Introduction to dichotomous/categorical outcomes with BART

Rodney Sparapani
Medical College of Wisconsin

September 15, 2025

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
- Categorical outcomes with BART
- Convergence diagnostics for dichotomous 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
- US 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*

$$y_i \overset{\text{ind}}{\sim} \mathrm{B}(p(x_i))$$

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

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

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

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

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*

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

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`:

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

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

predict.pbart input and output

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

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

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

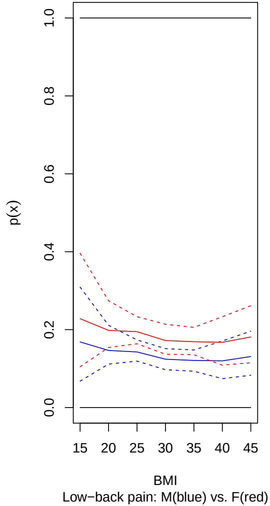
Output list with prob.test: $\hat{p}_{im} = \Phi(f_m(\boldsymbol{x}_i))$

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

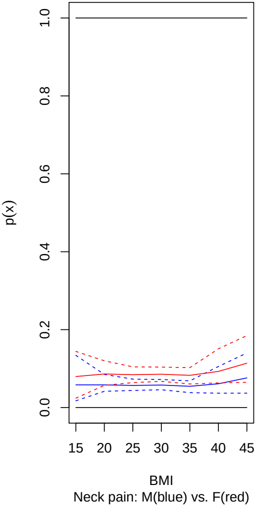
Demo: chronic spine pain and obesity

- 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')`
- See the `arq` data set (Arthritis Questionnaire)
- Covariates: sex (riagendr), age (ridageyr) and BMI (bmxbmi)
- riagendr: 1 for males, 2 for females

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

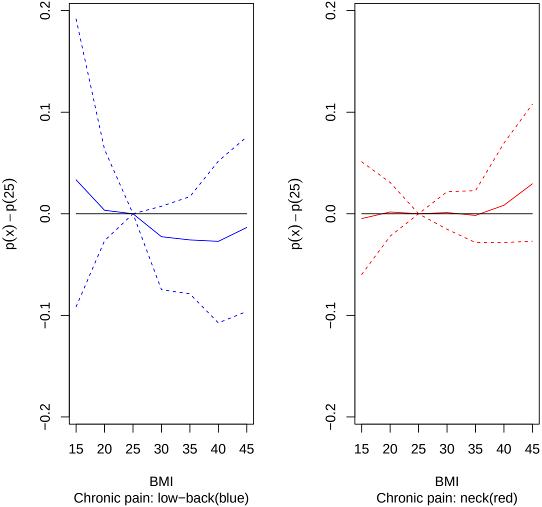


Low−back pain: M(blue) vs. F(red)



Neck pain: M(blue) vs. F(red)

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



Logistic BART for dichotomous outcomes

Logistic regression with latent variables

Devroye 1986 *Non-uniform random variate generation*

Holmes & Held 1993 *Bayesian Analysis*

Gramacy & Polson 2012 *Bayesian Analysis*

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

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

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

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

Continuous BART with heteroskedastic variance and z_i is the data

Categorical BART

Agarwal, Ranjan & Chipman 2013 *Can J Remote Sensing*

- This is referred to as the “one vs. all” approach
- Assume we have more than 2 categories $y_i \in \{1, \ldots, k\}$
- Fit a sequence of k probit (or logit) BART models

$$
\begin{aligned}
y_{ij} &= \mathrm{I}(y_i = j) & \bar{y}_{.j} &= N^{-1} \sum_i y_{ij} \\
\mu_j &= \Phi^{-1}(\bar{y}_{.j}) & & \\
\tilde{p}_{i1} &= \mathrm{P}[y_{i1} = 1] = \Phi(f_1(x_i)) & f_1 &\overset{\text{prior}}{\sim} \text{BART}\ (\mu_1) \\
\tilde{p}_{i2} &= \mathrm{P}[y_{i2} = 1] = \Phi(f_2(x_i)) & f_2 &\overset{\text{prior}}{\sim} \text{BART}\ (\mu_2) \\
&\vdots & & \\
\tilde{p}_{ik} &= \mathrm{P}[y_{ik} = 1] = \Phi(f_k(x_i)) & f_k &\overset{\text{prior}}{\sim} \text{BART}\ (\mu_k)
\end{aligned}
$$

- Prediction: $\tilde{y}_i = \arg\max_j \tilde{p}_{ij}$
- Let $p_{ij} = \tilde{p}_{ij} / \sum_{j'} \tilde{p}_{ij'}$

