Binary and categorical outcomes with BART

Rodney Sparapani
Medical College of Wisconsin

Robert McCulloch
Arizona State University

Abstract

This short article illustrates how to analyze binary and categorical outcomes with the **BART** R package.

Keywords: Bayesian Additive Regression Trees.

1. Binary and categorical outcomes with BART

The **BART** package supports binary outcomes via probit BART with Normal latents and logistic BART with Logistic latents. Categorical outcomes are supported with multinomial BART with Logistic latents. Convergence diagnostics are provided and variable selection as well.

1.1. Probit BART for binary outcomes

To extend BART to binary outcomes, we employ the technique of Albert and Chib (1993) to create what we call probit BART. Probit BART is provided by the **BART** package as the `pbart` function. In this case, the outcome, `y.train`, is provided as an integer with values 0 or 1. Given y_i, we introduce the truncated Normal latents, z_i; these auxiliary latents are efficiently sampled (Robert 1995) and recast as the outcome for a continuous BART with unit variance where i indexes subject and Φ is the standard Normal cumulative distribution function.

$$\begin{aligned} y_i|p_i &\overset{\text{ind}}{\sim} \mathrm{B}\left(p_i\right) \\ p_i|f &= \Phi(\mu_0 + f(\boldsymbol{x}_i)) \text{ where } f \overset{\text{prior}}{\sim} \text{BART} \\ z_i|y_i, f &\sim \mathrm{N}\left(\mu_0 + f(\boldsymbol{x}_i),\ 1\right) \begin{cases} \mathrm{I}\left(-\infty, 0\right) & \text{if } y_i = 0 \\ \mathrm{I}\left(0, \infty\right) & \text{if } y_i = 1 \end{cases} \end{aligned}$$

The z_i are centered around a known constant, μ_0, which is tantamount to centering the probabilities, p_i, around $p_0 = \Phi(\mu_0)$. If $\mu_0 = 0$, which is the default, then the p_i are centered around 0.5; to specify a different value, say -1, pass the argument `binaryOffset=-1` in the `pbart` call. The key insight into the probit BART technique is that the Gibbs conditional $f|z_i, y_i \overset{d}{=} f|z_i$, i.e., given z_i, y_i is unnecessary. This setup leads to the following Bernoulli likelihood: $[\boldsymbol{y}|f] = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$.

In the following, we assume that `binaryOffset=0` for convenience (which is the default). The `pbart` (`mc.pbart`) function is for serial (parallel) computation. The outcome `y.train` is a vector containing zeros and ones. The covariates for training (validation, if any) are `x.train` (`x.test`) which can be matrices or data frames containing factors; in the display below, we assume matrices for simplicity.

```
set.seed(99)
post <- pbart(x.train, y.train, x.test, ..., ndpost=M) or
post <- mc.pbart(x.train, y.train, x.test, ..., ndpost=M, mc.cores=2, seed=99)
```

Input matrices: `x.train` and, optionally, `x.test`: $\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$ or $\boldsymbol{x}_i$

`post`, of type `pbart`, which is essentially a list

`post$yhat.train` and `post$yhat.test`: $\begin{bmatrix} \hat{y}_{11} & \cdots & \hat{y}_{N1} \\ \vdots & \cdots & \vdots \\ \hat{y}_{1M} & \cdots & \hat{y}_{NM} \end{bmatrix}$ where $\hat{y}_{im} = f_m(\boldsymbol{x}_i)$

The columns of `post$yhat.train` and `post$yhat.test` represent different covariate settings and the rows, the `M` draws from the posterior. Although, `post$yhat.train` and `post$yhat.test`, when requested, are returned, generally, `post$prob.train` and `post$prob.test` are of more interest (and `post$prob.train.mean` and `post$prob.test.mean` which are the means of the posterior sample columns, not shown).

`post$prob.train` and `post$prob.test`: $\begin{bmatrix} \hat{p}_{11} & \cdots & \hat{p}_{N1} \\ \vdots & \cdots & \vdots \\ \hat{p}_{1M} & \cdots & \hat{p}_{NM} \end{bmatrix}$ where $\hat{p}_{im} = \Phi(f_m(\boldsymbol{x}_i))$

Often it is impractical to provide `x.test` in the call to `pbart` due to the number of predictions considered or all the settings to evaluate are simply not known at that time. To allow for this common problem, the **BART** package returns the trees encoded in an ASCII string, `treedraws$trees`, and provides a `predict` function to generate any predictions needed. Note that if you need to perform the prediction in some later R instance, then you can save the `pbart` object returned and reload it when needed, e.g., save with `saveRDS(post, 'post.rds')` and reload, `post <- readRDS('post.rds')` . The `x.test` input can be a matrix or a data frame; for simplicity, we assume a matrix below.

