# Log-Linear Bayesian Additive Regression Trees for Categorical and Count Responses

Jared S. Murray*
Department of Statistics, Carnegie Mellon University

January 2, 2017

**Abstract**

We introduce Bayesian additive regression trees (BART) for log-linear models including multinomial logistic regression and count regression with zero-inflation and overdispersion. BART has been applied to nonparametric mean regression and binary classification problems in a range of settings. However, existing applications of BART have been limited to models for Gaussian "data", either observed or latent. This is primarily because efficient MCMC algorithms are available for Gaussian likelihoods. But while many useful models are naturally cast in terms of latent Gaussian variables, many others are not – including models considered in this paper.

We develop new data augmentation strategies and carefully specified prior distributions for these new models. Like the original BART prior, the new prior distributions are carefully constructed and calibrated to be flexible while guarding against overfitting. Together the new priors and data augmentation schemes allow us to implement an efficient MCMC sampler outside the context of Gaussian models. The utility of these new methods is illustrated with several examples.

*Keywords:* Multinomial logistic regression, Poisson regression, Negative binomial regression, Zero inflation, Nonparametric Bayes

*The author gratefully acknowledges support from the National Science Foundation under grant numbers SES-1130706, SES-1631970 and DMS-1043903. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies. Thanks to P. Richard Hahn for helpful comments and suggestions on an early version of this manuscript.

# 1 Introduction

Since their introduction by Chipman et al. (2010) Bayesian additive regression trees (BART) have been applied to nonparametric regression and classification problems in a wide range of settings. To date these have been limited to models for Gaussian data, perhaps after data augmentation (as in probit BART for binary classification). Although many useful models are naturally cast in terms of latent Gaussian variables, many others are not or have other, more convenient latent variable representations. This paper extends BART to a much wider range of models via a novel log-linear formulation that is easily incorporated into regression models for categorical and count responses. Adapting BART to the log-linear setting while maintaining the computational efficiency of the original BART MCMC algorithm requires careful consideration of prior distributions, one of the main contributions of this paper.

The paper proceeds as follows: The remainder of this section reviews BART, including elements of the MCMC algorithm used for posterior inference. In Section 2 we introduce new log-linear BART models for categorical and count responses. In Section 3 we describe data augmentation and MCMC algorithms for these models. In Section 4 we introduce new prior distributions and give details of posterior computation. In Section 5 we present a large simulation study and an application to previously published data. In Section 6 we conclude with discussion of extensions and areas for future work.

## 1.1 Bayesian Additive Regression Trees (BART)

BART was introduced by Chipman et al. (2010) (henceforth CGM) as a nonparametric prior over a regression function $f(\cdot)$ designed to capture complex, nonlinear relationships and interactions. Our exposition in this section closely follows CGM. For observed data pairs $\{(y_i, \mathbf{x}_i); 1 \leq i \leq n\}$ CGM consider the regression model

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{1}$$

where $f$ is represented as the sum of many regression trees.

Each tree $T_h$ (for $1 \leq h \leq m$) consists of a set of interior decision nodes with splitting

rules of the form $x_{ij} < c$, and a set of $b_h$ terminal nodes. Each terminal node has an associated parameter, collected in the vector $M_h = (\mu_{h1}, \mu_{h2}, \ldots \mu_{hb_h})'$. We use $T = \{T_h : 1 \leq h \leq m\}$ and $M = \{M_h : 1 \leq h \leq m\}$ to refer to the collections of trees/parameters.

A tree and its associated decision rules induce a partition of the covariate space $\{\mathcal{A}_{h1}, \ldots, \mathcal{A}_{hb_h}\}$, where each element of the partition corresponds to a terminal node in the tree. Each pair $(T_h, M_h)$ parameterizes a step function $g$:

$$g(\mathbf{x}, T_h, M_h) = \mu_{ht} \text{ if } \mathbf{x} \in \mathcal{A}_{ht} \text{ (for } 1 \leq t \leq b_h). \tag{2}$$

An example tree and its corresponding step function are given in Figure 1. In BART a large number of these step functions are additively combined to obtain $f$:

$$f(\mathbf{x}) = \sum_{h=1}^{m} g(\mathbf{x}, T_h, M_h). \tag{3}$$

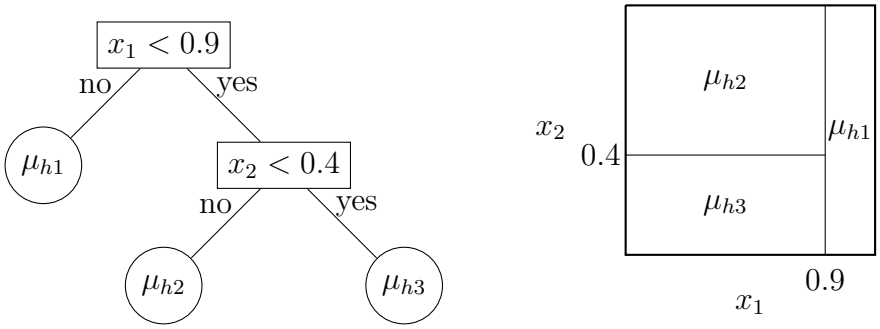


Figure 1: (Left) An example binary tree, with internal nodes labeled by their splitting rules and terminal nodes labeled with the corresponding parameters $\mu_{ht}$ (Right) The corresponding partition of the sample space and the step function $g(\mathbf{x}, T_h, M_h)$.

The prior on $(T_h, M_h)$ strongly favors small trees and leaf parameters that are near zero, constraining each term in the sum to be a "weak learner". Each tree is assigned an independent prior introduced by Chipman et al. (1998), where trees are grown iteratively: Starting from the root node, the probability that a node at depth $d$ splits (is not terminal) is given by

$$\alpha(1+d)^{-\beta}, \quad \alpha \in (0,1), \ \beta \in [0, \infty). \tag{4}$$

CGM propose $\alpha = 0.95$ and $\beta = 2$ as default values, which strongly favors small trees (of depth 2-3). A variable to split on is then selected uniformly at random, and given the selected variable a value to split at is selected according to a prior distribution defined over a grid. If the $j^{th}$ variable is continuous the grid for variable $j$ is either uniformly spaced or given by a collection of observed quantiles of $\{x_{ij} : 1 \leq i \leq n\}$. For binary or ordinal variables, the cutpoints can be defined as the collection of all possible values. Unordered categorical variables with $q$ levels are generally expanded as $q$ binary variables indicating each level, although alternative coding schemes could be used instead.

To set shrinkage priors on $M$ and avoid overfitting, CGM suggest scaling the data to lie in $\pm 0.5$ and assigning the leaf parameters independent priors:

$$\mu_{ht} \overset{iid}{\sim} N(0, \sigma_\mu^2) \quad \text{where } \sigma_\mu = 0.5/(k\sqrt{m}). \tag{5}$$

CGM recommend $1 \leq k \leq 3$, with $k = 2$ as a reasonable default choice. This prior shrinks the individual basis functions strongly toward zero and yields a $N(0, m\sigma_\mu^2)$ marginal prior for $f(\mathbf{x})$ at any covariate value. Since $\sqrt{m}\sigma_\mu = 0.5/k$ this prior assigns approximately 95% probability to the range of the transformed data ($\pm 0.5$) when $k = 2$, so $\sigma_\mu$ (through $k$) can be used to calibrate the prior.

## 1.2 MCMC for BART: "Bayesian backfitting"

A key ingredient in the MCMC sampler for BART is the "Bayesian backfitting" step, which we describe briefly here. (The Bayesian backfitting label is due to Hastie and Tibshirani (2000), who proposed a similar algorithm for MCMC sampling in additive models.) Let $T_{(h)} \equiv \{T_l : 1 \leq l \leq m,\ l \neq h\}$ denote all but the $h^{th}$ tree with $M_{(h)}$ defined similarly. CGM's MCMC algorithm updates $(T_h, M_h \mid T_{(h)}, M_{(h)}, -)$ in a block. This is simplified by the observation that

$$R_{hi} = \left( y_i - \sum_{l \neq h}^{m} g(\mathbf{x}_i, T_l, M_l) \right) \sim N(g(\mathbf{x}_i, T_h, M_h), \sigma^2), \tag{6}$$

so that $(T_h, M_h)$ only depends on the data through the vector of current partial residuals $R_h = (R_{h1}, R_{h2}, \ldots R_{hn})$. The partial residuals follow the Bayesian regression tree model described in Chipman et al. (1998), so the Metropolis-Hastings update given there can be can be adopted to sample from $(T_h, M_h \mid -)$ directly, treating $R_h$ as the observations.

Jointly updating the $(T_h, M_h)$ pairs obviates the need for transdimensional MCMC algorithms (to cope with the fact that the length of $M_h$ changes with the depth of $T_h$), which can be delicate to construct (Green, 1995). In addition, block updating parameters often accelerates the mixing of MCMC algorithms (Liu et al., 1994; Roberts and Sahu, 1997). The efficiency of this blocked MCMC sampler is a key feature of BART, and one of the contributions of this paper is to generalize this sampler to a wider range of models where backfitting is infeasible.

# 2 Log-linear BART Models

Extensions of the BART model in (1) have previously been limited to Gaussian models. CGM utilized BART for binary classification using a probit link and Albert and Chib (1993)'s data augmentation. Kindo et al. (2016) similarly extended BART to unordered categorical responses with latent Gaussian random variables in a multinomial probit regression model. Sparapani et al. (2016) use a clever reparameterization to adapt probit BART to survival analysis. The focus on Gaussian models seems to be motivated by the desire to utilize the Bayesian backfitting MCMC algorithm.

However, many models either lack a natural representation in terms of observed or latent Gaussian random variables or have a different, more convenient latent variable formulation. We consider several such models below. These models include one or more regression functions with positivity constraints. The natural extension of BART to this setting is obtained by expanding the log of the regression function into a sum of trees:

$$\log[f(\mathbf{x})] = \sum_{h=1}^{m} g(\mathbf{x}, T_h, M_h), \tag{7}$$

yielding log-linear Bayesian additive regression trees. We introduce log-linear BART models for categorical and count responses in the following subsections.

## 2.1 Multinomial logistic regression models

Suppose that for each covariate value $\mathbf{x}_i$ we observe $n_i$ observations falling into one of $1 \leq j \leq c$ categories. Often $n_i = 1$ for all $i$, as in the case with continuous covariates. Let $y_{ij}$

be the number of observations with covariate value $\mathbf{x}_i$ in category $j$ (so that $\sum_{j=1}^{c} y_{ij} = n_i$). We assume that the probability of observing category $j$ at a given covariate level is

$$\pi_j(\mathbf{x}_i) = \frac{f^{(j)}(\mathbf{x}_i)}{\sum_{l=1}^{c} f^{(l)}(\mathbf{x}_i)}, \tag{8}$$

or equivalently that the log odds in favor of category $j'$ over $j$ are given by

$$\log[f^{(j')}(\mathbf{x}_i)] - \log[f^{(j)}(\mathbf{x}_i)] \tag{9}$$

for any $j \neq j'$. Assuming $\log[f^{(j)}(\mathbf{x}_i)] = \sum_{h=1}^{m} g(\mathbf{x}, T^{(j)}, M^{(j)})$ induces a log-linear form for each of the log odds functions in (9), defining a multinomial logistic BART model:

$$\pi_j(\mathbf{x}_i) = \frac{\exp\left[\sum_{h=1}^{m} g(\mathbf{x}, T^{(j)}, M^{(j)})\right]}{\sum_{l=1}^{c} \exp\left[\sum_{h=1}^{m} g(\mathbf{x}, T^{(l)}, M^{(l)})\right]}. \tag{10}$$

As written this model is unidentified. Identification could be obtained by fixing some $f^{(l)}(\cdot) := 1$, in which case $f^{(l)}(\mathbf{x})$ gives the odds of category $l$ against category $j$ at covariate value $\mathbf{x}$. However, this prior depends on the arbitrary choice of a reference category. Instead, we use proper priors for each $f^{(j)}$ and work in the unidentified space. This avoids asymmetries in the prior arising from the arbitrary choice of the reference category, and has some computational benefits as well (see Section A.1 in the supplemental material). Post-processing MCMC samples yields estimates of identified quantities like predicted probabilities or odds ratios.

