Introduction to binary and categorical outcomes with BART

Rodney Sparapani
Medical College of Wisconsin

June 7 & 8: BART workshop
Medical College of Wisconsin, Milwaukee campus

Funding for this research was provided, in part, by the Advancing Healthier Wisconsin Research and Education Program under awards 9520277 and 9520364.

Outline

Sparapani, Spanbauer & McCulloch 2021
Journal of Statistical Software

- Motivation: chronic spine pain and obesity
- Dichotomous outcomes with probit BART
- Dichotomous outcomes with logistic BART
- Geweke convergence diagnostics for binary BART
- Categorical outcomes with logistic BART
- Categorical outcomes with probit BART

Motivation: chronic spine pain and obesity

- Hypothesis a: obesity is a risk factor for chronic lower back/buttock pain
- Hypothesis b: obesity is NOT a risk factor for chronic neck pain
- Data available from the National Health and Nutrition Examination Survey (NHANES) 2009-2010 Arthritis Questionnaire
- 5106 subjects were surveyed
- Demographics: age and gender
- Anthropometrics available: weight (kg), height (cm), body mass index (kg/m^2), waist circumference (cm)
- Sampling weights to estimate for the US as a whole
- For obesity quantified by BMI, see `demo/nhanes.pbart1.R` and `demo/nhanes.pbart2.R` in the **BART** R package
- For obesity quantified by waist circumference, see `demo/nhanes.pbart.R` in the **BART3** R package

Probit BART for binary outcomes

Probit regression with latent variables: Albert & Chib 1993 *JASA*

$$y_i | p_i \overset{\text{ind}}{\sim} \text{B}(p_i)$$

$$p_i | f = \Phi(\mu + f(x_i)) \text{ where } f \overset{\text{prior}}{\sim} \text{BART} \text{ and } \mu = \Phi^{-1}(\bar{y})$$

$$z_i | y_i, f \sim \text{N}(\mu + f(x_i),\ 1) \begin{cases} \text{I}(-\infty, 0) & \text{if } y_i = 0 \\ \text{I}(0, \infty) & \text{if } y_i = 1 \end{cases}$$

$$f | z_i, y_i \overset{d}{=} f | z_i$$

$$[y|f] = \prod_{i=1}^{N} p_i^{y_i} (1 - p_i)^{1-y_i} \qquad \text{Likelihood}$$

Continuous BART with unit variance, $\sigma^2 = 1$, and z_i are the data

Friedman's partial dependence function for probit BART

Friedman 2001 *AnnStat*

$$p(x) = p(x_S, x_C) \quad \text{BART function where } x = [x_S, x_C]$$

$$\begin{aligned}
p(x_S) &= \mathrm{E}_{x_C}\left[p(x_S, x_C) | x_S\right] \\
&\approx N^{-1} \sum_i p(x_S, x_{iC}) \equiv N^{-1} \sum_i \Phi(\mu + f(x_S, x_{iC})) \\
p_m(x_S) &\equiv N^{-1} \sum_i p_m(x_S, x_{iC}) \\
\hat{p}(x_S) &\equiv M^{-1} \sum_m p_m(x_S)
\end{aligned}$$

gbart and mc.gbart input and output

```
post <- gbart(x.train, y.train, type="pbart", ...,
        ndpost=M, keepevery=10) or
post <- mc.gbart(x.train, y.train, type="pbart", ...,
        ndpost=M, keepevery=10, mc.cores=2, seed=99)
```

Input matrices: `x.train` and, optionally, `x.test`: $\boldsymbol{x}_i$

$$\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$$

Output object, `post`, of type `pbart` (essentially a list)

Matrices: `post$prob.train` and, optionally, `post$prob.test`:

$$\hat{p}_{im} = \Phi(\mu + f_m(\boldsymbol{x}_i))$$

$$\begin{bmatrix} \hat{p}_{11} & \dots & \hat{p}_{N1} \\ \vdots & \vdots & \vdots \\ \hat{p}_{1M} & \dots & \hat{p}_{NM} \end{bmatrix}$$

`predict.pbart` input and output

```
pred <- predict(post, x.test, mc.cores=1, ...)
```