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

$$\text{Input: } \texttt{x.test}: \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_Q \end{bmatrix} \text{ or } \boldsymbol{x}_h$$

`pred`, of type `pbart`, which is essentially a list

$$\texttt{pred\$yhat.test}: \begin{bmatrix} \hat{y}_{11} & \cdots & \hat{y}_{Q1} \\ \vdots & \vdots & \vdots \\ \hat{y}_{1M} & \cdots & \hat{y}_{QM} \end{bmatrix} \text{ where } \hat{y}_{hm} = f_m(\boldsymbol{x}_h)$$

$$\texttt{pred\$prob.test}: \begin{bmatrix} \hat{p}_{11} & \cdots & \hat{p}_{Q1} \\ \vdots & \vdots & \vdots \\ \hat{p}_{1M} & \cdots & \hat{p}_{QM} \end{bmatrix} \text{ where } \hat{p}_{hm} = \Phi(f_m(\boldsymbol{x}_h))$$

$$\texttt{pred\$prob.test.mean}: [\hat{p}_1, \ldots, \hat{p}_Q] \text{ where } \hat{p}_h = M^{-1} \sum_{m=1}^{M} \hat{p}_{hm}$$

1.2. Friedman's partial dependence function

BART does not directly provide a summary of the effect of a single covariate, or a subset of covariates, on the outcome. This is also the case for black-box, or nonparametric regression, models in general which have had to deal with this issue. We recommend Friedman's partial dependence function (Friedman 2001) with BART to summarize the marginal effect due to a subset of the covariates, $\boldsymbol{x}_S$, by aggregating over the complement covariates, $\boldsymbol{x}_C$, i.e., $\boldsymbol{x} = [\boldsymbol{x}_S, \boldsymbol{x}_C]$. The marginal dependence function is defined by fixing $\boldsymbol{x}_S$ while aggregating over the observed settings of the complement covariates in the cohort: $f(\boldsymbol{x}_S) = N^{-1} \sum_{i=1}^{N} f(\boldsymbol{x}_S, \boldsymbol{x}_{iC})$. For probit BART, the f function is not directly of interest; rather, the probability of an event is more interpretable: $p(\boldsymbol{x}_S) = N^{-1} \sum_{i=1}^{N} \Phi(\mu_0 + f(\boldsymbol{x}_S, \boldsymbol{x}_{iC}))$. Other marginal functions can be obtained in a similar fashion. Estimates can be derived via functions of the posterior samples such as means, quantiles, e.g., $\hat{p}(\boldsymbol{x}_S) = M^{-1} N^{-1} \sum_{m=1}^{M} \sum_{i=1}^{N} \Phi(\mu_0 + f_m(\boldsymbol{x}_S, \boldsymbol{x}_{iC}))$ where m indexes posterior samples. Friedman's partial dependence function is a concept that is very flexible. So flexible that we are unable to provide abstract functional support in the **BART** package; rather, we provide examples of the many practical uses in the `demo` directory.

1.3. Logistic BART for binary outcomes

Note that the distribution of the latent z_i is not identifiable from the data so it is essentially a parametric assumption. This assumption can be relaxed by assuming the latents follow the Logistic distribution which has heavier tails and, therefore, is a better choice if the p_i can be very close to zero or one. For Logistic latents, we employ the technique of Holmes and Held (2006) to create what we call logistic BART. However, it is important to recognize that logistic BART is more computationally intensive than probit BART.

The outcome, `y.train`, is provided as an integer with values 0 or 1. Logistic BART is provided by the `lbart` function. Unlike probit BART where the auxiliary latents, z_i, have

a fixed variance $\sigma^2 = 1$; with Logisitic BART, we sample truncated Normal latents, z_i, with a random variance σ_i^2 (Robert 1995). If $\sigma_i^2 = 4\psi_i^2$ where ψ_i is sampled from the Kolmogorov-Smirnov distribution, then z_i follow the Logistic distribution. Sampling from the Kolmogorov-Smirnov distribution is described by Devroye (1986). So, the conditionally Normal latents, $z_i|\sigma_i^2$, are the outcomes for a continuous BART with a known heteroskedastic variance, σ_i^2. Since Logistic latents are more flexible, there is no centering parameter, i.e., $\mu_0 = 0$. Therefore, the probabilities are $p_i = F(f(\boldsymbol{x}_i))$ where F is the standard Logistic distribution function.

The input and output for `lbart` is essentially identical to `pbart`. Also, the `predict` function for objects of type `lbart` is analogous.

1.4. Multinomial BART for categorical outcomes

To extend BART to categorical outcomes, we employ as many logistic BARTs as there are categories to create what we call multinomial BART. Multinomial BART is provided by the **BART** package as the `mbart` function. In this case, the outcome, `y.train`, is provided as an integer with values $1, \ldots, C$ which generate the corresponding latents $z_{i1}, \ldots, z_{iC}$ and probabilities $p_{i1}, \ldots, p_{iC}$ which are constrained to sum to one.

The input for `mbart` is essentially identical to `pbart`. The output is slightly different.

```
set.seed(99)
post <- mbart(x.train, y.train, x.test, ..., ndpost=M) or
post <- mc.mbart(x.train, y.train, x.test, ..., ndpost=M, mc.cores=2, seed=99)
```

Input: `x.train` and, optionally, `x.test`: $\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$ or $\boldsymbol{x}_i$

`post`, of type `mbart`

`post$prob.train` and `post$prob.test`: $\begin{bmatrix} \hat{p}_{111} & \cdots & \hat{p}_{1C1} & \cdots & \hat{p}_{N11} & \cdots & \hat{p}_{NC1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{p}_{11M} & \cdots & \hat{p}_{1CM} & \cdots & \hat{p}_{N1M} & \cdots & \hat{p}_{NCM} \end{bmatrix}$

where $\hat{p}_{icm} = F(f_{cm}(\boldsymbol{x}_i))$

The columns of `post$prob.train` and `post$prob.test` represent different covariate settings crossed with the C categories. Also, the `predict` function for objects of type `mbart` is analogous.