## 2.2 Count regression models, with overdispersion and zero-inflation

For count responses we begin with Poisson or negative binomial models with mean function $\mathrm{E}(y_i \mid \mathbf{x}_i) = \mu_{0i} f(\mathbf{x}_i)$. Here $\mu_{0i}$ is a fixed offset such as an adjustment for unit-level exposure, or we may take $\mu_{i0} \equiv \mu_0$ to center the prior for the regression function at $\mu_0$. We induce a log-linear model for the mean function by assuming

$$\log[f(\mathbf{x})] = \sum_{h=1}^{m} g(\mathbf{x}, T_h, M_h). \tag{11}$$

The Poisson model is completely specified by the mean function. The negative binomial regression model has an additional parameter $\kappa$, which controls the degree of overdispersion relative to the Poisson. Under the negative binomial model,

$$\mathrm{Var}(y_i \mid \mathbf{x}_i) = \mathrm{E}(y_i \mid \mathbf{x}_i)\left(1 + \frac{\mathrm{E}(y_i \mid \mathbf{x}_i)}{\kappa}\right). \tag{12}$$

As $\kappa \to \infty$, the negative binomial model converges to the Poisson. The probability mass function under the Poisson model is

$$p_P(y_i \mid \mathbf{x}_i, f) = \frac{\exp[-\mu_{0i} f(\mathbf{x}_i)][\mu_{0i} f(\mathbf{x}_i)]^{y_i}}{y_i!}. \tag{13}$$

For the negative binomial model we have

$$p_{NB}(y_i \mid \mathbf{x}_i, f, \kappa) = \frac{\Gamma(\kappa + y_i)}{\Gamma(\kappa) y_i!} \left(\frac{\kappa}{\kappa + \mu_{0i} f(\mathbf{x}_i)}\right)^{\kappa} \left(\frac{\mu_{0i} f(\mathbf{x}_i)}{\kappa + \mu_{0i} f(\mathbf{x}_i)}\right)^{y_i}. \tag{14}$$

Many datasets exhibit an excess of zero values. Zero inflated variants of Poisson or negative binomial regression models accommodate the extra zeros by adding a point mass component:

$$\Pr(Y_i = y \mid \mathbf{x}_i) = \begin{cases} (1 - \omega(\mathbf{x}_i)) + \omega(\mathbf{x}_i) p(y \mid \mathbf{x}_i, f, \kappa) & \text{if } y = 0 \\ \omega(\mathbf{x}_i) p(y \mid \mathbf{x}_i, f, \kappa) & \text{if } y > 0 \end{cases} \tag{15}$$

where $p(y \mid \mathbf{x}_i, f, \kappa)$ is the probability mass function of a Poisson or negative binomial with mean $\mu_{0i} f(\mathbf{x})$ and dispersion $\kappa$ and $1 - \omega(\mathbf{x}_i)$ is the probability that a zero is due to the point mass component. We assume that

$$\text{logit}[1 - \omega(\mathbf{x})] = \log[1 - \omega(\mathbf{x})] - \log[\omega(\mathbf{x})] \tag{16}$$

has a log-linear expansion, which will be induced through the redundant parameterization

$$\omega(\mathbf{x}_i) = \frac{f^{(1)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(0)}(\mathbf{x}_i)}, \tag{17}$$

where $f^{(0)}$ and $f^{(1)}$ have independent log-linear BART priors as in the multinomial logistic regression model.

# 3 MCMC and Data Augmentation for Log-linear BART

Fitting the models in Section 2 is nontrivial: Some of the models lack a Gaussian representation, so CGM's Bayesian backfitting approach does not apply directly. However, the key element in CGM's MCMC sampler is actually a blocked MCMC update for each tree and its parameters, holding the other trees and parameters fixed. CGM derive this update

via Bayesian backfitting, but this is not strictly necessary. The general form of the update is summarized in Algorithm 1, using the following notation for the log-linear case:

We have one or more functions that have a sum-of-trees representation on the log scale, so that $\log[f(\mathbf{x})] = \sum_{h=1}^{m} g(\mathbf{x}, T_h, M_h)$. It will be convenient to work with $f$ directly, so we define the following transformed parameters:

$$\lambda_{ht} = \exp(\mu_{ht}), \quad \Lambda_h = (\lambda_{h1}, \lambda_{h2}, \ldots \lambda_{hb_h})', \tag{18}$$

and note that $g(\mathbf{x}, T_h, \Lambda_h) = \exp[g(\mathbf{x}, T_h, M_h)] = \lambda_{ht}$ if $\mathbf{x} \in \mathcal{A}_{ht}$ (for $1 \leq t \leq b_h$), so

$$f(\mathbf{x}) = \exp\left[\sum_{h=1}^{m} g(\mathbf{x}, T_h, M_h)\right] = \prod_{h=1}^{m} g(\mathbf{x}, T_h, \Lambda_h). \tag{19}$$

Additional parameters or latent variables are collected in a vector $\theta$. In models with more than one regression function we consider MCMC updates for each regression function conditional on the others, which we also collect in $\theta$.

---

**Algorithm 1** One step of the MCMC algorithm for updating a log-linear BART function parameterized by $T = \{T_h\}$ and $\Lambda = \{\Lambda_h\}$ $(1 \leq h \leq m)$

---

**Input:** Data and current values for $T$, $\Lambda$, and other parameters/latent variables (in $\theta$)

**Output:** New values of $T$, $\Lambda$

**for** $1 \leq h \leq m$ **do**

1. Propose $T_h^* \sim q(T_h^*;\ T_h)$
2. Set $a \leftarrow \frac{L(T_h^*;\ T_{(h)}, \Lambda_{(h)}, \theta, y) p(T_h^*)}{L(T_h;\ T_{(h)}, \Lambda_{(h)}, \theta, y) p(T_h)} \frac{q(T_h;\ T_h^*)}{q(T_h^*;\ T_h)}$
3. Set $T_h \leftarrow T_h^*$ with probability $\min(1, a)$
4. Sample $\Lambda_h \sim p(\Lambda_h \mid T_h, -)$

**end for**

---

Computing the conditional integrated likelihood function

$$L(T_h;\ T_{(h)}, \Lambda_{(h)}, \theta, y) = \int L(T_h, \Lambda_h;\ T_{(h)}, \Lambda_{(h)}, \theta, y) p(\Lambda_h) d\Lambda_h \tag{20}$$

is a key step in Algorithm 1. This is trivial in Gaussian BART models because CGM's normal prior is conjugate to the distribution of the observed or latent data. Efficiently computing this integral under CGM's original prior in log-linear BART models is not as

simple, since the prior is no longer conjugate. In particular, we will be concerned with likelihoods of the form

$$L(T, \Lambda; \Theta, y) = \prod_{i=1}^{n} w_i f(\mathbf{x}_i)^{u_i} \exp[v_i f(\mathbf{x}_i)] \tag{21}$$

where $w_i$, $u_i$,and $v_i$ are some functions of $\theta$ and $y_i$ that will vary depending on the model under consideration. To derive the corresponding conditional likelihood for $(T_h, \Lambda_h)$, define $f_{(h)}(\mathbf{x}) = \prod_{l \neq h} g(\mathbf{x}, T_h, \Lambda_h)$. This is the fit from all but the $h^{th}$ tree, and does not vary with $(T_h, \Lambda_h)$. Then we have

$$\begin{aligned}
L((T_h, \Lambda_h); T_{(h)}, \Lambda_{(h)}, y) &= \prod_{i=1}^{n} w_i f(\mathbf{x}_i)^{u_i} \exp[v_i f(\mathbf{x}_i)] && (22)\\
&= \prod_{i=1}^{n} w_i [f_{(h)}(\mathbf{x}_i) g(\mathbf{x}_i, T_h, \Lambda_h)]^{u_i} \exp[v_i f_{(h)}(\mathbf{x}_i) g(\mathbf{x}_i, T_h, \Lambda_h)] && (23)\\
&= \prod_{t=1}^{b_h} \prod_{i:\mathbf{x}_i \in \mathcal{A}_{ht}} w_i [f_{(h)}(\mathbf{x}_i) \lambda_{ht}]^{u_i} \exp[v_i f_{(h)}(\mathbf{x}_i) \lambda_{ht}] && (24)\\
&= c_h \prod_{t=1}^{b_h} \lambda_{ht}^{r_{ht}} \exp\left[-s_{ht} \lambda_{ht}\right], && (25)
\end{aligned}$$

where the outer product in (24) runs over the end nodes of $T_h$ and the inner product is over the observations with covariate values in the corresponding element of the partition (as defined in (2)), and

$$c_h = \prod_{i=1}^{n} w_i f_{(h)}(\mathbf{x}_i)^{v_i}, \quad r_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}} u_i, \quad s_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}} f_{(h)}(\mathbf{x}_i) v_i, \tag{26}$$

with $r_{ht}$ and $s_{ht}$ playing the role of conditional "sufficient" statistics.

To implement Algorithm 1, we need to compute the conditional integrated likelihood

$$L(T_h;\ T_{(h)}, \Lambda_{(h)}, y) = \int c_h \prod_{t=1}^{b_h} \lambda_{ht}^{r_{ht}} \exp\left[-\lambda_{ht} s_{ht}\right] p(\Lambda_h) d\Lambda_h \tag{27}$$

in step 2. The original BART prior for $M_h$ induces independent lognormal priors for $\lambda_{ht}$, and the integral (27) is unavailable under this prior. Before introducing a new conjugate prior in Section 4, we show how all the models in Section 2 admit simple data augmentation schemes that result in likelihood functions with multiple factors of the form (21). This will allow us to use one algorithm to fit all the models in Section 2.

## 3.1 Data Augmentation for Multinomial Logistic Models

The likelihood contribution for each distinct covariate value is

$$p_{MN}(y_i) = \binom{n_i}{y_{i1}y_{i2}\dots y_{ic}} \frac{\prod_{j=1}^{c} f^{(j)}(\mathbf{x}_i)^{y_{ij}}}{(\sum_{l=1}^{c} f^{(l)}(\mathbf{x}_i))^{n_i}}. \tag{28}$$

We augment the likelihood function by introducing a new latent variable $\phi_i$, and defining a joint model for $(\phi_i, y_i)$ where the marginal probability mass function of $y_i$ is (28) and $(\phi_i \mid y_i, -) \sim Gamma(n_i, \sum_{j=1}^{c} f^{(j)}(\mathbf{x}_i))$ (recall that $n_i = \sum_{j=1}^{c} y_{ij}$). This yields the following augmented likelihood:

$$p_{MN}(y_i, \phi_i) = \binom{n_i}{y_{i1}y_{i2}\dots y_{ic}} \left( \prod_{j=1}^{c} f^{(j)}(\mathbf{x}_i)^{y_{ij}} \right) \frac{\phi_i^{n_i-1}}{\Gamma(n_i)} \exp\left[ -\phi_i \sum_{j=1}^{c} f^{(j)}(\mathbf{x}_i) \right] \tag{29}$$

$$= \binom{n_i}{y_{i1}y_{i2}\dots y_{ic}} \frac{\phi_i^{n_i-1}}{\Gamma(n_i)} \prod_{j=1}^{c} f^{(j)}(\mathbf{x}_i)^{y_{ij}} \exp\left[ -\phi_i f^{(j)}(\mathbf{x}_i) \right]. \tag{30}$$

Note that given $\phi_i$ the augmented model (30) factors into separate terms for each $f^{(j)}(\cdot)$, with each taking the form of (21).

