehat{Q}_{\text{(S)BART}}^{(t)} = \frac{1}{N}\left[\sum_{i\in U} \tilde{\theta}_i^{(t)} + \left(\sum_{i\in s} y_i - \sum_{i\in s} \tilde{\theta}_i^{(t)}\right)\right]$

Step 3 Obtain $\widehat{Q}_{\text{(S)BART}}$: point estimates using (posterior) median of $\widehat{Q}_{\text{(S)BART}}^{(t)}$ with credible intervals constructed using quantiles splitting the tails of posterior distribution equally.

BART and SBART propensity prediction

- A doubly robust approach
  - Little and An (2004) proposed including the logit-transformed response propensity as a covariate using spline in the imputation models to yield robust estimates of sample means when the imputation model is misspecified.
- We extend the BART and SBART prediction by adding the sample inclusion propensity as a covariate in the prediction models
  - estimate $\pi = \Pr(I = 1|\mathbf{z}, \mathbf{x})$ via using probit BART
  - model y using $y = G(\mathbf{z}, \mathbf{x}, \hat{\pi}) + \epsilon$

Simulation studies

- Population data
  - $N = 3000$, $n = 600$
  - binary variable $\{z_l\}_{l=1,\ldots,p}$ were obtained such that $\Pr(z_l = 1)$ falls in the range of $(0.34, 0.66)$
  - continuous variable $\{x_l\}_{l=1,\ldots,r}$ were generated from Unif(0,1)
- Four simulation scenarios
  - S1: low-dimensional covariates ($p = 3, r = 1$)
  - S2: high-dimensional covariates ($p = 30, r = 10$)
  - S3: high-dimensional covariates ($p = 30, r = 10$) with data sparsity on the auxiliary variables associated with y
  - S4: high-dimensional covariates ($p = 30, r = 10$) with data sparsity on the auxiliary variables NOT associated with y