1.5. Converegence diagnostics for dichotomous and categorical outcomes

How do you perform convergence diagnostics for BART? For continuous outcomes, convegence can easily be determined from the trace plots of the the error variance, σ^2. However, for probit BART with Normal latents, the error variance is fixed at 1 so this is not an option. Similarly, for logistic and multinomial BART, σ_i^2, are auxiliary latent variables not suitable for convergence diagnostics. Therefore, we adapt traditional MCMC diagnostic approaches

to BART. We perform graphical checks via auto-correlation, trace plots and an approach due to Geweke (1992).

Geweke diagnostics are based on earlier work which characterizes MCMC as a time series (Hastings 1970). Once this transition is made, auto-regressive, moving-average (ARMA) process theory is employed (Silverman 1986). Generally, we define our Bayesian estimator as $\hat{\theta}_M = M^{-1}\sum_{m=1}^{M}\theta_m$. We represent the asymptotic variance of the estimator by $\sigma^2_{\hat{\theta}} = \lim_{M\to\infty} \mathrm{V}\left[\hat{\theta}_M\right]$. If we suppose that θ_m is an ARMA(p,q) process, then the spectral density of the estimator is defined as $\gamma(w) = (2\pi)^{-1}\sum_{m=-\infty}^{\infty}\mathrm{V}\left[\theta_0,\theta_m\right]\mathrm{e}^{\mathrm{i}mw}$ where $\mathrm{e}^{\mathrm{i}tw} = \cos(tw) + \mathrm{i}\sin(tw)$. This leads us to an estimator of the asymptotic variance which is $\hat{\sigma}^2_{\hat{\theta}} = \hat{\gamma}^2(0)$. We divide our chain into two segments, A and B, as follows: $m \in A = \{1,\ldots,M_A\}$ where $M_A = aM$; and $m \in B = \{M - M_B + 1,\ldots,M\}$ where $M_B = bM$. Note that $a + b < 1$. Geweke suggests $a = 0.1,\ b = 0.5$ and recommends the following Normal test for convergence.

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

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

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

In our **BART** package, we supply R functions adapted from the **coda** R package (Plummer, Best, Cowles, and Vines 2006) to perform Geweke diagnostics: `spectrum0ar` and `gewekediag`. But, how do we apply Geweke's diagnostic to BART? We can check convergence for any estimator of the form $\theta = h(f(\boldsymbol{x}))$, but often setting h to the identify function will suffice, i.e., $\theta = f(\boldsymbol{x})$. However, BART being a Bayesian nonparametric technique means that we have many potential estimators to check, i.e., essentially one estimator for every possible choice of $\boldsymbol{x}$.

We have supplied Figures 1, 2 and 3 generated by the example `geweke.pbart2.R`: `system.file('demo/geweke.pbart2.R', package='BART')`. The data are simulated by Friedman's five-dimensional test function (Friedman 1991) where 50 covariates are generated as $x_{ij}\sim\mathrm{U}\left(0,1\right)$ but only the first 5 covariates have an impact on the outcome at sample sizes $N = 100, 1000, 10000$.

$$\begin{aligned} f(\boldsymbol{x}_i) &= -1.5 + \sin(\pi x_{i1}x_{i2}) + 2(x_{i3} - 0.5)^2 + x_{i4} + 0.5x_{i5} \\ z_i &\sim \mathrm{N}\left(f(\boldsymbol{x}_i),\ 1\right) \\ y_i &= \mathrm{I}\left(z_i > 0\right) \end{aligned}$$

The convergence for each of these data sets is graphically displayed in Figures 1, 2 and 3 where each figure is broken into four quadrants. In the upper left quadrant, we have plotted Friedman's partial dependence function for $f(x_{i4})$ vs. x_{i4} for 10 values of x_{i4}. This is a check that can't be performed for real data, but it is informative in this case. Notice that $f(x_{i4})$ vs. x_{i4} is directly proportional in each figure as expected. In the upper right quadrant, we plot

the auto-correlations of $f(\boldsymbol{x}_i)$ for 10 randomly selected $\boldsymbol{x}_i$ where i indexes subjects. Notice that there is very little auto-correlation for $N = 100, 1000$, but a more notable amount for $N = 10000$. In the lower left quadrant, we display the corresponding trace plots for these same settings. The traces demonstrate that samples of $f(\boldsymbol{x}_i)$ appear to adequately traverse the sample space for $N = 100, 1000$, but less notably for $N = 10000$. In the lower right quadrant, we plot the Geweke Z_{AB} statistics for each subject i. Notice that for $N = 100$, the Z_{AB} exceed the 95% limits only a handful of times. Although, there are 10 times more comparisons, $N = 1000$ has seemingly more than 10 times as many values exceeding the 95% limits. And, for $N = 10000$, there are dramatically more values exceeding the 95% limits. Based on these figures, we conclude that the chains have converged for $N = 100$; for $N = 1000$, convergence is questionable; and, for $N = 10000$, convergence has not been attained. We would suggest that more thinning be employed for $N = 1000, 10000$ via the `keepevery` argument to `pbart`; perhaps, `keepevery=50` for $N = 1000$ and `keepevery=250` for $N = 10000$.