The "gamma trick" as a tool for dealing with sums or integrals in the denominator has appeared in other settings as well (e.g. (Nieto-Barajas et al., 2004; Walker, 2011; Caron and Doucet, 2012)). The same likelihood (up to an irrelevant constant) can also be derived via the Poisson-multinomial transformation (Baker, 1994; Forster, 2010), which adds an artificial Poisson distribution for the cell total $n_i$ parameterized by $\phi_i$ and $\sum_{j=1}^{c} f^{(j)}(\mathbf{x}_i)$ (with a further prior on $\phi_i$, $p(\phi_i) \propto \phi_i^{-1}$). Since $n_i$ is often fixed by design, in our view casting the augmented model directly in terms of a proper joint probability model for $(y_i, \phi_i)$ is more transparent and removes any questions about the propriety of the posterior.

Our data augmentation has some advantages over alternatives for logistic models: There is a single latent variable with a simple distribution for each distinct covariate value (not necessarily each observation). Additionally, the functions $f^{(j)}$ are conditionally independent given $\phi$ allowing for parallel updates to speed up the most computationally intensive step during MCMC. No other augmentation for logistic models has all these features. In addition to proposing the current state-of-the-art Polya-Gamma data augmentation for logistic likelihoods, Polson et al. (2013) give a recent review and comparison of several choices (including e.g. Holmes and Held (2006); Frühwirth-Schnatter and Frühwirth (2010)). While

these augmentations yield Gaussian models, they either require multiple latent variables per observation or latent variables with non-standard distributions. None yield conditional independence of the $f^{(j)}$'s.

In related work Kindo et al. (2016) proposed a multinomial probit BART model using Albert and Chib (1993)'s data augmentation, which requires sampling from a truncated multivariate normal latent variable for each observation. It also requires the specification of a reference category and a prior for the covariance matrix over the latent Gaussian random variables, neither of which is easy or inconsequential (see Burgette and Hahn (2010) for discussion about reference categories, and Burgette and Nordheim (2012) on covariance matrix priors in linear regression settings). It also does not result in conditional independence of the $f^{(j)}$'s.

## 3.2 Data Augmentation for Count Models

The Poisson model requires no data augmentation. The negative binomial and zero-inflated Poisson data augmentation schemes can be obtained via restrictions of the data augmentation for the zero-inflated negative binomial (ZINB) model, which we describe below. The likelihood contribution of a single observation under the ZINB model is

$$
\begin{aligned}
p_{ZINB}(y_i \mid \mathbf{x}_i, f, f^{(0)}, f^{(1)}, \kappa) =& \frac{f^{(1)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} p_{NB}(y_i \mid \mathbf{x}_i, f, \kappa) && (31)\\
&+ \left( \frac{f^{(0)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} \right) \mathbb{1}(y_i = 0) && (32)
\end{aligned}
$$

Introducing $\xi_i, \phi_i \in (0, \infty)$ and $Z_i \in \{0, 1\}$ we can define the data augmented likelihood:

$$
\begin{aligned}
p_{ZINB}(y_i, Z_i, \phi_i, \xi_i \mid f^{(0)}, f_i, \kappa, f) =& f^{(0)}(\mathbf{x}_i)^{1-Z_i} \exp[-\phi_i f^{(0)}(\mathbf{x}_i)] && (33)\\
&\times f^{(1)}(\mathbf{x}_i)^{Z_i} \exp[-\phi_i f^{(1)}(\mathbf{x}_i)] && (34)\\
&\times f(\mathbf{x}_i)^{Z_i y_i} \exp\left[-Z_i \xi_i \mu_{0i} f(\mathbf{x}_i)\right] && (35)\\
&\times \left\{ \frac{1}{\Gamma(\kappa) y_i!} \kappa^{\kappa} \mu_{0i}^{y_i} \xi_i^{\kappa + y_i - 1} \exp\left[-\xi_i \kappa\right] \right\}^{Z_i} && (36)\\
&\times \mathbb{1}(Z_i = 1 \text{ if } y_i > 0). && (37)
\end{aligned}
$$

**Proposition 3.1.** Integrating over $\xi_i, \phi_i$, and $Z_i$ in (33)-(37) yields (32).

Note that given values for all the latent variables, the likelihood factors into terms of the form (21) for each of the log-linear functions (Eq. (33)-(35)). The augmented likelihood function for the negative binomial model *without* zero-inflation is obtained by fixing $Z_i = 1$ for all $i$ and removing terms (33) and (34). An augmented likelihood for the zero-inflated Poisson model is recovered by setting $\xi_i = 1$ for all $i$ and dropping the remaining terms involving $\kappa$. Applying both restrictions leads to the Poisson likelihood function.

# 4 Prior choice and posterior computation

Given the conditional likelihood

$$L((T_h, \Lambda_h); T_{(h)}, \Lambda_{(h)}, \Theta, y) = c_h \prod_{t=1}^{b_h} \lambda_{ht}^{r_{ht}} \exp\left[-\lambda_{ht} s_{ht}\right], \tag{38}$$

from the previous section we would prefer a prior for $\lambda_{ht}$ that is

1. Symmetric about 0 on the log scale, since

$$\log[f(\mathbf{x})] = \sum_{h=1}^{m} \log[g(\mathbf{x}, T_h, \Lambda_h)] = \sum_{h=1}^{m} \sum_{t=1}^{b_h} \log(\lambda_{ht}) \mathbb{1}(\mathbf{x} \in \mathcal{A}_{ht}). \tag{39}$$

   Each term in the sum should contribute a small amount to the overall fit, in either direction with equal prior probability, in the same spirit as the original CGM prior.

2. Conjugate to (38), so we can compute the integrated likelihood (27) in closed form and easily sample the end node parameters from their full conditional $p(\Lambda_h \mid T_h, -)$.

Independent lognormal priors on $\lambda_{ht}$ satisfy 1, but not 2. Independent Gamma priors satisfy 2, but not 1 - they are asymmetric on the log scale. Exact symmetry and conditional conjugacy requires a new prior, which we introduce below.

## 4.1 A symmetric, conditionally conjugate prior

Our strategy for deriving the new prior on $\lambda_{ht}$ is to ensure that in addition to symmetry and conjugacy, we have $\log[f(\mathbf{x})] \overset{approx}{\sim} N(0, a_0^2)$ marginally at any covariate value $\mathbf{x}$. This allows

us to use $a_0$ to calibrate the log-linear prior the same way that $\sigma_\mu$ parameter calibrates the original CGM prior. (Nonzero means for the log-linear regression function are handled via multiplicative offsets.) So with independent priors for $\lambda_{ht}$, we require that $\mathrm{E}(\log[\lambda_{ht}]) = 0$ and $\mathrm{Var}(\log[\lambda_{ht}]) = a_0^2/m$. Typically $m$ is large, so the normal approximation to the marginal distribution of $\log[f(\mathbf{x})]$ will be accurate by the central limit theorem. The specific prior below is somewhat complex, but the end result is very similar to CGM's leaf prior and has a single, interpretable tuning parameter (for a fixed $m$).

Our proposed leaf prior is a mixture of generalized inverse Gaussian (GIG) distributions. GIG distributions are characterized by their density function

$$p_{GIG}(\lambda \mid \eta, \chi, \psi) = \frac{\lambda^{\eta-1} \exp\left[-\frac{1}{2}\left(\chi/\lambda + \psi\lambda\right)\right]}{Z(\eta, \chi, \psi)}, \tag{40}$$

with normalizing constant

$$Z(\eta, \chi, \psi) = \begin{cases} \Gamma(\eta)\left(\frac{2}{\psi}\right)^{\eta} & \text{if } \eta > 0,\ \chi = 0,\ \psi > 0 \\ \Gamma(-\eta)\left(\frac{2}{\chi}\right)^{-\eta} & \text{if } \eta < 0,\ \chi > 0,\ \psi = 0 \\ \frac{2K_\eta(\sqrt{\psi\chi})}{(\psi/\chi)^{(\eta/2)}} & \text{if } \chi > 0,\ \psi > 0, \end{cases} \tag{41}$$

where $K_\eta(x)$ is the modified Bessel function of the second kind. The gamma and inverse gamma distributions are recovered when $\chi = 0$ and $\psi = 0$, respectively. This distribution is also conjugate to (38). Our mixture prior is given by

$$p_\lambda(\lambda_{ht} \mid c, d) = \frac{1}{2} p_{GIG}(\lambda_{ht} \mid -c, 2d, 0) + \frac{1}{2} p_{GIG}(\lambda_{ht} \mid c, 0, 2d). \tag{42}$$

where $c$ and $d$ are parameters that will be determined by $a_0$. As a mixture of GIG distributions this prior is also conjugate to (38). We refer to this as the $P_\lambda(c, d)$ distribution.

The $P_\lambda(c, d)$ distribution has the following simple stochastic representation:

$$W_{ht} \sim Gamma(c, d) \tag{43}$$

$$U_{ht} \sim Bernoulli(1/2) \tag{44}$$

$$\lambda_{ht} = U_{ht} W_{ht} + (1 - U_{ht})(1/W_{ht}), \tag{45}$$

(The $W$ and $U$ random variables are never instantiated and only introduced here for exposition.) By construction the implied prior on $\mu_{ht}$ is symmetric about 0 since $\mu_{ht} = \log(W_{ht})$ or $-\log(W_{ht})$ with equal probability.

The parameters $c, d$ can be set from user-supplied values of $a_0$ and $m$. The optimal values are not available in closed form (although they are easy to obtain numerically) but for a large number of trees and/or a small value of $a_0$, the values of $c$, $d$ also have simple approximate values. These results are summarized in Propositions 4.1 and 4.2.

**Proposition 4.1.** If $\lambda \sim P_\lambda(c, d)$, $\text{Var}(\lambda) = a_0^2/m$ when $\psi''(c) = a_0^2/m$ and $d = \exp(\psi'(c))$, where $\psi(c) = \log[\Gamma(c)]$. The function $\psi''(c)$ is monotone decreasing and hence invertible on $\mathbb{R}^+$, so the solutions to these equations are unique.

**Proposition 4.2.** For small values of $a_0^2/m$, the values of $c$ and $d$ from Proposition 4.1 are approximately $c \approx m/a_0^2 + 0.5$ and $d \approx m/a_0^2$.

One could calibrate a gamma prior similarly, and in fact the shape and rate parameters will be the same as $c$ and $d$ in Proposition 4.1 (respectively). Figure 2 compares the calibrated $P_\lambda$ and log-gamma priors to CGM's normal priors for $m = 25$ and $a_0 = 3.5/\sqrt{2}$, which are actual parameter settings we will use later. The log-gamma prior is asymmetric, compared to the log-$P_\lambda$ prior which is symmetric and has slightly heavier tails than the normal. The log-gamma and log-$P_\lambda$ priors both become increasingly close to the normal distribution as $a_0^2/m \to 0$, but the asymmetry in the log-gamma prior for small values of $m$ is undesirable. The $P_\lambda$ prior is a more reasonable default choice for the entire range of $a_0$ and $m$ values.