Simulated population and sample data

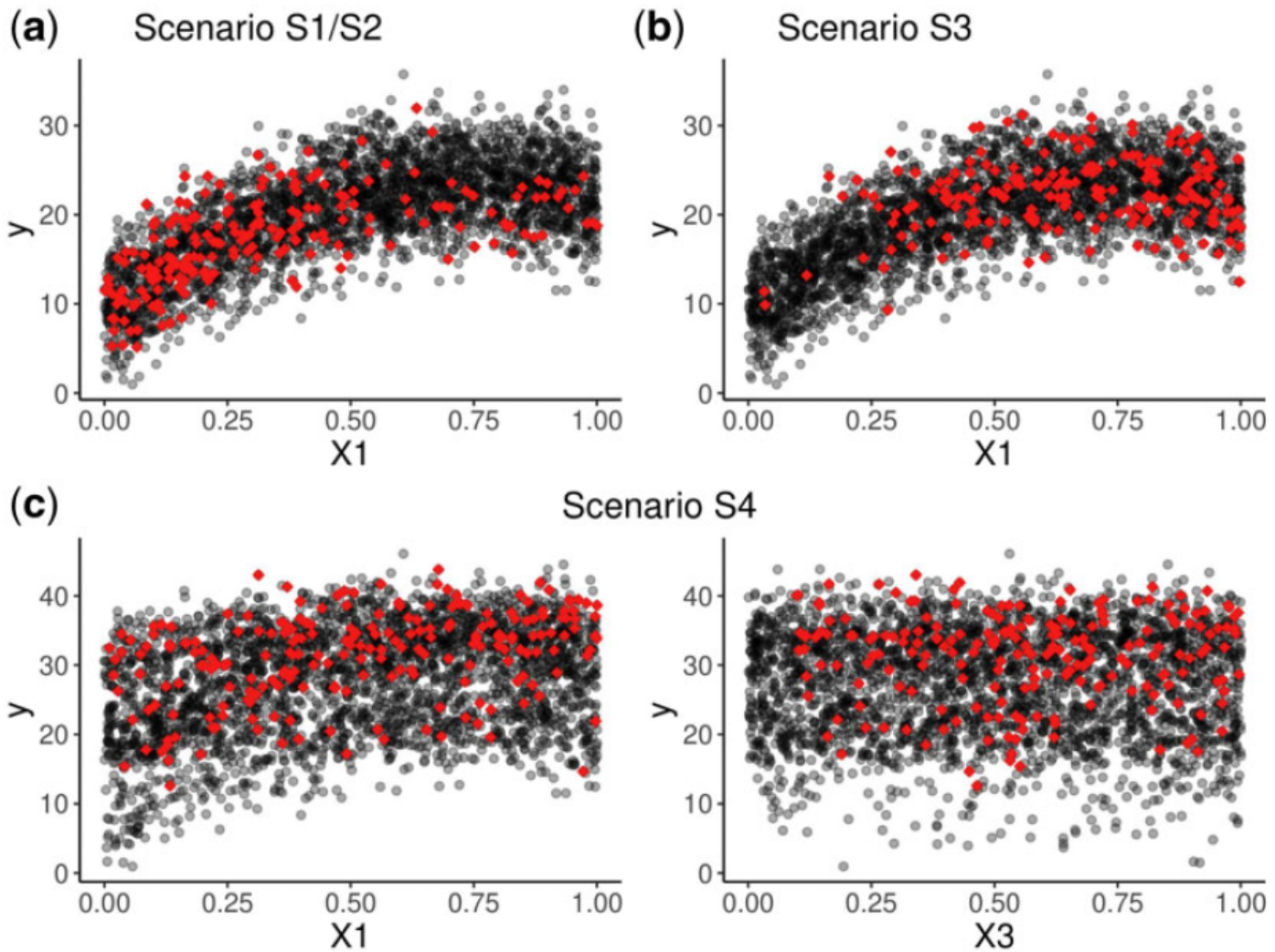


Figure 2. Scatterplots of Outcomes *y* Versus Continuous Auxiliary Variables of Units in the Population (in Gray Dots) and a Selected Sample (in Red Diamonds) for (a) Scenario S1–S2, (b) Scenario S3, and (c) Scenario S4.

Simulation results: bias and RMSE

Table 1. Simulation Results: Empirical Bias and RMSE of Various Methods in Estimating Population Means, from 500 Simulation Replicates, for Each Simulation Setting

Method	S1		S2		S3		S4	
	$Q = 19.88$						$Q = 27.74$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Raw	−2.99	2.99	−2.99	2.99	2.43	2.43	3.13	3.14
PS[a]	−0.37	0.43						
Raking[b]	−0.16	0.22						
BART	−0.08	0.17	−0.17	0.22	0.37	0.43	0.07	0.17
BART-P	−0.06	0.18	−0.12	0.20	0.30	0.38	0.06	0.17
SBART	−0.08	0.17	−0.10	0.19	0.24	0.32	0.04	0.16
SBART-P	−0.07	0.18	−0.10	0.19	0.24	0.32	0.04	0.16

NOTE.—The standard errors of empirical bias from 500 simulation replicates are $< 7.5 \times 10^{-3}$ for all methods.

[a]PS is based on Z_1, Z_2, Z_3, and X_1 discretized using tertiles.

[b]Ranking is based on Z_1, Z_2, Z_3, and X_1 discretized using quintiles.