1.6. BART and variable selection

Several methods have been proposed for variable selection with BART (Chipman, George, and McCulloch 2010; Bleich, Kapelner, George, and Jensen 2014; Hahn and Carvalho 2015; McCulloch, Carvalho, and Hahn 2015; Linero 2016). The **BART** package supports the sparse prior of Linero (2016) by specifying `sparse=TRUE` (the default is `sparse=FALSE`). Let's represent the variable selection probabilities by s_j where $j = 1, \ldots, P$. Now, replace the uniform variable selection prior in BART with a Dirichlet prior. Also, place a Beta prior on the θ parameter.

$$[s_1, \ldots, s_P] \overset{\text{prior}}{\sim} \text{Dirichlet}\,(\theta/P, \ldots, \theta/P)$$
$$\frac{\theta}{\theta + \rho} \overset{\text{prior}}{\sim} \text{Beta}\,(a,\ b)$$

Typical settings are $b = 1$ and $\rho = P$ (the defaults) which you can over-ride with the `b` and `rho` arguments respectively. The value $a = 0.5$ (the default) is a sparse setting whereas an alternative setting $a = 1$ is not sparse or dense; you can specify this parameter with argument `a`. Linero discusses two assumptions: Assumption 2.1 and Assumption 2.2 (see Linero (2016) for more details). Basically, Assumption 2.2 (2.1) is more (less) friendly to binary/ordinal covariates and is (not) the default corresponding to `augment=FALSE` (`augment=TRUE`).

Let's return to the simulated probit BART example explored above which is in the **BART** package: `system.file('demo/sparse.pbart.R', package='BART')`. For sample sizes of $N = 100, 1000, 10000$, there are $P = 100$ covariates, but only the first 5 are active. In Figure 4, the 5 (95) active (inactive) covariates are red (black) and circles (dots) are $>$ ($\leq$) P^{-1} which is chance association represented by a black line. For $N = 100$, only $s_2 \leq P^{-1}$, but notice that there are 34 false positives. For $N = 1000$, all five active covariates are identified, but notice that there are 18 false positives. For $N = 10000$, all five active covariates are identified and notice that there are only two false positives.

We are often interested in the inter-relationship between covariates within our model. We can assess these relationships by inspecting the binary trees. For example, we can ascertain how often x_1 is chosen as a branch decision rule leading to a branch decision rule with x_2 further

up the tree or vice versa. In this case, we call x_1 and x_2 a concordant pair and we denote by $x_1 \leftrightarrow x_2$ which is a symmetric relationship, i.e., $x_1 \leftrightarrow x_2$ implies $x_2 \leftrightarrow x_1$. If B_h is the number of branches in tree T_h, then the concordant pair probability is: $\kappa_{ij} = \mathrm{P}\left[x_i \leftrightarrow x_j \in T_h | B_h > 1\right]$ for $i = 1, \ldots, P-1$ and $j = i+1, \ldots, P$. See an example of calculating these probabilities in `system.file('demo/trees.pbart.R', package='BART')`.

1.7. Motivating example: chronic pain and obesity

We want to test the hypothesis that obesity is a risk factor for chronic lower back pain (which includes buttock pain in this definition). A corollary to this hypothesis is that obesity is not considered to be a risk factor for chronic neck pain. A good source of data for this question is available in the National Health and Nutrition Examination Survey (NHANES) 2009-2010 Arthritis Questionnaire. 5106 subjects were surveyed. We will use probit BART to analyze the dichotomous outcomes of chronic lower back pain and chronic neck pain. We restrict our attention to the following covariates: age, gender and anthropometric measurements including weight (kg), height (cm), body mass index (kg/m^2) and waist circumference (cm). Also, note that sampling weights are available to extrapolate the rates from the survey to the US as a whole. We will concentrate on body mass index (BMI) and gender, $\boldsymbol{x}_S$, while utilizing Friedman's partial dependence function as defined above and also incorporating the sampling weights, i.e., $p_S(\boldsymbol{x}_S) = \sum_{i=1}^N w_i \Phi(\mu_0 + f(\boldsymbol{x}_S, \boldsymbol{x}_{iC}))/\sum_{i'=1}^N w_{i'}$.

The **BART** package provides two examples:
`system.file('demo/nhanes.pbart1.R', package='BART')` for chronic lower back pain and `system.file('demo/nhanes.pbart2.R', package='BART')` for chronic neck pain. In Figure 5, the unweighted relationship between chronic lower back pain, BMI and gender are displayed: males (females) are represented by blue (red) lines. As you can see, there is a non-linear relationship between the probability of chronic lower back pain and BMI for both genders where females have a parallel higher probability than males. For frail and underweight, the probability is high and drops as BMI increases until about 35 kg/m^2 and afterwards increases until about 65 kg/m^2 and then is flat. Based on sampling weights, the results are basically the same (not shown). In Figure 6, the unweighted relationship between chronic neck pain, BMI and gender are displayed: males (females) are represented by blue (red) lines. As you can see, there appears to be no relationship between the probability of chronic neck pain and BMI for both genders where females have a nearly parallel higher probability than males. Based on sampling weights (not shown), the results are basically the same.

References

Albert J, Chib S (1993). "Bayesian analysis of binary and polychotomous response data." *JASA*, **88**, 669–79.

Bleich J, Kapelner A, George EI, Jensen ST (2014). "Variable selection for BART: an application to gene regulation." *The Annals of Applied Statistics*, **8**(3), 1750–1781.

Chipman HA, George EI, McCulloch RE (2010). "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics*, **4**, 266–98.

Devroye L (1986). *Non-Uniform Random Variate Generation*. Springer.