## 4.2 Posterior computation

With the prior specified we can now fill in the details of Algorithm 1:

1-3. We utilize the grow, prune, change and swap proposal moves described by CGM (originally introduced in Chipman et al. (1998)) but any proposals could be used (see e.g. Denison et al. (1998); Wu et al. (2007); Pratola (2016) for other possibilities). The integrated likelihood function that appears in the acceptance ratio is

$$L(T_h; T_{(h)}, \Lambda_{(h)}, \Theta, y) = c_h \prod_{t=1}^{b_h} \int \lambda_{ht}^{r_{ht}} \exp\left[-\lambda_{ht} s_{ht}\right] p_\lambda(\lambda_{ht} \mid c, d) d\lambda_{ht} \tag{46}$$

$$= c_h \prod_{t=1}^{b_h} \frac{Z(-c + r_{ht}, 2d, 2s_{ht}) + Z(c + r_{ht}, 0, 2[d + s_{ht}])}{2Z(c, 0, 2d)} \tag{47}$$

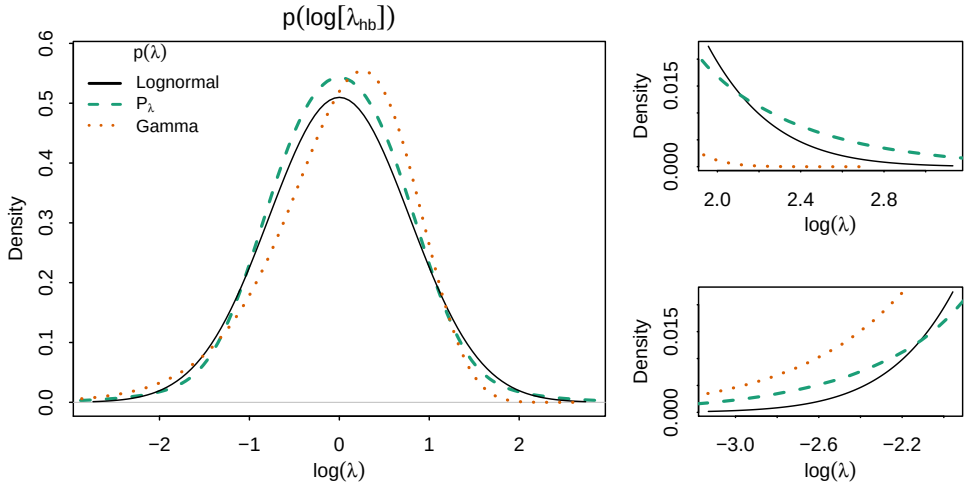


Figure 2: The proposed $P_\lambda$ node-parameter prior (green dashed line) compared to CGM's log-normal prior on $\lambda_{ht}$ (solid black) and a Gamma prior calibrated to have the same moments on the log scale (dashed orange). Here $m = 25$ and $a_0 = 3.5/\sqrt{2}$

using the fact that $Z(c, 0, 2d) = Z(-c, 2d, 0)$. The leading term $c_h$ cancels in the Metropolis-Hastings acceptance ratio, but the denominator in (47) does not when the proposal changes the dimension of the partition (e.g. grow/prune moves).

4. Sample $(\Lambda_h \mid T_h, T_{(h)}, \Lambda_{(h)})$ from its full conditional. The components of $\Lambda_h$ are conditionally independent with full conditional distributions

$$p(\lambda_{ht} \mid -) \propto \lambda_{ht}^{(-c+r_{ht})} \exp\left[-\frac{1}{2}\left(2d/\lambda_{ht} + 2s_{ht}\lambda_{ht}\right)\right] \quad + \lambda_{ht}^{(c+r_{ht})} \exp\left[-\frac{1}{2}\left(2d + 2s_{ht}\right)\lambda_{ht}\right]. \tag{48}$$

This distribution is a mixture of GIG distributions:

$$p(\lambda_{ht} \mid -) = \pi_{ht} p_{GIG}(-c + r_{ht}, 2d, 2s_{ht}) + (1 - \pi_{ht}) p_{GIG}(c + r_{ht}, 0, 2[d + s_{ht}]) \tag{49}$$

where

$$\pi_{ht} = \frac{Z(-c + r_{ht}, 2d, 2s_{ht})}{Z(-c + r_{ht}, 2d, 2s_{ht}) + Z(c + r_{ht}, 0, 2[d + s_{ht}])}. \tag{50}$$

Algorithm 1 forms the backbone of MCMC in log-linear BART models, with additional parameters or latent variables sampled from their conditional distributions in further MCMC steps. In the following subsections we describe how to calibrate the $P_\lambda$ prior for the models in Section 2 and outline posterior sampling.

## 4.3 Prior choice and posterior computation for multinomial logistic models

In the multinomial logistic BART model, for any two outcome categories $j \neq j'$ the log odds in favor of $j'$ are given by

$$\log[f^{(j')}(\mathbf{x}_i)] - \log[f^{(j)}(\mathbf{x}_i)], \tag{51}$$

and each function $f^{(l)}(\cdot)$ has an independent log-linear BART prior parameterized by $(T^{(l)}, \Lambda^{(l)})$ (for $1 \leq l \leq c$). We assume that the prior on each $f^{(l)}(\cdot)$ uses the same number of trees $m$ and parameter $a_0$ in the $P_\lambda$ prior. Then the induced prior on (51) is approximately $N(0, 2a_0^2)$, so $a_0$ can be chosen to reflect prior beliefs about the plausible range of the log odds functions. Since the log-odds lie within $(-2\sqrt{2}a_0, 2\sqrt{2}a_0)$ at any covariate value with probability approximately 0.95 under the prior, $a_0 = 3.5/\sqrt{2}$ is a reasonable default choice.

A single step of the MCMC sampler proceeds as follows:

1. For $1 \leq i \leq n$, draw $\phi_i \sim Gamma(n_i, \sum_{j=1}^{c} f^{(j)}(\mathbf{x}_i))$. This is a direct consequence of the data augmentation, which was conditional on $y_i$ and the regression functions.

2. For $1 \leq j \leq c$, *independently* update the parameters of $f^{(j)}$ using Algorithm 1 and the expressions in Section 4.2 with

$$r_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}^{(j)}} y_{ij}, \quad s_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}^{(j)}} \phi_i f_{(h)}^{(j)}(\mathbf{x}_i)$$

   where $f_{(h)}^{(j)}(\mathbf{x}_i) = \prod_{l \neq h} g(\mathbf{x}, T_h^{(j)}, \Lambda_h^{(j)})$ is the fit from all but the $h^{th}$ tree.

The augmentation in (30) yields a very convenient MCMC algorithm: There is a *single* augmented variable for each covariate value, regardless of the number of categories or observations, and it has a standard, untruncated distribution. Further, the $c$ regression functions are conditionally independent given the latent variable.

## 4.4 Prior choice and posterior computation for count models

We describe prior specification and MCMC sampling for the most complex case, the zero-inflated negative binomial. Prior specification is similar in negative binomial or zero-inflated

Poisson models. Specializations of the MCMC algorithm to the negative binomial or zero-inflated Poisson follow from the discussion at the end of Section 3.2.

Recall that the probability of observing an "excess" zero is

$$1 - \omega(\mathbf{x}_i) = \frac{f^{(0)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)}. \tag{52}$$

Similar to the previous subsection, independent log-linear BART priors on $f^{(0)}$ and $f^{(1)}$ with common values of the concentration parameter and number of trees (say $a_{z0}$ and $m_z$) induce a log-linear BART logistic regression model:

$$\text{logit}[1 - \omega(\mathbf{x})] = \log[f^{(0)}(\mathbf{x})] - \log[f^{(1)}(\mathbf{x})] \tag{53}$$

The log-odds of observing an excess zero at any covariate value (53) is approximately distributed $N(0, 2a_{z0}^2)$ marginally, so $a_{z0}$ may be chosen based on plausible values for the odds function. As defaults we suggest $m_z = 100$ and $a_0 = 3.5/\sqrt{2}$.

In the zero-inflated model, $\mu_{0i} f(\mathbf{x}_i)$ is the mean of the non-point mass component of the zero-inflated model and $f(\cdot)$ has a log-linear BART prior with $m$ trees and concentration parameter $a_0$. Assuming $\mu_{0i} = \mu_0$, a reasonable default prior is obtained by positing a near-maximum value for $y$, say $y^*$, and setting $a_0 = 0.5[\log(y^*) - \mu_0]$. Then $\Pr(f(\mathbf{x}_i) \leq y^*) \approx 0.95$ marginally, since $\log[f(\mathbf{x}_i)] \overset{approx}{\sim} N(0, a_0^2)$. For large values of $\mu_0$ it may also be necessary to specify a near-minimum as well. For $\kappa$, we use beta prime priors: $p(\kappa) \propto \kappa^{a_\kappa - 1}(1+\kappa)^{-a_\kappa + b_\kappa}$. This is a heavy-tailed prior which is equivalent to a $Beta(a_\kappa, b_\kappa)$ prior on $\kappa/(1+\kappa)$. Gamma priors are another reasonable choice (e.g. Zhou et al. (2012)).

Posterior sampling for the ZINB model has many more steps than the multinomial logistic regression model, and is outlined in Section A.3 of the supplemental material. The primary innovation is three applications of Algorithm 1 that can be run in parallel, with all the remaining parameters updated in a single block for efficiency.

# 5 Illustrations and applications

## 5.1 Simulation: Multinomial Logistic Regression

We compared default and cross-validated multinomial logistic BART models (BART-default and BART-CV, respectively) with several classification methods using 20 datasets taken

from the UCI repository and processed as in Fernández-Delgado et al. (2014). The primary purpose of this exercise is to establish multinomial logistic BART as having reasonable classification performance. We do not expect BART to necessarily outperform other machine learning methods designed and tuned for classification accuracy. We would also like to compare the performance of default and CV BART models, as default variants require less computation and yield valid posterior inference.

For our comparison we selected datasets with 3-6 outcome categories and between 100 and 3,000 observations. We consider two variants of BART-default: one that sets the number of trees per category to 100, so that the log-odds functions involve 200 trees, and one that sets the number of trees per category such that the *total* number of trees is as close to 200 as possible. Both set $a_0 = 3.5/\sqrt{2}$. BART-CV was cross validated over range of $m$ that included both default rules for the number of trees and 25 trees per category. Possible values for $a_0$ included $2/\sqrt{2}, 3.5/\sqrt{2}$ (the default choice) and $6/\sqrt{2}$. Competing methods included random forests, gradient boosted models, penalized multinomial probit regression, a support vector machine using radial basis functions, and a single layer neural net[1]. Each method was cross-validated over its default parameter grid in the R package `caret`. Classifiers were compared on the basis of a 10-fold CV estimate of classification accuracy.

Table 3 gives CV estimates of out of sample classification accuracy, along with its standard deviation. BART-CV is generally competitive and has the best accuracy in 5 of the 20 examples, although the differences in accuracy are typically small relative to the standard deviation. To get a better sense of the variability across splits, Figure 3 shows the relative out-of-sample accuracy of each method against the best performer across the 10 folds for each of the 20 datasets. Once again, for most problems no single method clearly dominates. BART-CV is not only competitive, but its out of sample accuracy tends to be

[1]We tried to include Kindo et al. (2016)'s multinomial probit BART, but the accompanying R package routinely crashed during simulations. We expect that it would perform similar to multinomial logistic BART in cross-validation, at substantially increased computational cost due to the need to update several latent Gaussian variables per covariate value as well as a latent covariance matrix, and to cross-validate the choice of reference category in addition to $m$ and the parameters of the covariance matrix prior. (Kindo et al. (2016) propose no default settings for reference category or prior on the covariance matrix.)

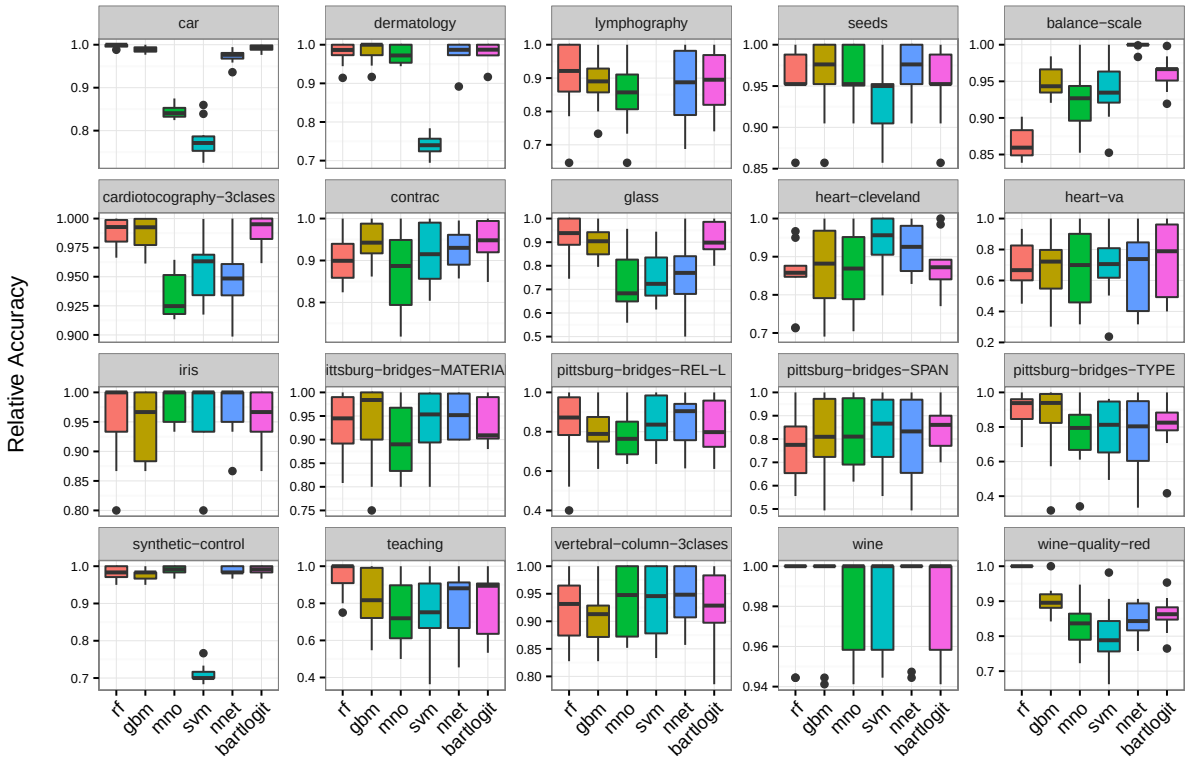


Figure 3: Relative accuracy over the 10 folds for each dataset in the classification simulation.

stable across the different splits.