Input matrices: `x.test`: $\boldsymbol{x}_i$

$$\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_Q \end{bmatrix}$$

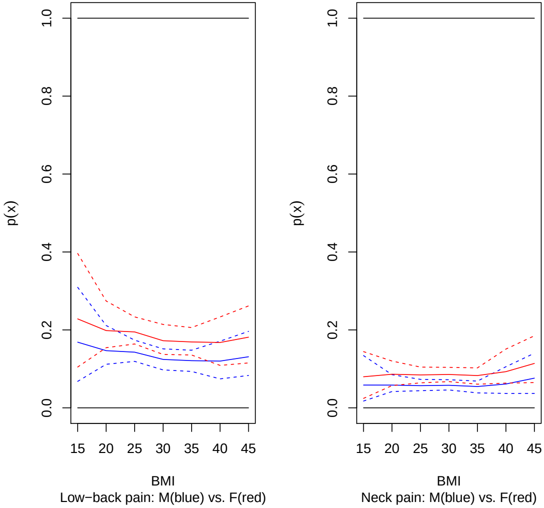
Output list with `prob.test`: $\hat{p}_{im} = \Phi(\mu + f_m(\boldsymbol{x}_i))$

$$\begin{bmatrix} \hat{p}_{11} & \dots & \hat{p}_{Q1} \\ \vdots & \vdots & \vdots \\ \hat{p}_{1M} & \dots & \hat{p}_{QM} \end{bmatrix}$$

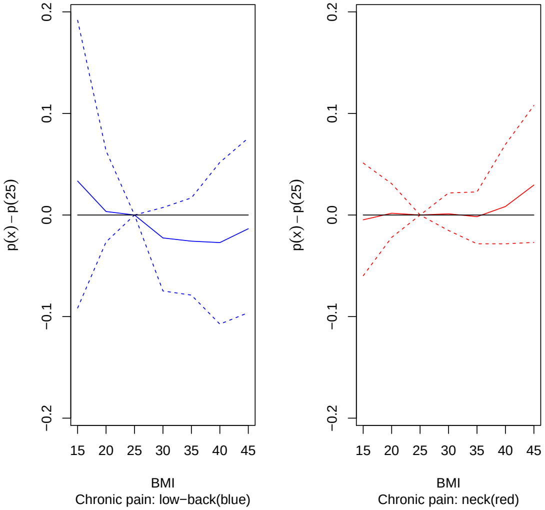
Demo: chronic spine pain and obesity

- Hypothesis a: obesity is a risk factor for chronic lower back/buttock pain
- Hypothesis b: obesity is NOT a risk factor for chronic neck pain
- `system.file('demo/nhanes.pbart1.R', package='BART')`
- `system.file('demo/nhanes.pbart2.R', package='BART')`

Friedman's partial dependence function: Probability of chronic pain vs. BMI



Friedman's partial dependence function: Probability of chronic pain vs. BMI



Logistic BART for binary outcomes

Logistic regression with latent variables

Devroye 1986 *Non-uniform random variate generation*

Holmes & Held 1993 *Bayesian Analysis*

Gramacy & Polson 2012 *Bayesian Analysis*

$$y_i|p_i \overset{\text{ind}}{\sim} \mathrm{B}(p_i)$$

$$p_i|f = \Phi(\mu + f(x_i)) \text{ where } f \overset{\text{prior}}{\sim} \text{BART } (\mu) \text{ and } \mu = \Phi^{-1}(\bar{y})$$

$$z_i|y_i, f, \sigma_i \sim \mathrm{N}\left(\mu + f(x_i),\ \sigma_i^2\right) \begin{cases} \mathrm{I}(-\infty, 0) & \text{if } y_i = 0 \\ \mathrm{I}(0, \infty) & \text{if } y_i = 1 \end{cases}$$

$$\sigma_i^2 = 4\psi_i^2 \text{ where } \psi_i \sim \text{Kolmogorov-Smirnov (see Devroye)}$$

Continuous BART with heteroskedastic variance and z_i is the data

Geweke convergence diagnostics for binary BART

Hastings 1970 *Biometrika*, Silverman 1986 *Chapman and Hall*

$$\hat{\theta}_M = M^{-1} \sum_{m=1}^{M} \theta_m$$

Bayesian estimator

$$\sigma^2_{\hat{\theta}} = \lim_{M \to \infty} \mathrm{V}\left[\hat{\theta}_M\right]$$

Asymptotic variance

Suppose θ_m is an $\mathrm{ARMA}\ (p, q)$

$$\gamma(w) = (2\pi)^{-1} \sum_{m=-\infty}^{\infty} \mathrm{V}\left[\theta_0, \theta_m\right] \mathrm{e}^{imw}$$