Friedman JH (1991). "Multivariate adaptive regression splines (with discussion and a rejoinder by the author)." *Annals of Statistics*, **19**, 1–67.

Friedman JH (2001). "Greedy function approximation: a gradient boosting machine." *Annals of Statistics*, **29**, 1189–1232.

Geweke J (1992). *Bayesian Statistics*, chapter Evaluating the accuracy of sampling-based approaches to calculating posterior moments. fourth edition. Clarendon Press, Oxford, UK.

Hahn P, Carvalho C (2015). "Decoupling Shrinkage and Selection in Bayesian Linear Models: a Posterior Summary Perspective." *JASA*, **110**, 435–48.

Hastings W (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, **57**, 97–109.

Holmes C, Held L (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian Analysis*, **1**, 145–68.

Linero AR (2016). "Bayesian regression trees for high dimensional prediction and variable selection." *Journal of the American Statistical Association*, (<doi:10.1080/01621459.2016.1264957>).

McCulloch R, Carvalho C, Hahn R (2015). "A General Approach to Variable Selection Using Bayesian Nonparametric Models." Joint Statistical Meetings, Seattle, 08/09/15-08/13/15.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. [https://journal.r-project.org/archive].

Robert CP (1995). "Simulation of truncated normal variables." *Statistics and computing*, **5**(2), 121–125.

Silverman B (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Affiliation:

Rodney Sparapani rsparapa@mcw.edu
Division of Biostatistics, Institute for Health and Equity
Medical College of Wisconsin, Milwaukee campus

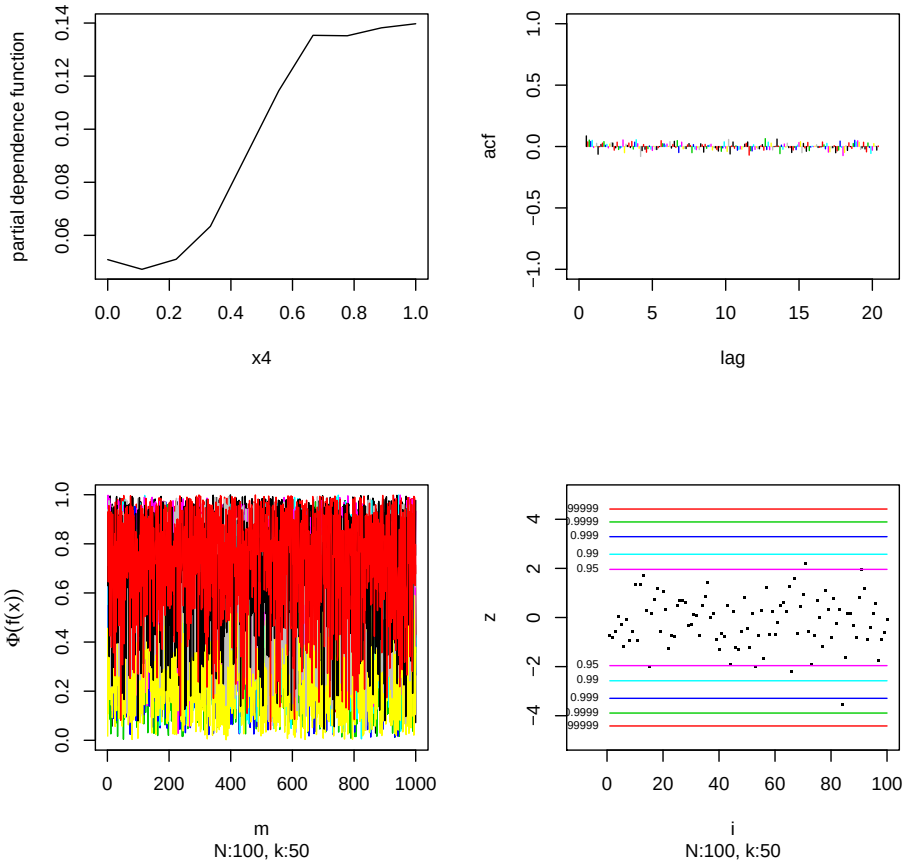

Figure 1: Geweke convergence diagnostics for probit BART: $N = 100$. In the upper left quadrant, we have plotted Friedman's partial dependence function for $f(x_{i4})$ vs. x_{i4} for 10 values of x_{i4}. This is a check that can't be performed for real data, but it is informative in this case. Notice that $f(x_{i4})$ vs. x_{i4} is directly proportional as expected. In the upper right quadrant, we plot the auto-correlations of $f(\boldsymbol{x}_i)$ for 10 randomly selected $\boldsymbol{x}_i$ where i indexes subjects. Notice that there is very little auto-correlation. In the lower left quadrant, we display the corresponding trace plots for these same settings. The traces demonstrate that samples of $f(\boldsymbol{x}_i)$ appear to adequately traverse the sample space. In the lower right quadrant, we plot the Geweke Z_{AB} statistics for each subject i. Notice that the Z_{AB} exceed the 95% limits only a handful of times. Based on this figure, we conclude that the chains have converged.