The cross-validated BART models were generally not default choices: Two out of twenty datasets yielded a cross-validated model that was also one of the defaults. In ten cases cross-validation selected a smaller set of trees (25 per level). In four cases cross-validation yielded lower values of $a_0$ $(2/\sqrt{2})$, and in twelve cases larger values of $a_0$ $(6/\sqrt{2})$ were selected. There was no clear relationship between the cross-validated parameters and the number of outcome categories, covariates, or the difficulty of the problem. However, the differences in classification accuracy where generally mild: Figure 4 compares the relative accuracy of the two default BART prior settings against BART-CV. The default choices were nearly as accurate as the CV model at a fraction of the computational cost, and also maintain their Bayesian validity. The default prior with 100 trees per category tended to have a slight edge over the prior with 200 total trees, at some computational cost.

In summary, default versions of BART have competitive predictive performance. More

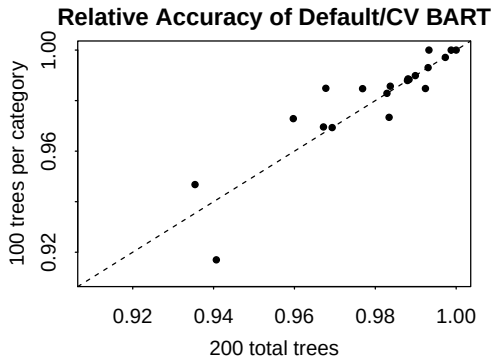


Figure 4: Relative accuracy of the two BART-default variants versus BART-CV.

importantly, these are proper, fully Bayesian models that give valid posterior inference and may be incorporated into more complex models where cross-validation would be difficult even if it were desirable. An immediate example of this is the logistic BART model incorporated into the zero-inflated count regression model.

## 5.2 Example: Patent Citations

When applying for new patents inventors must cite related existing patents, so the number of citations a patent receives is a (crude) measure of the invention's influence. We consider predicting citation counts using data from the European Patent Office (EPO), as presented in Klein et al. (2015). Several covariates are available; these are summarized in Table 1. Klein et al. (2015) provide compelling evidence that these data cannot be adequately modeled without zero inflation and overdispersion, so we compare the ZINB-BART regression model to the semiparametric Bayesian ZINB regression models introduced in that paper.

Klein et al. (2015) select a model based on stepwise selection using DIC under semiparametric regression models for the dispersion, zero-inflation, and mean parameters. Their

Table 1: Summary of variables in the patent citation dataset.

| Variable | Description | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| opp | Patent was opposed (1=yes, 0=no) | 0.41 | - | 0 | 1 |
| biopharm | Patent from biopharmaceutical sector (1=yes, 0=no) | 0.44 | - | 0 | 1 |
| ustwin | U.S. "twin" patent exists (1=yes, 0=no) | 0.61 | - | 0 | 1 |
| patus | Patent holder is from U.S (1=yes, 0=no) | 0.33 | - | 0 | 1 |
| patgsgr | Patent holder is from Germany, Switzerland, or Great Britain (1=yes, 0=no) | 0.24 | - | 0 | 1 |
| year | Grant year | - | - | 1980 | 1997 |
| ncountry | Number of designated states for the patent | 7.8 | 4.12 | 1 | 17 |
| nclaims | Number of claims against the patent | 12.3 | 8.13 | 1 | 50 |
| ncit | Number of citations of the patent | 1.6 | 2.71 | 0 | 40 |

selected model (StAR-1) is as follows:

$$\log[f(x)] = \beta_0^\mu + \beta_1^\mu \text{opp} + \beta_2^\mu \text{biopharm} + \beta_3^\mu \text{patus} + \beta_4^\mu \text{patsgr} \tag{54}$$

$$+ f_1^\mu(\text{ncountry}) + f_2^\mu(\text{year}) + f_3^\mu(\text{nclaims}) \tag{55}$$

$$\text{logit}[1 - \omega(x)] = \beta_0^\omega + \beta_1^\omega \text{biopharm} + \beta_2^\omega(\text{year-1991}) + f_1^\omega(\text{ncountry}) \tag{56}$$

$$\log[\kappa(x)] = \beta_0^\kappa + \beta_1^\kappa \text{patus} + \beta_2^\kappa \text{patgsgr}. \tag{57}$$

The functions $f_1^\mu, f_2^\mu, f_3^\mu$, and $f_1^\omega$ are modeled via cubic B-spline expansions using 20 knots, with shrinkage priors on the coefficients (Klein et al., 2015). We also consider two other specifications: A model that has the same specifications for $f(\mathbf{x})$ and $\omega(\mathbf{x})$ as above but a constant $\kappa$ (StAR-2), and a "saturated" model that has a constant $\kappa$, and additive models for $f(\mathbf{x})$ and $\omega(\mathbf{x})$ that include main effects for all categorical covariates and univariate B-spline basis expansions for each of the three continuous variables (StAR-3). We consider constant $\kappa$ models to compare results with ZINB-BART, which also uses a single dispersion parameter, and the "saturated" model is included to give some indication of the necessity of selection in this class of models. Prior distributions for the nonparametric components are the same as in Klein et al. (2015). Posterior sampling was carried out via MCMC using the BayesX software package (Belitz et al., 2016).

As an alternative we consider a single ZINB-BART model with reasonable defaults -

$f(\mathbf{x})$ has a log-linear BART prior with 200 trees and $a_0 = 2$, so that the marginal prior on $\mu(\mathbf{x})$ puts approximately 95% probability over the range $(0.02, 50)$. The excess zero probability $1 - \omega(\mathbf{x})$ has a logistic BART prior with 200 total trees and $a_0 = 3.5\sqrt{2}$, so that $\Pr(|\text{logit}[1 - \omega(\mathbf{x})]| < 7) \approx 0.95$. The dispersion parameter $\kappa$ has a beta-prime prior with $a_\kappa = 5,\ b_\kappa = 3$, yielding a prior mode of 1, $E(\kappa) = 2.5$, and $Var(\kappa) = 8.75$.

### 5.2.1 Results

We apply the same outlier removal rule as Klein et al. (2015), deleting observations with over 50 claims against them. (B-spline models are sensitive to outliers; ZINB-BART's tree-based basis functions are not and ZINB-BART's fits are essentially unchanged when including these points.) The models are evaluated based on the Watanabe-Akaike/"widely applicable" information criterion (WAIC) (Watanabe, 2010, 2013), defined as

$$WAIC = -2\sum_{i=1}^{n} \log\left(\mathrm{E}[p(y_i \mid \mathbf{x}_i, \Theta)]\right) + 2\sum_{i=1}^{n} \mathrm{Var}[\log\{p(y_i \mid \mathbf{x}_i, \Theta)\}], \tag{58}$$

where the expectations and variances are with respect to the posterior over $\Theta$ (overloading $\Theta$ for the moment to represent *all* the parameters, including any trees and their parameters). The first term is the log of the predictive density at each data point (LPD), and the second term is a measure of the effective number of parameters ($p_{waic}$). The WAIC has a number of desirable features over other information criteria: As noted by Gelman et al. (2014), it averages over the posterior rather than conditioning on a point estimate, is invariant to reparameterization, and is more readily justified outside of regular parametric models.

Table 2 shows that ZINB-BART has the lowest WAIC of all models considered, despite StAR-1 being chosen via stepwise selection and being somewhat more flexible in allowing the dispersion parameter $\kappa$ to vary with covariates. The estimated values of $p_{waic}$ show that all three StAR models have similar complexity, with the saturated model having approximately 11 additional effective parameters due to the additional nonlinear partial effects. However, this saturated model underperforms all the others – the extra complexity swamps the mild increase in estimated predictive log likelihood. ZINB-BART has significantly more effective parameters (about 132 compared to 43-54) but a much higher predictive likelihood. The effective number of parameters is also far fewer than the actual number of parameters

Table 2: Comparison of the four competing models of the patent citation data.

| | LPD | $p_{waic}$ | WAIC |
|---|---|---|---|
| StAR-1 (stepwise DIC) | -7783.5 | 43.6 | 15654.24 |
| StAR-2 (stepwise DIC, constant $\kappa$) | -7801.7 | 43.9 | 15691.14 |
| StAR-3 (saturated additive model, constant $\kappa$) | -7793.6 | 54.2 | 15695.48 |
| ZINB-BART | -7688.2 | 131.5 | **15639.47** |

- a total of 400 regression trees and their associated leaf parameters, plus $\kappa$, due to the strong regularizing priors.

It is difficult to pinpoint exactly why ZINB-BART outperforms the other models, since summarizing the posterior distribution of nonparametric regression models like ZINB-BART is challenging. However, there are some interactions that may be important and are sensible based on subject matter considerations. For example, there seems to be an interaction effect between biopharm and year. This is supported by the existing literature; due to regulatory hurdles, biopharmaceutical innovations take more time to reach the market and be generally recognized (Jaffe and Trajtenberg, 1996). Therefore we would expect to see a higher probability of an excess zero in recent years for biopharmaceutical patents.