Spectral density

$$\hat{\sigma}^2_{\hat{\theta}} = \hat{\gamma}^2(0)$$

Variance estimator

Geweke convergence diagnostics for binary BART

Geweke 1992 *Bayesian Statistics*

- Divide your chain into two segments: A and B
- $m \in A = \{1, \ldots, M_A\}$ where $M_A = aM$
- $m \in B = \{M - M_B + 1, \ldots, M\}$ where $M_B = bM$
- $a + b < 1$, Geweke suggests $a = 0.1$ and $b = 0.5$

$$\hat{\theta}_A = M_A^{-1} \sum_{m \in A} \theta_m \qquad \hat{\theta}_B = M_B^{-1} \sum_{m \in B} \theta_m$$

$$\hat{\sigma}^2_{\hat{\theta}_A} = \hat{\gamma}^2_{m \in A}(0) \qquad \hat{\sigma}^2_{\hat{\theta}_B} = \hat{\gamma}^2_{m \in B}(0)$$

$$z = \frac{\sqrt{M}(\hat{\theta}_A - \hat{\theta}_B)}{\sqrt{a^{-1}\hat{\sigma}^2_{\hat{\theta}_A} + b^{-1}\hat{\sigma}^2_{\hat{\theta}_B}}} \qquad \sim \mathrm{N}(0,\ 1)$$

Geweke convergence diagnostics for binary BART

- We have a z_i corresponding to each $\theta_i = h(\mu + f(x_i))$
- In the **BART** R package, we created the `gewekediag` function which was adapted from the **coda** R package
  Plummer, Best et al. 2006

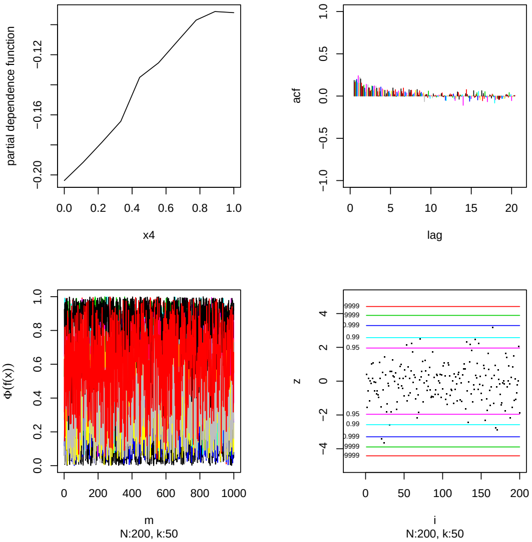
```
system.file('demo/geweke.pbart2.R', package='BART')
```

Geweke convergence diagnostics for binary BART: simulated data scenario

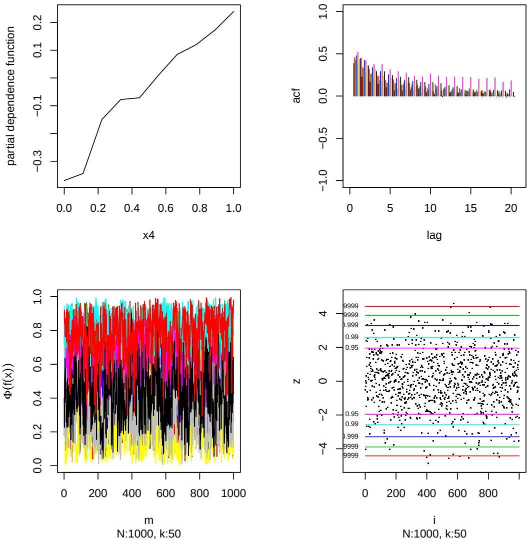
```
system.file('demo/geweke.pbart2.R', package='BART')
```

$$
\begin{aligned}
N &= 200, 1000, 10000 && \text{sample sizes} \\
K &= 50 && \text{number of covariates} \\
f(x_i) &= -1.5 + \sin(\pi x_{1i} x_{2i}) + 2(x_{3i} - 0.5)^2 + x_4 + 0.5x_5 \\
z_i &\sim \mathrm{N}(f(x_i),\ 1) \\
y_i &= \mathrm{I}(z_i > 0)
\end{aligned}
$$

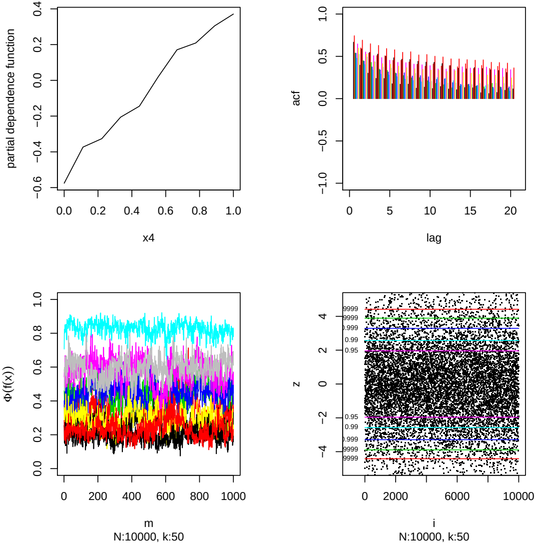
Geweke convergence diagnostics for binary BART: $N = 200$