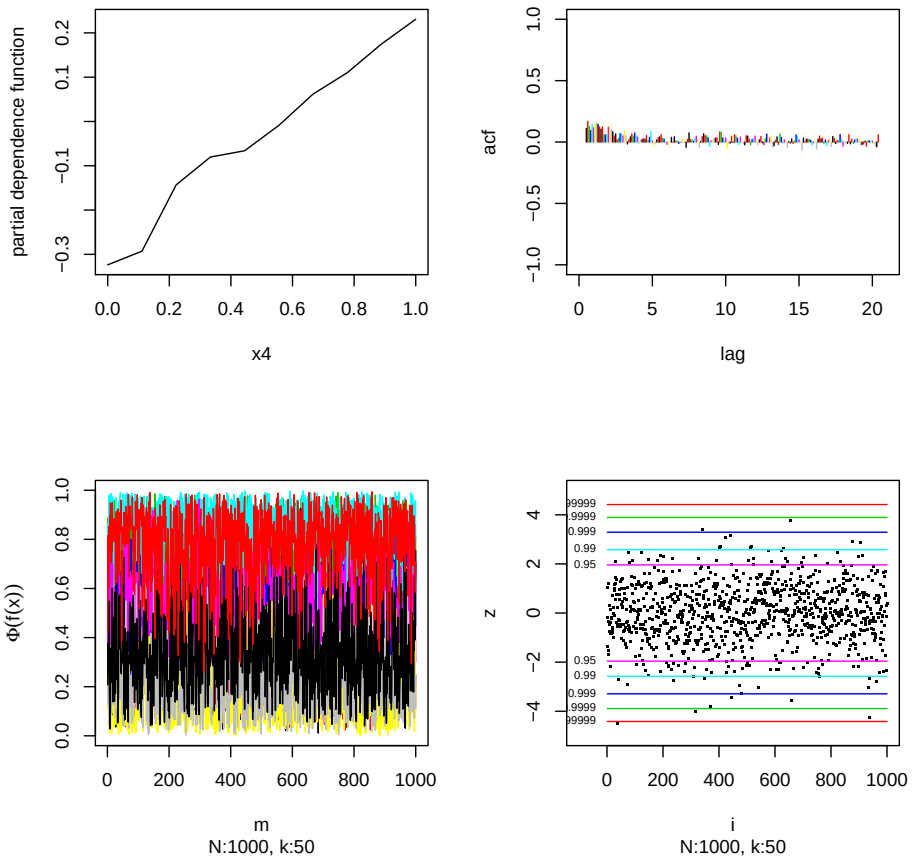

Figure 2: Geweke convergence diagnostics for probit BART: $N = 1000$. In the upper left quadrant, we have plotted Friedman's partial dependence function for $f(x_{i4})$ vs. x_{i4} for 10 values of x_{i4}. This is a check that can't be performed for real data, but it is informative in this case. Notice that $f(x_{i4})$ vs. x_{i4} is directly proportional as expected. In the upper right quadrant, we plot the auto-correlations of $f(\boldsymbol{x}_i)$ for 10 randomly selected $\boldsymbol{x}_i$ where i indexes subjects. Notice that there is very little auto-correlation. In the lower left quadrant, we display the corresponding trace plots for these same settings. The traces demonstrate that samples of $f(\boldsymbol{x}_i)$ appear to adequately traverse the sample space. In the lower right quadrant, we plot the Geweke Z_{AB} statistics for each subject i. Notice that there appear to be a considerable number exceeding the 95% limits. Based on this figure, we conclude that convergence is questionable. We would suggest that more thinning be employed via the `keepevery` argument to `pbart`; perhaps, `keepevery=50`.