This effect is captured in the ZINB-BART fit. The first row of Figure 5 displays summaries of the posterior over $\text{logit}[1-\omega(\mathbf{x})]$. In the leftmost plot the solid center line is the *partial dependence* (PD) function (Friedman, 2001) defined as

$$\hat{f}_j(t) = \frac{1}{n}\sum_{i=1}^{n} \text{logit}[1-\omega(\tilde{\mathbf{x}}_i)], \tag{59}$$

where $\tilde{x}_{ik} = x_{ik}$ for $k \neq j$ and $x_{ij} = t$. Here the $j^{th}$covariate is year. As suggested by Goldstein et al. (2015), we also plot a 10% sample of the individual response functions $f(\tilde{\mathbf{x}}_i)$, with dots indicating the actual year (PD plots alone can be misleading in the presence of interactions). The middle plot centers each of the curves at their 1980 value, which makes the interaction apparent: Recent biopharm patents are more likely to have excess zeros than non-biopharm patents. The rightmost plot displays mean-centered PD functions computed across the sample (in gray) and separately for biopharm/non-biopharm patents. On average, older biopharm patents are *less* likely to have excess zeros than contemporaneous non-biopharm patents. The pattern is reversed for recent patents. The second row

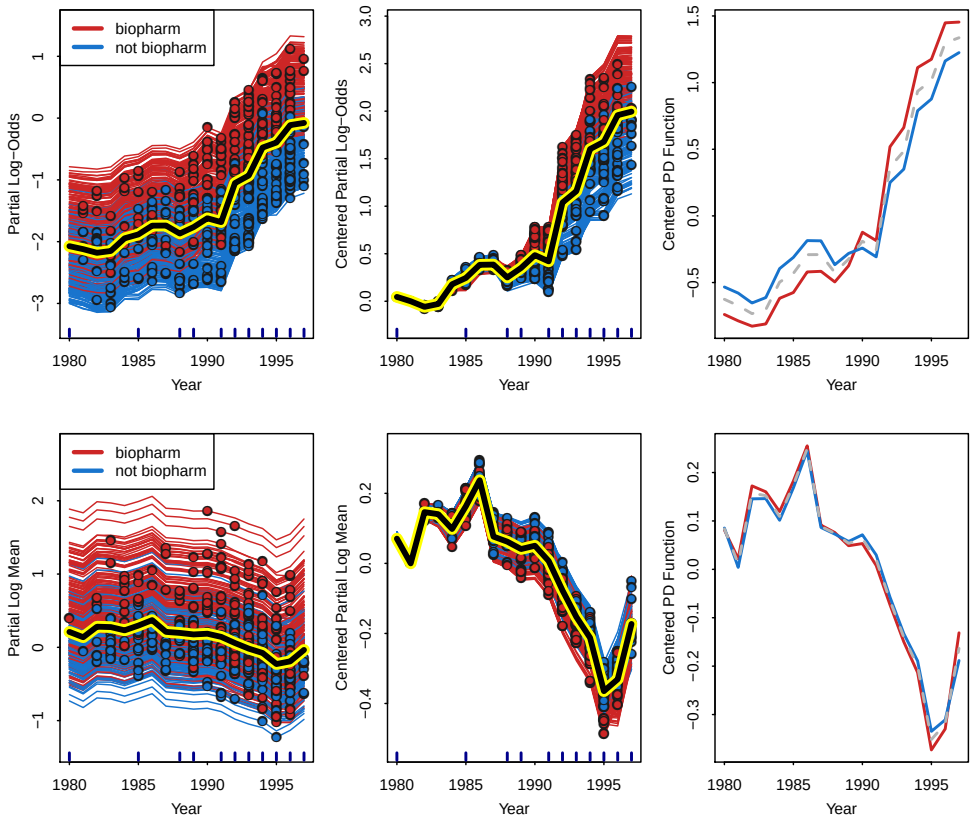


Figure 5: Log-odds functions $\text{logit}[1-\omega(\mathbf{x})]$ are given in the top row, log mean functions $\log[f(\mathbf{x})]$ are in the bottom row. (Left) Partial dependence function (solid line) and a 10% sample of the functions as year varies. (Center) The same as the right panel, except all curves are centered at their value in 1980. (Right) Partial dependence functions computed in the entire sample (dashed gray) and in biopharm/non-biopharm subgroups separately

of Figure 5 shows the same set of plots for $\log[\mu(\mathbf{x}_i)]$, where no such pattern is apparent.

In summary, the ZINB-BART model fits much better than additive semiparametric alternatives. This comes at some cost in summarizing and interpreting the fit, which would seem to be an advantage of the additive model. However, the results proposed by Klein et al. (2015) utilize stepwise selection on the entire dataset to select a model. Subsequent inferences are not strictly valid from a Bayesian perspective due to the double use of the data, and we should not expect them to have frequentist validity either (see e.g. Berk et al. (2013)). Fitting a single nonparametric model like ZINB-BART avoids this issue, and

despite the challenge of summarizing the posterior distribution we have seen that ZINB-BART can detect meaningful, interpretable interactions and nonlinearities that were not specified *a priori* without relying on explicit model selection. In this example the computational costs for each method are similar; MCMC for ZINB BART took approximately 8 minutes, while fitting a single StAR model took about 4 minutes to obtain similar effective sample sizes for the linear predictors (not accounting for the stepwise model selection).

# 6 Conclusion

We have introduced a novel prior and MCMC sampler that allow us to efficiently extend BART to log-linear models for unordered categorical and count responses. We expect that these models will be useful in a variety of settings, given the range of applied problems where the original BART model and its extensions have been successfully deployed. Like the original BART model, log-linear BART is highly modular and amenable to embedding within larger models for more complex applications. The use of a logistic regression BART model in the context of zero-inflated count data is just the one step in this direction.

These priors and algorithms can be used to fit a wide range of models including ordinal models like the continuation ratio logit, as well as hurdle versions of the Poisson and negative binomial models, with different data augmentation techniques. As another concrete example, in the supplemental material (Section A.5) we describe how to fit models for continuous data with covariate-dependent heteroscedasticity using the methods in this paper. (This model using a different prior distribution was presented by McCulloch (2015), concurrently with a preliminary version of this paper.)

There are a some important areas for future work: Summarizing the fit of complicated nonparametric models like BART is difficult. Other authors – beginning with CGM – have proposed variable selection procedures for BART that could be applied in log-linear BART directly (Bleich et al., 2014). Additionally, Linero (2016+) recently introduced a modification of CGM's tree prior that is more suitable for high-dimensional settings and provides a measure of variable importance. This prior is immediately applicable to log-linear BART models. However, detecting and summarizing interesting features like interactions and nonlinearities from a BART fit remains an open problem.

# References


J. H. Albert and S. Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669, June 1993. ISSN 01621459.

S. G. Baker. The multinomial-poisson transformation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(4):495–504, 1994.

C. Belitz, A. Brezger, T. Kneib, S. Lang, and N. Umlauf. Bayesx-software for Bayesian inference in structured additive regression models (version 3.0.2), 2016. URL `http://www.bayesx.org`.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Annals of statistics*, 41(2):802–837, April 2013. ISSN 0090-5364, 2168-8966.

J. Bleich and A. Kapelner. Bayesian additive regression trees with parametric models of heteroskedasticity. *arXiv preprint arXiv:1402.5397*, 2014.

J. Bleich, A. Kapelner, E. I. George, and S. T. Jensen. Variable selection for BART: An application to gene regulation. *The annals of applied statistics*, 8(3):1750–1781, September 2014. ISSN 1932-6157, 1941-7330.

L. F. Burgette and P. R. Hahn. Symmetric Bayesian multinomial probit models. *Duke University Statistical Science Technical Report*, pages 1–20, 2010.

L. F. Burgette and E. V. Nordheim. The trace restriction: An alternative identification strategy for the Bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410, 2012.

F. Caron and A. Doucet. Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012. ISSN 1061-8600.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, March 2010. ISSN 1941-7330.

D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1 June 1998. ISSN 0006-3444.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

J. J. Forster. Bayesian inference for poisson and multinomial log-linear models. *Statistical Methodology*, 7(3):210–224, 2010.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.

S. Frühwirth-Schnatter and R. Frühwirth. Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz, editors, *Statistical Modelling and Regression Structures*, pages 111–132. Physica-Verlag HD, 2010. ISBN 9783790824124, 9783790824131.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016, 2014. ISSN 0960-3174, 1573-1375.

A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. ISSN 1061-8600.

P. J. Green. Reversible jump markov chain monte carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1 December 1995. ISSN 0006-3444.

T. Hastie and R. Tibshirani. Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223, 1 August 2000. ISSN 0883-4237, 2168-8745.

C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, March 2006. ISSN 1936-0975, 1931-6690.

A. B. Jaffe and M. Trajtenberg. Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographic boundaries. *Proceedings of the National Academy of Sciences*, 93(23):12671–12677, 1996.

B. P. Kindo, H. Wang, and E. A. Peña. Multinomial probit Bayesian additive regression trees. *Stat*, 5(1):119–131, 1 January 2016. ISSN 0038-9986, 2049-1573.

N. Klein, T. Kneib, and S. Lang. Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and OverdispersedCountData. *Journal of the American Statistical Association*, 110(509):405–419, April 2015. ISSN 0162-1459.

A. R. Linero. Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association (accepted)*, 0(ja):0–0, 2016+.

J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1 March 1994. ISSN 0006-3444.

R. McCulloch. Nonparametric heteroscedastic regression modeling, Bayesian regression trees and MCMC sampling. Presented at the 10th Conference on Bayesian Nonparametrics, Raleigh, NC, 2015.

L. E. Nieto-Barajas, I. Prünster, and S. G. Walker. Normalized random measures driven by increasing additive processes. *Annals of statistics*, 32(6):2343–2360, December 2004. ISSN 0090-5364, 2168-8966.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108 (504):1339–1349, 2013. ISSN 0162-1459.

M. T. Pratola. Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian analysis*, 11(3):885–911, September 2016. ISSN 1936-0975, 1931-6690.

G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 59(2):291–317, 1 January 1997. ISSN 1369-7412, 1467-9868.

R. A. Sparapani, B. R. Logan, R. E. McCulloch, and P. W. Laud. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in medicine*, 7 February 2016. ISSN 0277-6715, 1097-0258.

S. G. Walker. Posterior sampling when the normalizing constant is unknown. *Communications in Statistics - Simulation and Computation*, 40(5):784–792, 2011.

S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research: JMLR*, 11(Dec):3571–3594, 2010. ISSN 1532-4435.

S. Watanabe. A widely applicable Bayesian information criterion. *Journal of machine learning research: JMLR*, 14(Mar):867–897, 2013. ISSN 1532-4435.

Y. Wu, H. Tjelmeland, and M. West. Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66, 2007.

M. Zhou, L. Li, D. Dunson, and L. Carin. Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access, 2012.

# A Supplemental Material

## A.1 Parameterizing Logistic BART Models

When $n_i = 1$ for all $i$ and $c = 2$, so that $y$ is a binary vector, we recover the binary regression model

$$p_B(y_i) = \left(\frac{f^{(0)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)}\right)^{(1-y_i)} \left(\frac{f^{(1)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)}\right)^{y_i}, \tag{60}$$

which is a logistic regression model with log odds of success $\log[f^{(1)}(\mathbf{x}_i)] - \log[f^{(0)}(\mathbf{x}_i)]$. If $f^{(0)}$ and $f^{(1)}$ have the same number of trees (say $m$) and precision parameter $a_0^2$ then under our prior $\tilde{f}^{(1)}(\mathbf{x}_i) \stackrel{d}{=} f^{(1)}(\mathbf{x}_i)/f^{(0)}(\mathbf{x}_i)$, where $\tilde{f}^{(1)}$ has $2m$ trees and concentration parameter $2a_0^2$, so can write the model equivalently as

$$p_B(y_i) = \left(\frac{1}{1+\tilde{f}^{(1)}(\mathbf{x}_i)}\right)^{(1-y_i)} \left(\frac{\tilde{f}^{(1)}(\mathbf{x}_i)}{1+\tilde{f}^{(1)}(\mathbf{x}_i)}\right)^{y_i}, \tag{61}$$

in terms of the identified parameter $\tilde{f}^{(1)}(\mathbf{x}_i)$. The prior and likelihood (and therefore the posterior) are identical, but the performance of the data augmented MCMC algorithm can be substantially different for extreme probabilities.

To illustrate we consider a simple synthetic example. The probabilities are given by

$$\Pr(y_i \mid \mathbf{x}_i = x) = \exp[f^*(\mathbf{x}_i)]/(1+\exp[f^*(\mathbf{x}_i)]), \quad f^*(x) = 12(\mathbf{x}_i - 0.5) \tag{62}$$

so that the true log odds range over $\pm 6$, yielding probabilities in $(0.0025, 0.9975)$. The covariates are placed (not sampled) uniformly over $(0,1)$. In addition to the identified and unidentified logit models we compare the BART probit model introduced by CGM, which assumes that

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Phi[f_{CGM}(\mathbf{x}_i)] \tag{63}$$

where $f_{CGM}(\cdot)$ has the original BART prior with leaf parameters $\mu_{ht} \sim N(0, 1.5^2/m)$, so that $\Pr(|f(\mathbf{x}_i)| < 3) \equiv \Pr(\Phi[-3] < \Phi[f(\mathbf{x}_i)] < \Phi[3]) = 0.95$ *a priori*. This is approximately the true range of the probabilities in our synthetic example, and we use the same condition to set $a_0$ in the logistic models.