Geweke convergence diagnostics for binary BART: $N = 1000$



Geweke convergence diagnostics for binary BART: $N = 10000$



Multinomial BART with logit link

`mbart2` function for a larger number of categories

Sparapani, Spanbauer and McCulloch 2021 *JSS*

- $y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \sim \textbf{Multinomial}\,(n,\ p)\ \ \textbf{where}\ p = \begin{bmatrix} p_1 \\ \vdots \\ p_k \end{bmatrix}$
- $n = \sum_j y_j$ and $\sum_j p_j = 1$
- If $n = 1$, computing Multinomial BART is facilitated by modeling the binary outcomes with k logistic BARTs

$$y_{ij} \sim \mathrm{B}(p_{ij})\ \ \textbf{where}\ f_j \overset{\text{prior}}{\sim} \textbf{BART}\,(\mu_j) \text{ and } p_{ij} \propto F(\mu_j + f_j(x_i))$$

- And then combining the inference as follows

$p_{ij} = \frac{\exp(\mu_j+f_j(x_i))}{\sum_{j'} \exp(\mu_j+f_j(x_i))}$ (but each fit is slow and we need k of them)

- This would work with the probit link (and it would be much faster), but there is no theoretical basis for combining probits in this way
- Or another alternative (that also doesn't follow from theory)
- $\tilde{p}_{ij} = \frac{\Phi(\mu_j+f_j(x_i))}{\sum_{j'} \Phi(\mu_j+f_j(x_i))}$

Multinomial BART with probit link

`mbart` function for a smaller number of categories

Sparapani, Spanbauer and McCulloch 2021 *JSS*

- If $n = 1$, fit a sequence of binary probit models (this bears some resemblance to continuation-ratio logits)
- assume k categories where each are represented by mutually exclusive binary indicators: $y_{i1}, \dots, y_{ik}$
- the probability of these outcomes, p_{ij}, where $j = 1, \dots, k$

$$\begin{aligned}
p_{i1} &= \mathrm{P}[y_{i1} = 1] \\
p_{i2} &= \mathrm{P}[y_{i2} = 1 | y_{i1} = 0] \\
p_{i3} &= \mathrm{P}[y_{i3} = 1 | y_{i1} = y_{i2} = 0] \\
&\vdots \\
p_{i,k-1} &= \mathrm{P}[y_{i,k-1} = 1 | y_{i1} = \cdots = y_{i,k-2} = 0] \\
p_{ik} &= \mathrm{P}[y_{i,k-1} = 0 | y_{i1} = \cdots = y_{i,k-2} = 0]
\end{aligned}$$

Notice that $p_{ik} = 1 - p_{i,k-1}$ so we can specify the k conditional probabilities via $k - 1$ parameters

Multinomial BART with probit link

`mbart` function for a smaller number of categories

- these conditional probabilities are, by construction, defined for subsets of subjects: let $S_1 = \{1, \dots, N\}$ and $S_j = \{i : y_{i1} = \dots = y_{i,j-1} = 0\}$ where $j = 2, \dots, k-1$
- the unconditional probability of these outcomes, π_{ij}, can be defined in terms of the conditional probablities and their complements, $q_{ij} = 1 - p_{ij}$, for all subjects

$$
\begin{aligned}
\pi_{i1} &= \mathrm{P}[y_{i1} = 1] = p_{i1} \\
\pi_{i2} &= \mathrm{P}[y_{i2} = 1] = p_{i2} q_{i1} \\
\pi_{i3} &= \mathrm{P}[y_{i3} = 1] = p_{i3} q_{i2} q_{i1} \\
&\vdots \\
\pi_{i,k-1} &= \mathrm{P}[y_{i,k-1} = 1] = p_{i,k-1} q_{i,k-2} \cdots q_{i1} \\
\pi_{ik} &= \mathrm{P}[y_{ik} = 1] = q_{i,k-1} q_{i,k-2} \cdots q_{i1}
\end{aligned}
$$

N.B. the rules of probability ensure that $\sum_{j=1}^{k} \pi_{ij} = 1$

Multinomial BART with probit link
Alligator food choice: `demo/alligator.R`

- 219 alligators were taken by hunters in 1985 from 4 Florida lakes
- From 1 to 4 meters long, their stomachs were removed for study
- Each gator's primary food choice was determined
  5 categories: bird, fish, invertebrate, reptile or other
- Covariates: lake, sex, and size (small vs. large)