Simulation results: coverage rate

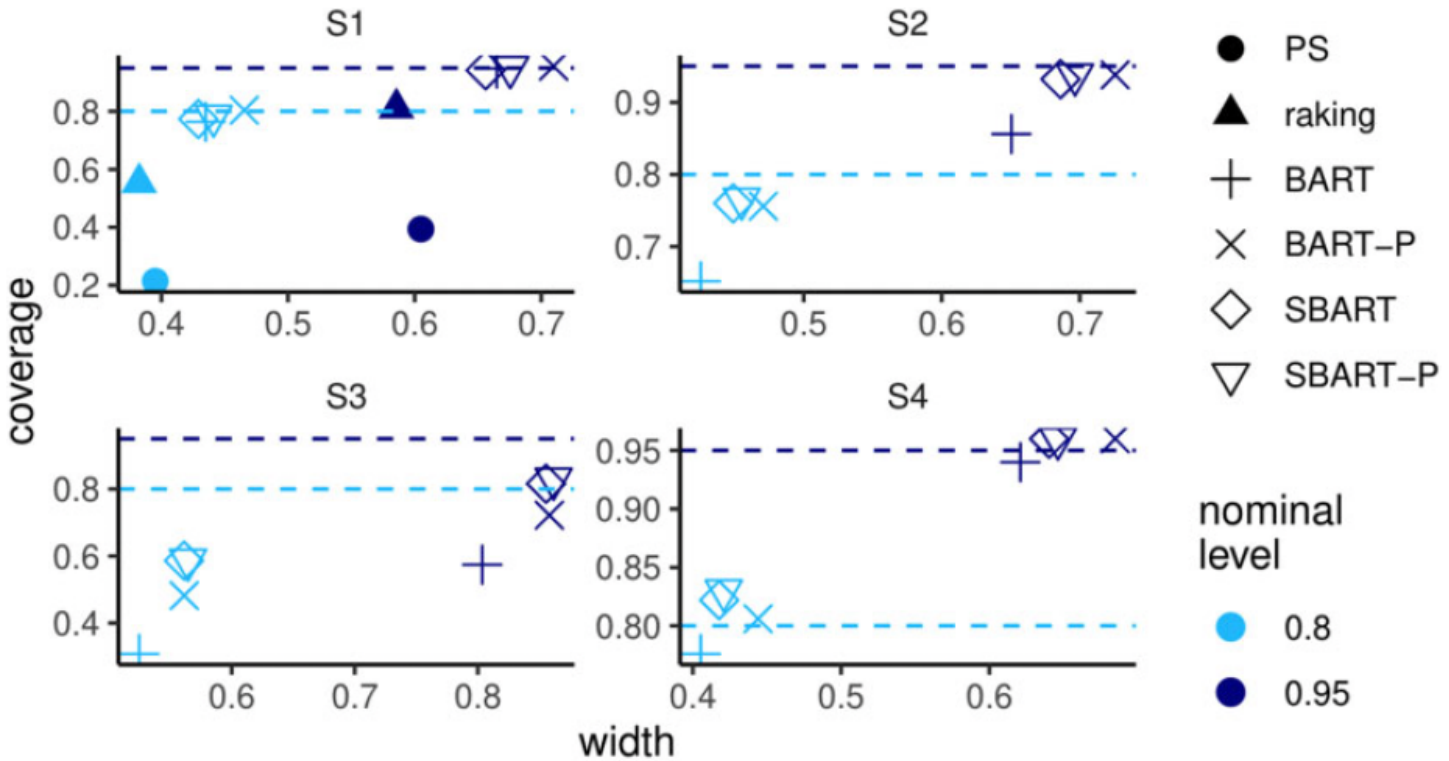
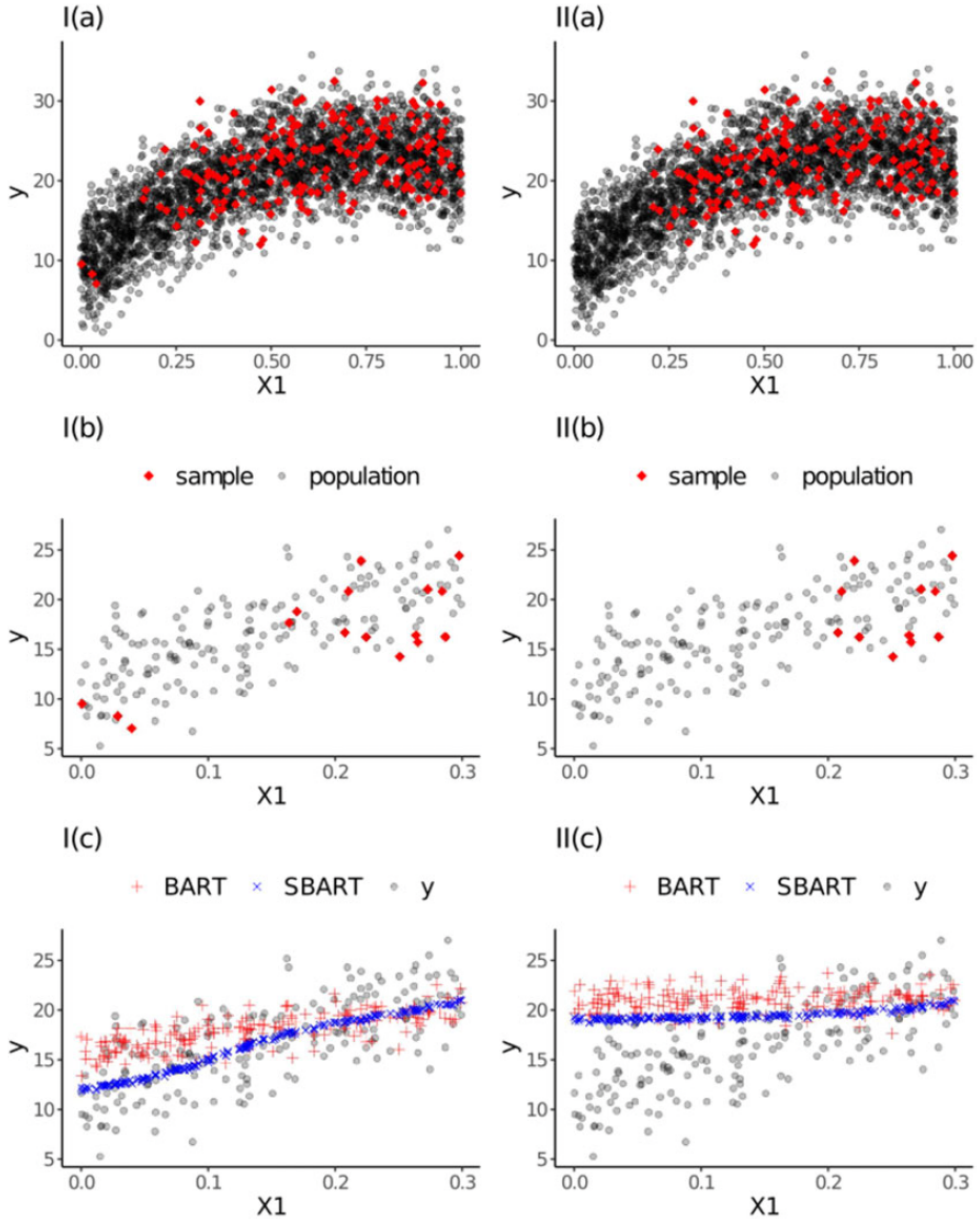