We generated 25 datasets of size $n = 100$ and ran the MCMC algorithm for 6,000 iterations, discarding the first 1,000 as burn-in. We estimate the log odds function at each covariate value. They are given by $\log[f^{(1)}(\mathbf{x}_i)] - \log[f^{(0)}(\mathbf{x}_i)]$ for the unidentified logit model, $\log[\tilde{f}^{(1)}(\mathbf{x}_i)]$ for the identified logit model, and $\log(\Phi[f(\mathbf{x}_i)]) - \log(1 - \Phi[f(\mathbf{x}_i)])$ for BART probit. We compare the average effective sample size (computed using the R package `coda`) over the 25 replicates. Unlike the two logit models, BART probit has a different target distribution and therefore the effective sample sizes are not directly comparable. We include it in the comparison primarily to illustrate the operating characteristics of a similar, well-known data augmentation scheme.

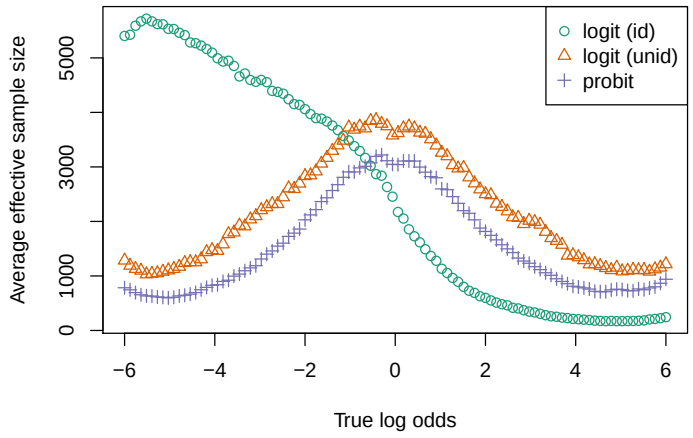


Figure 6: Average effective sample size for estimates of the log odds for the simulated datasets in Section A.1. The labels "logit (id)" and "logit (unid)" refer to the identified/unidentified parameterizations of the logistic regression model (Eqs (61) and (60), respectively).

Figure 6 shows the results. The two logit models perform much differently; the identified model mixes extremely well when the log odds are small and extremely poorly when they are large. The unidentified model mixes best near 0, degrading as the log odds increase in magnitude. But the unidentified parameterization has a minimum average ESS of about 1,030, compared to 170 for the identified parameterization, so the unidentified paramterization has the benefit of performing adequately everywhere.

Examining the two data augmentation schemes sheds some light on this behavior. In the identified parameterization, the latent variables are sampled $\phi_i \sim Gamma(1, 1 + \tilde{f}^{(1)}(\mathbf{x}))$ prior to updating $\tilde{f}^{(1)}$. When $\tilde{f}^{(1)}(\mathbf{x}_i)$ is large, the variance of this full conditional is small and $\phi_i$ will make small moves leading to high cross-correlation between the parameters determining $f^{(1)}(\mathbf{x}_i)$ and $\phi_i$. When $\tilde{f}^{(1)}(\mathbf{x}_i)$ is small the full conditional is much more dispersed, reducing the crosscorrelation and leading to the excellent behavior in Fig 6. In the unidentified parameterization, $\phi_i \sim Gamma(1, f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}))$. For probabilities near 1, increases in $f^{(1)}(\mathbf{x}_i)$ are offset by compensatory decreases in $f^{(0)}(\mathbf{x}_i)$ (which is fixed

at 1 in the identified sampler), since our prior for both functions is centered at 1. However, for probabilities at either extreme $|f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x})|$ will tend to be large and induce a greater degree of crosscorrelation. For probabilities near 0.5, $f^{(0)}(\mathbf{x}_i) \approx f^{(1)}(\mathbf{x}_i)$ and fit is freely allocated between $f^{(1)}$ and $f^{(0)}$. Hence MCMC in the unidentified parameterization behaves similarly to the BART-probit sampler, which is constructed using Albert and Chib (1993)'s data augmentation:

1. Sample $\phi_i \sim N(f(\mathbf{x}_i), 1)\mathbb{1}(\phi_i > 0)$ if $y_i = 1$ or $\phi_i \sim (f(\mathbf{x}_i), 1)\mathbb{1}(\phi_i < 0)$ if $y_i = 0$
2. Update $f(\cdot)$ via the CGM MCMC algorithm

Here the crosscorrleation between the parameters determining $f(\mathbf{x}_i)$ and $\phi_i$ is weakest when $f(\mathbf{x}_i) \approx 0$, or 0 on the log odds scale.

Our results suggest that the identified parmaterization of the logit model mixes more efficiently for log odds less than about $-1$ (probabilities less than about 0.27). Therefore the most efficient parameterization will depend on factors including the balance of the outcome as well as the distribution of the covariates and their discriminative power. If it is known that the outcome is rare and the predictors relatively weak working in the identified parameterization may be more efficient. In the absence of such strong prior knowledge the unidentified parameterization yields good results across a range of settings, and is more sensible as a default. In the case of the multinomial regression model it also avoids the risk of accidentally specifying a poor prior through an inappropriate choice of reference category, since all outcome values are treated symmetrically.

## A.2 Proof of propositions

**Proposition 3.1:** Collecting terms in the augmented variables, we have:

$$p_{ZINB}(y_i, Z_i, \phi_i, \xi_i \mid f^{(0)}, f_i, \kappa, f) \tag{64}$$

$$= f^{(0)}(\mathbf{x}_i)^{1-Z_i} f^{(1)}(\mathbf{x}_i)^{Z_i} \exp[-\phi_i\{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)\}] \tag{65}$$

$$\left\{\frac{1}{\Gamma(\kappa)y_i!}\kappa^{\kappa}[\mu_{0i}f(\mathbf{x}_i)]^{y_i}\xi_i^{\kappa+y_i-1}\exp\left[-\xi_i(\kappa + \mu_{0i}f(\mathbf{x}_i))\right]\right\}^{Z_i} \tag{66}$$

$$\mathbb{1}(Z_i = 1 \text{ if } y_i > 0). \tag{67}$$

Since $\int_0^\infty t^{u-1}\exp(-st)dt = \Gamma(u)/s^u$,

$$\int_0^\infty \int_0^\infty p_{ZINB}(y_i, Z_i, \phi_i, \xi_i \mid f^{(0)}, f_i, \kappa, f)d\phi_i d\xi_i \tag{68}$$

$$= f^{(0)}(\mathbf{x}_i)^{1-Z_i} f^{(1)}(\mathbf{x}_i)^{Z_i} \int_0^\infty \exp[-\phi_i\{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)\}]d\phi_i \tag{69}$$

$$\times \left\{ \frac{1}{\Gamma(\kappa)y_i!}\kappa^\kappa[\mu_{0i}f(\mathbf{x}_i)]^{y_i} \int_0^\infty \xi_i^{\kappa+y_i-1}\exp\left[-\xi_i(\kappa + \mu_{0i}f(\mathbf{x}_i))\right] d\xi_i \right\}^{Z_i} \tag{70}$$

$$\times \mathbb{1}(Z_i = 1 \text{ if } y_i > 0) \tag{71}$$

$$= \frac{f^{(0)}(\mathbf{x}_i)^{1-Z_i} f^{(1)}(\mathbf{x}_i)^{Z_i}}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} \tag{72}$$

$$\times \left\{ \frac{1}{\Gamma(\kappa)y_i!}\kappa^\kappa[\mu_{0i}f(\mathbf{x}_i)]^{y_i} \frac{\Gamma(\kappa + y_i)}{(\kappa + \mu_{0i}f(\mathbf{x}_i))^{\kappa+y_i}} \right\}^{Z_i} \tag{73}$$

$$\times \mathbb{1}(Z_i = 1 \text{ if } y_i > 0) \tag{74}$$

$$= \frac{f^{(0)}(\mathbf{x}_i)^{1-Z_i} f^{(1)}(\mathbf{x}_i)^{Z_i}}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} \tag{75}$$

$$\times \left\{ \frac{\Gamma(\kappa + y_i)}{\Gamma(\kappa)y_i!} \left(\frac{\kappa}{\kappa + \mu_{0i}f(\mathbf{x}_i)}\right)^\kappa \left(\frac{\mu_{0i}f(\mathbf{x}_i)}{\kappa + \mu_{0i}f(\mathbf{x}_i)}\right)^{y_i} \right\}^{Z_i} \tag{76}$$

$$\times \mathbb{1}(Z_i = 1 \text{ if } y_i > 0). \tag{77}$$

To sum over $Z_i$, consider the two cases $y_i > 0$ and $y_i = 0$. If $y_i = 0$ then

$$\sum_{Z_i=0}^{1} \int_0^\infty \int_0^\infty p_{ZINB}(y_i, Z_i, \phi_i, \xi_i \mid f^{(0)}, f_i, \kappa, f)d\phi_i d\xi_i \tag{78}$$

$$= \frac{f^{(1)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} \frac{\Gamma(\kappa + y_i)}{\Gamma(\kappa)y_i!} \left(\frac{\kappa}{\kappa + \mu_{0i}f(\mathbf{x}_i)}\right)^\kappa \left(\frac{\mu_{0i}f(\mathbf{x}_i)}{\kappa + \mu_{0i}f(\mathbf{x}_i)}\right)^{y_i} \tag{79}$$

due to the indicator function. Otherwise if $y_i = 1$ then

$$\sum_{Z_i=0}^{1} \int_0^\infty \int_0^\infty p_{ZINB}(y_i, Z_i, \phi_i, \xi_i \mid f^{(0)}, f_i, \kappa, f)d\phi_i d\xi_i \tag{80}$$

$$= \frac{f^{(0)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} \tag{81}$$

$$+ \frac{f^{(1)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)} \frac{\Gamma(\kappa + y_i)}{\Gamma(\kappa)y_i!} \left(\frac{\kappa}{\kappa + \mu_{0i}f(\mathbf{x}_i)}\right)^\kappa \left(\frac{\mu_{0i}f(\mathbf{x}_i)}{\kappa + \mu_{0i}f(\mathbf{x}_i)}\right)^{y_i} \tag{82}$$

So we have

$$\sum_{Z_i=0}^{1}\int_0^\infty\int_0^\infty p_{ZINB}(y_i, Z_i, \phi_i, \xi_i \mid f^{(0)}, f_i, \kappa, f)d\phi_i d\xi_i \tag{83}$$

$$=\frac{f^{(0)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i)+f^{(1)}(\mathbf{x}_i)}\mathbb{1}(y_i=0) \tag{84}$$

$$+\frac{f^{(1)}(\mathbf{x}_i)}{f^{(0)}(\mathbf{x}_i)+f^{(1)}(\mathbf{x}_i)}\frac{\Gamma(\kappa+y_i)}{\Gamma(\kappa)y_i!}\left(\frac{\kappa}{\kappa+\mu_{0i}f(\mathbf{x}_i)}\right)^{\kappa}\left(\frac{\mu_{0i}f(\mathbf{x}_i)}{\kappa+\mu_{0i}f(\mathbf{x}_i)}\right)^{y_i} \tag{85}$$

as required.

**Proposition 4.1:** To set the parameters $c$ and $d$ from $a_0$ and $m$, note that

$$\mathrm{E}[\log(\lambda_{ht})] = 0 \tag{86}$$

$$\mathrm{Var}[\log(\lambda_{ht})] = \psi''(c) + [\psi'(c) - \log(d)]^2, \tag{87}$$

where $\psi(c) = \log[\Gamma(c)]$. Enforcing $\mathrm{Var}[\log(\lambda_{ht})] = a_0^2/m$ requires that $c,\ d$ solve

$$\psi''(c) + [\psi'(c) - \log(d)]^2 - a_0^2/m = 0. \tag{88}$$

The real roots of (88) are given by

$$d = \exp\left[\sqrt{a_0^2/m - \psi''(c)} \pm \psi'(c)\right], \tag{89}$$

subject to $a_0^2/m - \psi''(c) \geq 0$. Taking $a_0^2/m - \psi''(c) = 0$ minimizes $d$, which concentrates more mass around zero on the log scale and is an appropriate choice for a strong regularizing prior. So $c$ is obtained numerically as the solution to $\psi''(c) = a_0^2/m$, which is trivial as $\psi''(c)$ is monotonically decreasing, and $d = \exp[\psi'(c)]$.