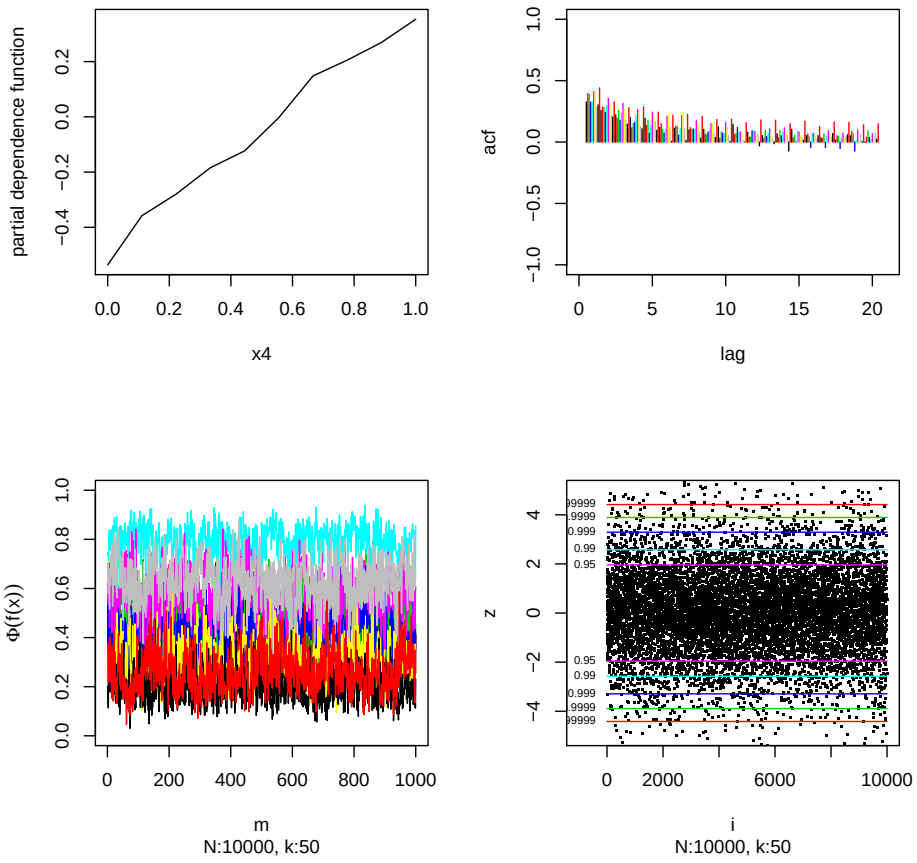

Figure 3: Geweke convergence diagnostics for probit BART: $N = 10000$. In the upper left quadrant, we have plotted Friedman's partial dependence function for $f(x_{i4})$ vs. x_{i4} for 10 values of x_{i4}. This is a check that can't be performed for real data, but it is informative in this case. Notice that $f(x_{i4})$ vs. x_{i4} is directly proportional as expected. In the upper right quadrant, we plot the auto-correlations of $f(\boldsymbol{x}_i)$ for 10 randomly selected $\boldsymbol{x}_i$ where i indexes subjects. Notice that there is some auto-correlation. In the lower left quadrant, we display the corresponding trace plots for these same settings. The traces demonstrate that samples of $f(\boldsymbol{x}_i)$ appear to traverse the sample space, but there are some slower oscillations. In the lower right quadrant, we plot the Geweke Z_{AB} statistics for each subject i. Notice that there appear to be far too many exceeding the 95% limits. Based on these figures, we conclude that convergence has not been attained. We would suggest that more thinning be employed via the `keepevery` argument to `pbart`; perhaps, `keepevery=250`.

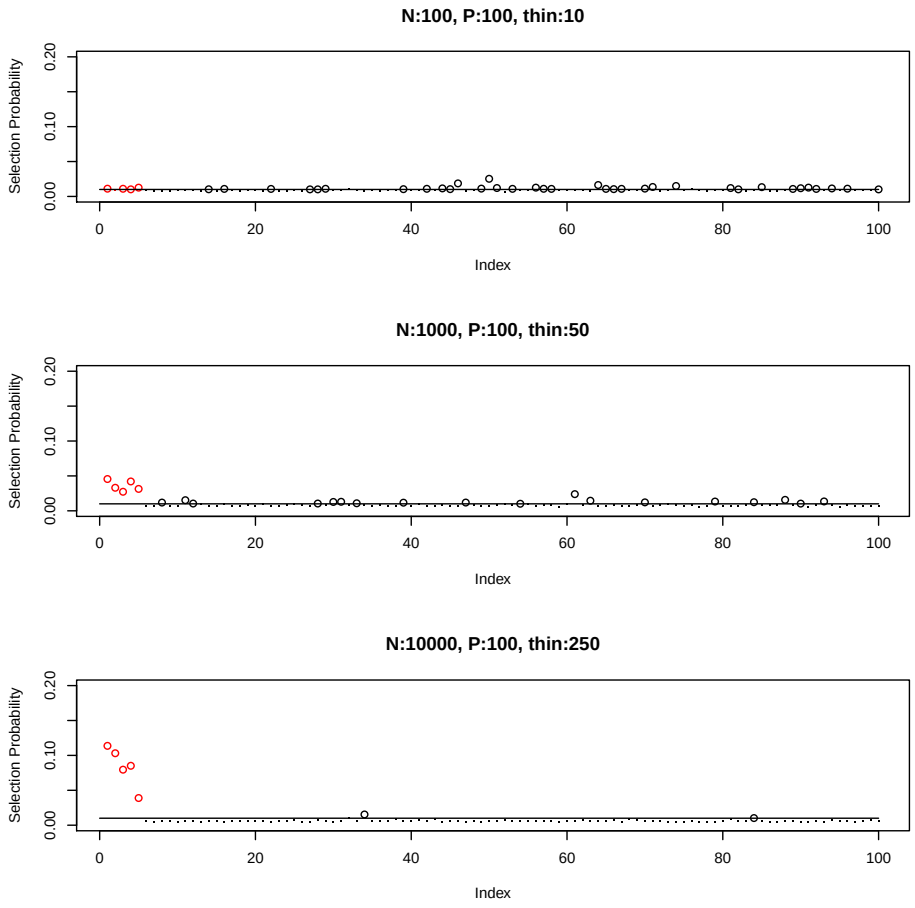

Figure 4: Probit BART and variable selection example. For sample sizes of $N = 100, 1000, 10000$, there are $P = 100$ covariates, but only the first 5 are active. The 5 (95) active (inactive) covariates are red (black) and circles (dots) are $>$ ($\leq$) P^{-1} which is chance association represented by a black line. For $N = 100$, only $s_2 \leq P^{-1}$, but notice that there are 34 false positives. For $N = 1000$, all five active covariates are identified, but notice that there are 18 false positives. For $N = 10000$, all five active covariates are identified and notice that there are only two false positives.