Figure 3. Simulation Results—Empirical Coverage Rates of 80% and 95% Probability Intervals (with the Horizontal Dashed Lines Denoting the Nominal Levels) Against Average Probability Interval Widths, from 500 Simulation Replicates, for Each Simulation Setting.

Comparison of BART and soft BART

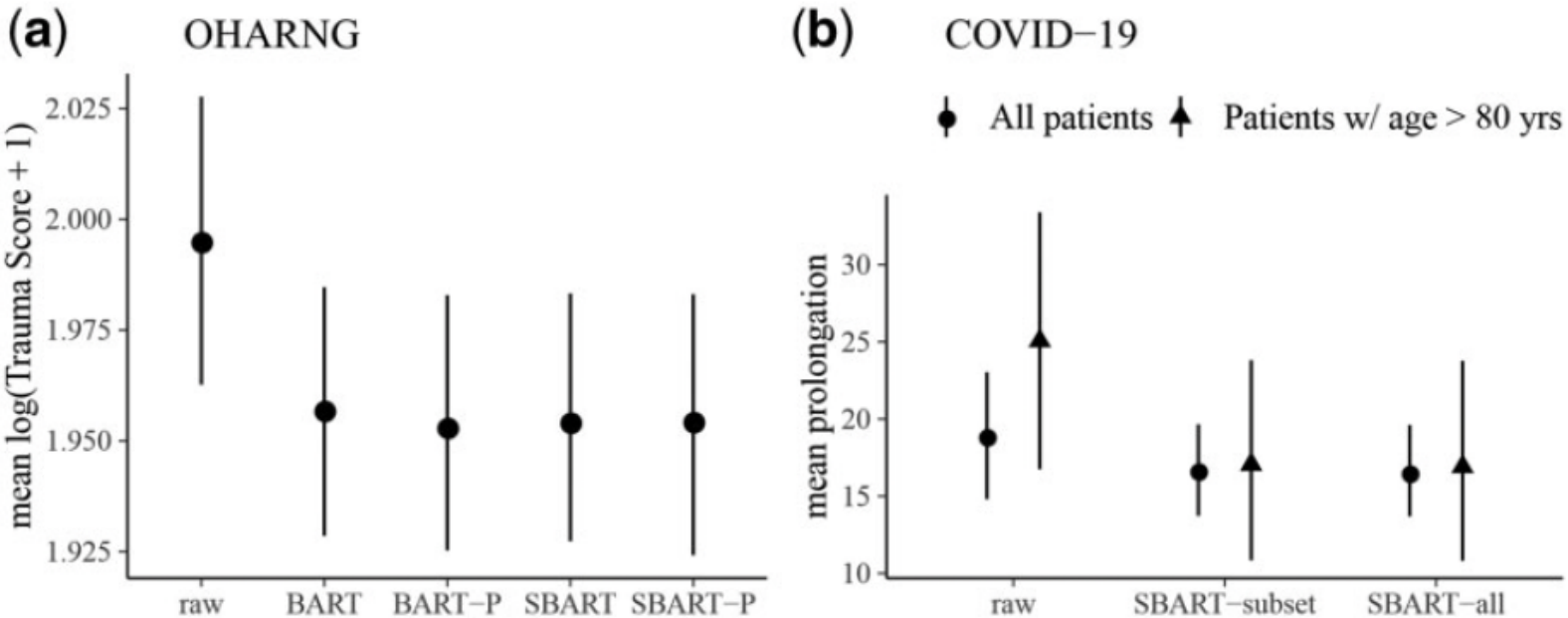


Sensitivity analyses under Scenario 2

- Robustness of the (S)BART-P by excluding important continuous auxiliary variable from (S)BART prediction models
  - (S)BART performs poorly, but (S)BART-P performs well
- Using estimated versus true π
  - results are similar
- Robust to the positivity assumption, allowing some population units to have zero chance of being selected
  - the positivity assumption is not necessary for the proposed methods as long as the three required assumptions are satisfied.