Convergence diagnostics for dichotomous BART

Hastings 1970 *Biometrika*

Silverman 1986 *Density Estimation for Statistics and Data Analysis*

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

Bayesian estimator

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

Asymptotic variance

Suppose θ_m is an **ARMA** (p, q)

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

Spectral density

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

Variance estimator

Convergence diagnostics for dichotomous BART

Geweke 1992 *Bayesian Statistics*

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

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

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

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

Convergence diagnostics for dichotomous BART

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

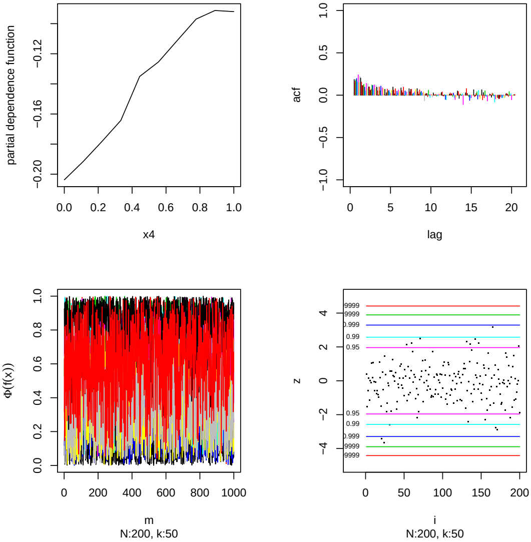
```
system.file('demo/geweke.pbart2.R', package='BART')
```

Convergence diagnostics for dichotomous BART: simulated data scenario

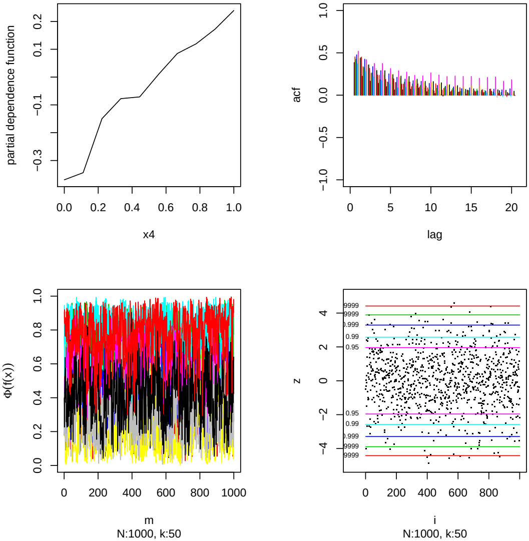
```
system.file('demo/geweke.pbart2.R', package='BART')
```

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

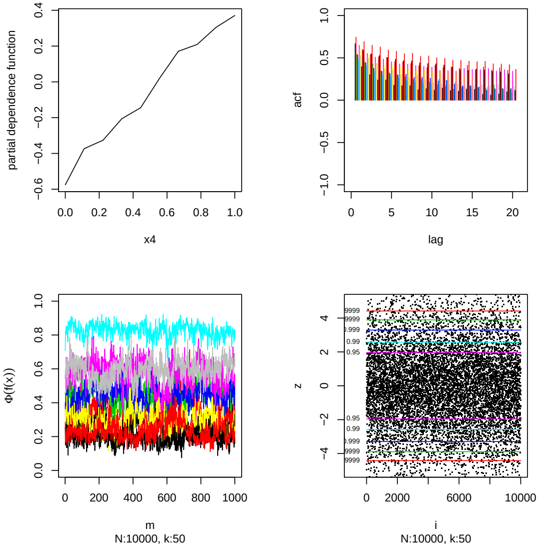
Convergence diagnostics for dichotomous BART: $N = 200$



Convergence diagnostics for dichotomous BART: $N = 1000$



Convergence diagnostics for dichotomous BART: $N = 10000$



Convergence diagnostics for dichotomous BART: a modern alternative?

- the Geweke method is time-tested
- but it feels a bit dated today
- inspired by FPD: I have used the following approach instead
- $\theta_m^{\text{max}} = \max f_m(x_i)$ assessed with `maxRhat`
- this also works well for time-to-event outcomes
- however, it might be sensitive to outliers
- so you should consider several quantities
- such as $\theta_m^{\text{min}} = \min f_m(x_i)$
- and $\theta_m^{\text{median}} = \text{median}\, f_m(x_i)$