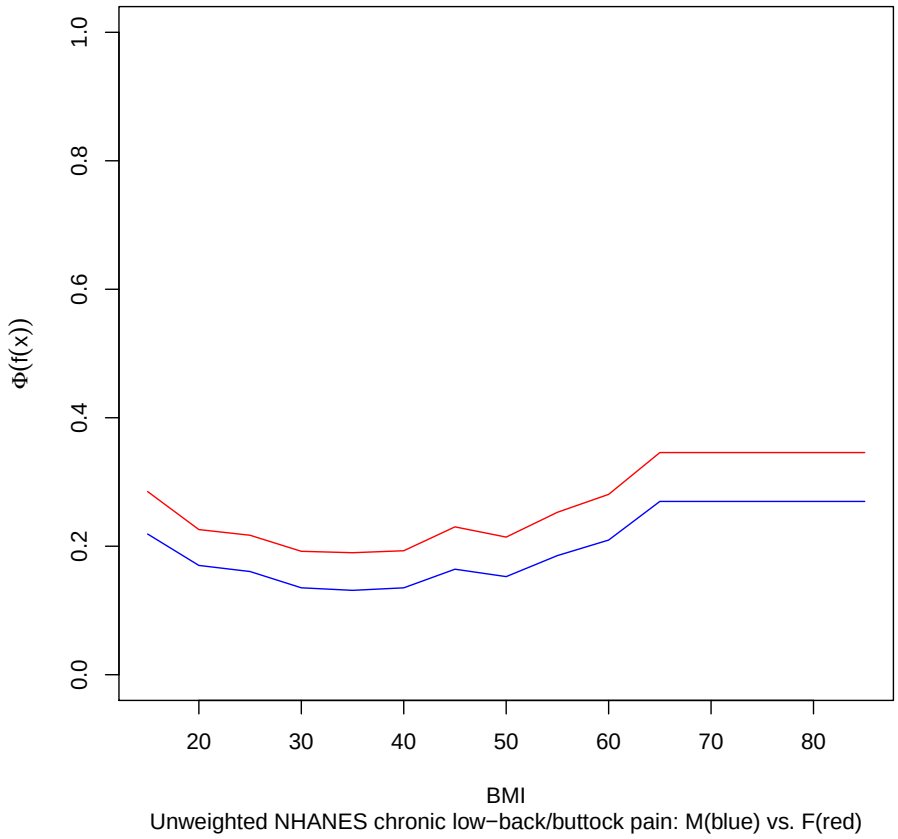

Figure 5: Friedman's partial dependence function: BMI and probability of chronic lower back pain. The unweighted relationship between chronic lower back pain, BMI and gender are displayed: males (females) are represented by blue (red) lines. As you can see, there is a non-linear relationship between the probability of chronic lower back pain and BMI for both genders where females have a parallel higher probability than males. For frail and underweight, the probability is high and drops as BMI increases until about 35 kg/m^2 and afterwards increases until about 65 kg/m^2 and then is flat. Based on sampling weights (not shown), the results are basically the same

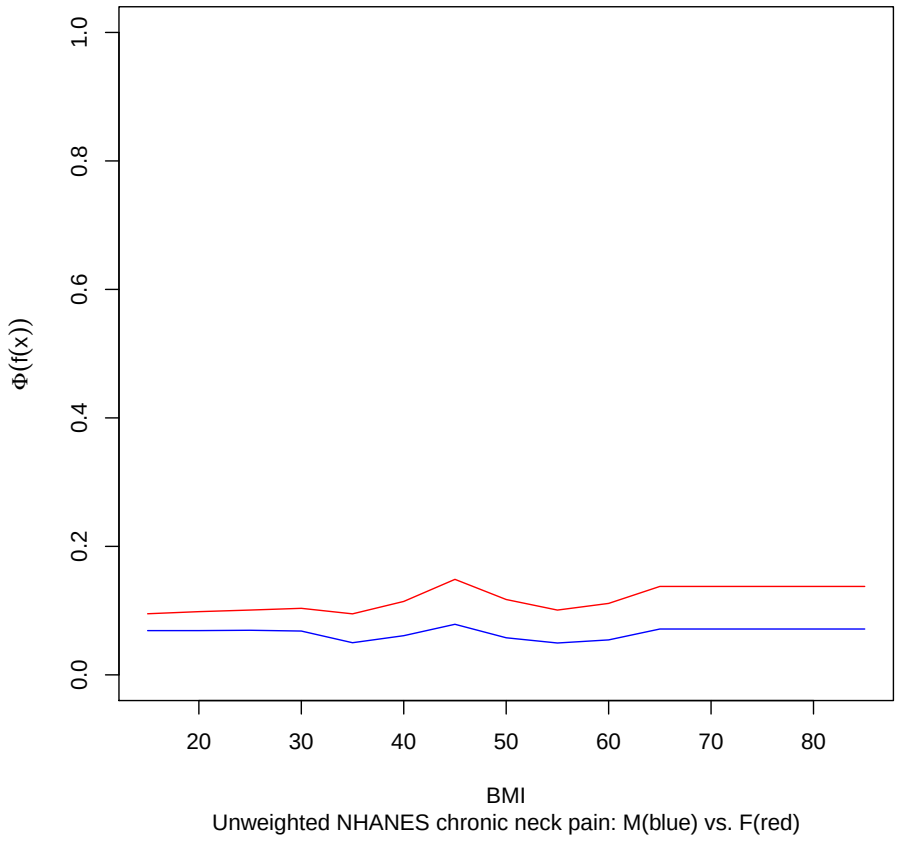

Figure 6: Friedman's partial dependence function: BMI and probability of chronic neck pain. The unweighted relationship between chronic neck pain, BMI and gender are displayed: males (females) are represented by blue (red) lines. As you can see, there appears to be no relationship between the probability of chronic neck pain and BMI for both genders where females have a nearly parallel higher probability than males. Based on sampling weights (not shown), the results are basically the same.