Real data applications

- We demonstrate the application using two real data
  - a mental health survey of Ohio Army National Guards
  - a real world data of New York City COVID-19 study



Conclusions

- Our regularized prediction methods using (soft) BART
  - effectively reduce selection bias in the nonrandom sample
  - yield efficient estimates of population quantities
  - with close to the nominal level coverage rate
- In highdimensional setting with sparse signals, SBART, with soft decision trees and sparsity-inducing priors, is less biased and more efficient than BART.
- When BART underperforms, including propensity score in BART could reduce bias and improve credible interval coverage while such benefit is not obvious for SBART.
- SBART-P can offer protection from model misspecification in SBART, when important predictor is omitted from the model.

Use with caution

- The Bayesian additive-trees-based methods need to be used with caution
  - Both BART and SBART prediction fail when selection bias results in no data points at the tails of the continuous covariates that are associated with the outcomes.
  - For important continuous variables associated with the outcomes, the range and distribution in the sample and population need to be checked before using the methods.

A common setting: only population margins of the auxiliary variables are available

Background

- Unit-level population data on auxiliary variables are not always available
- Aggregated data of population characteristics are more accessible
  - joint distribution of population characteristics
  - population margins
- Multilevel regression and poststraitfication (MRP) can provide efficient estimate of population quantities, while reducing selection bias
  - requires the complete population joint distribution of post-stratifiers
- We proposed an adapted MRP method
  - only requires the population margins of post-stratifiers

Multilevel Regression and Poststratification

- The MRP method consists of two steps (Gelman 1997).
  - First, fit a multilevel regression model on y.

$$g\left(\mathrm{E}(y_i|\mathbf{x})\right) = \alpha_0 + \sum_{q=1}^{Q} \alpha_q x_{iq} + \sum_{k=1}^{K} \alpha_{\ell[i]}^{(k)}, \tag{4}$$

$$\alpha_{\ell[i]}^{(k)} \mid \sigma^{(k)} \sim \text{normal}\left(0, \sigma^{(k)}\right),$$

  - Second, obtain the estimated mean of y in the j-th population subgroup, denoted using $\hat{\theta}_j$.

$$\hat{\theta}^{\text{MRP}} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j}. \tag{5}$$

where N_j is the population count for subgroup j and is assumed to be known.

Adaptation of MRP

- When N_j is unknown, it needs to be estimated.
- We proposed a two-step approach to estimate N_j
  - First, model the sample count c_j for cells in the poststratification table formed by the auxiliary variables using a Bayesian model
    - Poisson
    - Negative binomial
    - BART
  - Then, apply raking to the estimated c_j drew from their posterior predictive distributions based on the fitted models using the known population margins of post-stratifiers

The adapted MRP estimator

- To obtain an adapted MRP estimate
  - obtain $\hat{N}_{j,d}$, for $d = 1, \ldots, 1000$ draws
  - obtain $d = 1, \ldots, 1000$ draws of $\hat{\theta}_{j,d}$ from model (4)
  - obtain draws of

$$\hat{\theta}_d^{\text{MRP-adapted}} = \frac{\sum_{j=1}^{J} \hat{N}_{j,d} \hat{\theta}_{j,d}}{\sum_{j=1}^{J} \hat{N}_{j,d}} \tag{6}$$

  - These draws are averaged to obtain the point estimate of θ.

Simulation & application

- The primary take-home message
  - The MRP adaptations perform similarly to MRP, although MRP exhibits slightly smaller bias and RMSE by using true N_j
  - BART is an attractive alternative to parametric models for estimating N_j when there are high dimensional auxiliary variables.
- We apply the MRP adaptations to a survey in NYC among persons with HIV, in which in-person data collection was disrupted by COVID-19 pandemic.

- **Reference**
  - Liu, Y. Gelman, A., Chen, Q. (2023). “Inference from nonrandom samples using Bayesian machine learning”, *J*ournal of Survey Statistics and Methodology, 11, 433-435.
  - Pitts, A.J., Yomogida, M., Aidala, A., Gelman, A., Chen, Q. “Multilevel regression and poststratification using margins of post-stratifiers: improving inference for HIV outcomes during the COVID-19 pandemic”, *to be submitted*.
- **Acknowledgement**
  - This work was supported by the US National Institutes of Health (R01AG067149, R21ES029668).