MCMC convergence diagnostics with $\hat{R}$ or `Rhat`

- we have roughly 3 generations of `Rhat`
- the original development of the statistic we call `oldRhat`
- Gelman and Rubin 1992; Brooks and Gelman 1998;
  Bayesian Data Analysis (BDA) 1st/2nd ed. by Gelman et al.
- `oldRhat` $< \mathbf{1.1}$ convergence **DON'T USE: too liberal**
- an improvement we call `splitRhat`: BDA 3rd ed.
- the latest and greatest which we call `maxRhat`
  Vehtari, Gelman et al. 2021 *Bayesian Analysis*
- see `Rhat.R` in the **BART3** R package
- we should use `maxRhat` which is the most robust
- `maxRhat` $< \mathbf{1.01}$ convergence (1.1 might be better for BART?)
- standard advice: to get M samples, we generate $2M$ samples and discard the first half M (called burn-in) since the beginning may be sensitive to initial starting values
- but the point is to check convergence with diagnostics

MCMC convergence diagnostics with `splitRhat`

- Compute at least $C = 2$ chains and split each chain into two halves: $D = 2C$ sub-chains
- Each sub-chain with L samples for a total of $DL = M$
- For ALL θ: converged if $\hat{R} < 1.01$ (not proof, but probable)

$$\bar{\theta}_{.j} = L^{-1} \sum_{i=1}^{L} \theta_{ij} \qquad \bar{\theta}_{..} = D^{-1} \sum_{j=1}^{D} \bar{\theta}_{.j}$$

$$B = \frac{L}{D-1} \sum_{j} (\bar{\theta}_{.j} - \bar{\theta}_{..})^2$$

$$W = D^{-1} \sum_{j} s_j^2 \qquad s_j^2 = (L-1)^{-1} \sum_{i} (\theta_{ij} - \bar{\theta}_{.j})^2$$

$$\hat{R} = \sqrt{\widehat{\mathrm{var}}/W}$$

$$\text{where } \widehat{\mathrm{var}} = L^{-1}\left[(L-1)W + B\right]$$

MCMC convergence diagnostics: `maxRhat`

- `splitRhat` is essentially ANOVA based on Normal errors
- θ might not be Normal, i.e., the posterior is not necessarily Normal with respect to θ which is a key tenet of Bayesianism small sample size or non-Normal due to the prior/likelihood
- Compute at least $C = 4$ chains and split each chain into two halves: $D = 2C$ sub-chains
- Compute `splitRhat` with rank Normalized $\tilde{\theta}$
- $\tilde{\theta}_{ij} = \Phi^{-1}\left(\frac{\text{rank}(\theta_{ij})-0.5}{DL}\right)$
- Compute Folded `splitRhat` with rank Normalized $\tilde{\zeta}$
- $\zeta_{ij} = |\theta_{ij} - Q_2| \text{ where } Q_2 = \text{median}\ \theta_{ij}$
- $\tilde{\zeta}_{ij} = \Phi^{-1}\left(\frac{\text{rank}(\zeta_{ij})-0.5}{DL}\right)$
- `maxRhat` = max (`splitRhat` for $\tilde{\theta}_{ij}$, `splitRhat` for $\tilde{\zeta}_{ij}$)

MCMC convergence diagnostics with Effective Sample Size (ESS)

- we have roughly 3 generations of ESS corresponding to `Rhat`
- (ESS not to be confused with Emacs Speaks Statistics)
- the orginal development of the statistic we call N_{eff}
- BDA 1st/2nd ed.
- an improvement we call S_{eff} or `Seff`: BDA 3rd ed.
- the latest and greatest which we call `minSeff` inspired by `maxRhat`
- `Seff` and `minSeff` are calculated with functions `splitRhat` and `maxRhat` respectively
- we should use `minSeff` which is the most robust
- `minSeff` = min (`Seff` for $\tilde{\theta}_{ij}$, `Seff` for $\tilde{\zeta}_{ij}$)

MCMC convergence diagnostics
Effective Sample Size: N_{eff} and S_{eff}

S_{eff} is more conservative than previous formulas such as N_{eff}

$$N_{\text{eff}} = \frac{L}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{L}{1 + 2\sum_{t=1}^{\infty} \rho_t} \text{ NO LONGER RECOMMENDED}$$

$$S_{\text{eff}} = DL\hat{\tau}^{-1}$$

$$\hat{\tau} = 1 + 2\sum_{t=1}^{2k+1} \hat{\rho}_t \text{ where } \hat{\rho}_t = 1 - \frac{W - D^{-1}\sum_{j=1}^{D} \hat{\rho}_{tj}}{\widehat{\text{var}}}$$

$$= -1 + 2\sum_{t'=0}^{k} \hat{P}_{t'} \text{ where } \hat{P}_{t'} = \hat{\rho}_{2t'} + \hat{\rho}_{2t'+1}$$

N.B. choose the largest k such that $\hat{P}_{t'} > 0$

I find it much easier to program S_{eff} in terms of $\hat{\rho}_t$ rather than $\hat{P}_{t'}$