**Proposition 4.2:** The exact solutions for the parameters in Proposition 4.1 can be approximated by $c \approx m/a_0^2 + 0.5$ and $d \approx m/a_0^2$. Let $v = a_0^2/m$. Typically $v$ will be quite small, so $c$ will be large. The Laurent series of $\psi''(c)$ at $c = \infty$ is

$$\frac{1}{c} + \frac{1}{2c^2} + O\left(\frac{1}{c^3}\right) \tag{90}$$

Using the first two terms of the series to approximate $\psi''(c)$ we want to solve $v = \frac{1}{c} + \frac{1}{2c^2}$. Since $v, c$ are both positive, the only solution is

$$c = \frac{1 + \sqrt{1 + 2v}}{2v} \tag{91}$$

We can obtain a simpler expression with one more approximation:

$$\begin{aligned} c &= \frac{1+\sqrt{1+2v}}{2v} \\ &= \frac{1}{2v} + \sqrt{\frac{1}{4v^2} + \frac{1}{2v}} \\ &\approx \frac{1}{2v} + \sqrt{\frac{1}{4v^2} + \frac{1}{2v} + \frac{1}{4}} \\ &= \frac{1}{2v} + \sqrt{\left(\frac{1}{2v} + \frac{1}{2}\right)^2} \\ &= \frac{1}{v} + \frac{1}{2} \\ &= \frac{m}{a_0^2} + \frac{1}{2} \end{aligned}$$

The expansion of $\exp[\psi'(c)]$ at $c = \infty$ is $c - 0.5 + O(1/c)$, so $d \approx m/a_0^2$. Thus when $m >> a_0^2$, we have $c \approx m/a_0^2 + 0.5$ and $d \approx m/a_0^2$. For all the settings of $m$ and $a_0$ considered in this paper, the largest relative error under these approximations is less than 2% for both $c$ and $d$. These include some extreme settings from the cross validation exercise, however, and the approximation is usually much better. For example, the multinomial logistic regression default parameter setting $m = 100, a_0 = 3.5/\sqrt{2}$ yields an approximation with less then 0.03% relative error.

## A.3 MCMC for ZINB-BART

A single step of the ZINB MCMC algorithm proceeds as follows:

1. Block update $(\kappa, Z, \xi, \phi \mid -)$ by composition. (These steps are order-dependent.)

   (a) First sample $\kappa$ from

   $$p(\kappa \mid y, \mathbf{x}, f, f^{(0)}, f^{(1)}, \kappa) \propto p(\kappa) \prod_{i=1}^{n} p_{ZINB}(y_i \mid \mathbf{x}_i, f, f^{(0)}, f^{(1)}, \kappa) \tag{92}$$

   using a Metropolis-Hastings step.

   (b) Given the new value for $\kappa$, sample $Z$. The $Z_i$'s are mutually independent given $\kappa, \omega$ and $f$. If $y_i > 0$, $Z_i = 1$. Otherwise $p(Z_i \mid \kappa, \omega, f)$ is Bernoulli with probability

   $$\frac{\omega(\mathbf{x}_i) p_{NB}(0 \mid \mathbf{x}_i, \kappa, f)}{1 - \omega(\mathbf{x}_i) + \omega(\mathbf{x}_i) p_{NB}(0 \mid \mathbf{x}_i, \kappa, f)}. \tag{93}$$

(c) Finally $(\xi, \phi)$ are sampled from their joint full conditional. This is particularly simple due to their conditional independence: For all observations with $Z_i = 1$, sample $\xi_i$ independently from

$$(\xi_i \mid -) \sim G(\kappa + y_i, \kappa + \mu_{0i} f(\mathbf{x}_i)), \tag{94}$$

and for each $1 \leq i \leq n$ sample

$$(\phi_i \mid -) \sim Exp(f^{(0)}(\mathbf{x}_i) + f^{(1)}(\mathbf{x}_i)). \tag{95}$$

2. Update $f^{(j)}$ for $j \in \{0, 1\}$ using Algorithm 1 and the expressions in Section 4.2 with

$$r_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}^{(j)}} \mathbb{1}(Z_i = j), \quad s_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}^{(j)}} \phi_i f_{(h)}^{(j)}(\mathbf{x}_i)$$

where $f_{(h)}^{(j)}(\mathbf{x}_i) = \prod_{l \neq h} g(\mathbf{x}, T_h^{(j)}, \Lambda_h^{(j)})$ is the fit from all but the $h^{th}$ tree.

3. Update $f$ using Algorithm 1 and the expressions in Section 4.2 with

$$r_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}} Z_i y_i, \quad s_{ht} = \sum_{i:\mathbf{x}_i \in A_{ht}} Z_i \xi_i \mu_{0i} f_{(h)}(\mathbf{x}_i)$$

where $f_{(h)}(\mathbf{x}_i) = \prod_{l \neq h} g(\mathbf{x}, T_h, \Lambda_h)$ is the fit from all but the $h^{th}$ tree.

Note that all three regression functions can be updated in parallel, as they are conditionally independent given the latent variables.

## A.4 Additional Classification Study Results

| Dataset | rf | gbm | mno | svm | nnet | BART-cv | BART-default 200 | BART-default 100 per |
|---|---|---|---|---|---|---|---|---|
| car | 0.985 (0.0095) | 0.976 (0.0057) | 0.833 (0.0129) | 0.768 (0.0387) | 0.961 (0.0185) | 0.979 (0.0082) | 0.962 (0.0107) | 0.953 (0.0124) |
| dermatology | 0.978 (0.0292) | 0.981 (0.0292) | 0.975 (0.0239) | 0.743 (0.0285) | 0.978 (0.0332) | 0.981 (0.0263) | 0.978 (0.0254) | 0.978 (0.0286) |
| lymphography | 0.859 (0.1232) | 0.84 (0.0674) | 0.812 (0.1022) | 0 (0) | 0.834 (0.1144) | 0.846 (0.1078) | 0.826 (0.115) | 0.833 (0.1108) |
| seeds | 0.948 (0.0351) | 0.952 (0.0449) | 0.957 (0.0351) | 0.924 (0.0402) | 0.962 (0.0376) | 0.943 (0.0376) | 0.933 (0.046) | 0.933 (0.046) |
| balance-scale | 0.842 (0.0207) | 0.923 (0.0068) | 0.899 (0.0349) | 0.912 (0.0407) | 0.97 (0.0205) | 0.933 (0.0123) | 0.922 (0.0045) | 0.922 (0.0045) |
| cardiotocography-3clases | 0.948 (0.0201) | 0.948 (0.0133) | 0.896 (0.0178) | 0.917 (0.0157) | 0.909 (0.0254) | 0.948 (0.0097) | 0.937 (0.0155) | 0.937 (0.0133) |
| contrac | 0.54 (0.0344) | 0.564 (0.0215) | 0.521 (0.0475) | 0.548 (0.0389) | 0.556 (0.0269) | 0.566 (0.0449) | 0.565 (0.0462) | 0.566 (0.0449) |
| glass | 0.809 (0.1031) | 0.794 (0.0814) | 0.634 (0.0976) | 0.657 (0.084) | 0.664 (0.1353) | 0.8 (0.061) | 0.748 (0.0662) | 0.757 (0.0543) |
| heart-cleveland | 0.579 (0.0431) | 0.591 (0.0794) | 0.592 (0.0789) | 0.637 (0.0516) | 0.63 (0.0633) | 0.598 (0.0383) | 0.578 (0.0553) | 0.589 (0.0579) |
| heart-va | 0.344 (0.0413) | 0.346 (0.0875) | 0.339 (0.1066) | 0.35 (0.1077) | 0.34 (0.1503) | 0.38 (0.1615) | 0.365 (0.143) | 0.37 (0.1512) |
| iris | 0.953 (0.0706) | 0.947 (0.0613) | 0.98 (0.0322) | 0.96 (0.0644) | 0.973 (0.0466) | 0.96 (0.0466) | 0.953 (0.0549) | 0.953 (0.0549) |
| pittsburg-bridges-MATERIAL | 0.852 (0.0837) | 0.852 (0.0692) | 0.822 (0.101) | 0.858 (0.0513) | 0.869 (0.0415) | 0.858 (0.0504) | 0.848 (0.0817) | 0.848 (0.0817) |
| pittsburg-bridges-REL-L | 0.689 (0.1834) | 0.671 (0.0982) | 0.658 (0.072) | 0.721 (0.1173) | 0.709 (0.096) | 0.694 (0.1684) | 0.671 (0.147) | 0.673 (0.135) |
| pittsburg-bridges-SPAN | 0.655 (0.0991) | 0.708 (0.1827) | 0.695 (0.0871) | 0.712 (0.1373) | 0.687 (0.165) | 0.73 (0.0694) | 0.718 (0.0856) | 0.718 (0.0856) |
| pittsburg-bridges-TYPE | 0.65 (0.0724) | 0.618 (0.1861) | 0.55 (0.1185) | 0.568 (0.1006) | 0.55 (0.1533) | 0.586 (0.1176) | 0.568 (0.0825) | 0.568 (0.093) |
| synthetic-control | 0.983 (0.0176) | 0.978 (0.0158) | 0.988 (0.0137) | 0.712 (0.0236) | 0.988 (0.0112) | 0.99 (0.0117) | 0.983 (0.0136) | 0.99 (0.0086) |
| teaching | 0.676 (0.1149) | 0.59 (0.1108) | 0.531 (0.1074) | 0.544 (0.151) | 0.556 (0.1155) | 0.563 (0.1024) | 0.53 (0.0908) | 0.516 (0.0772) |
| vertebral-column-3clases | 0.845 (0.0451) | 0.829 (0.0374) | 0.858 (0.0732) | 0.855 (0.0532) | 0.868 (0.0492) | 0.848 (0.0681) | 0.842 (0.0653) | 0.835 (0.0671) |
| wine | 0.989 (0.0234) | 0.989 (0.0241) | 0.983 (0.0274) | 0.983 (0.0268) | 0.989 (0.0228) | 0.983 (0.0274) | 0.983 (0.0274) | 0.983 (0.0274) |
| wine-quality-red | 0.715 (0.0277) | 0.645 (0.0243) | 0.596 (0.0415) | 0.575 (0.049) | 0.605 (0.0283) | 0.615 (0.03) | 0.605 (0.0335) | 0.607 (0.0332) |

Table 3: Results of the classification study. The cross-validated estimate of out of sample accuracy is given along with its standard deviation in parantheses. "BART-default 200" sets $a_0 = 3.5/\sqrt{2}$ and $m$ set so that there are approximately 200 total trees. "BART-default 100 per" sets $a_0 = 3.5/\sqrt{2}$ and $m = 100$.

## A.5 Covariate-dependent heteroscedastic regression

CGM's BART regression model for continuous data assumed homoscedastic, normally-distributed errors:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{96}$$

CGM demonstrated that this model performed well relative to competitors in a range of simulations. However, when the data exhibit heteroscedasticity this model may over- or under-fit the mean function, and predictive intervals computed from the posterior predictive $p(y_{n+1} \mid \mathbf{x}_{n+1}, \{y_i, \mathbf{x}_i : 1 \leq i \leq n\})$ will be poorly calibrated. Further, the effect of covariates on the variance may be of interest itself. Heteroscedastic BART models with parametric variance functions were introduced in Bleich and Kapelner (2014), where the authors provide the necessary expressions for the integrated likelihood and full conditionals to update $f(\cdot)$ under heteroscedasticity. Here we extend the heteroscedastic BART model to utilize log-linear BART priors for the variance function.

Specifically we consider the following regression model:

$$y_i = f(\mathbf{x}_i) + \sigma(\mathbf{x}_i)\epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma_0^2). \tag{97}$$

where $\sigma^2(\cdot)$ is given a log-linear BART prior. Due to the symmetry of our prior distribution this is equivalent to a log-linear BART prior on $\sigma^{-2}(\cdot)$. We give the mean function $f(\cdot)$ a BART prior with normal priors on the end node parameters as in (5). Rather than centering and scaling $y$ to $\pm 0.5$, we scale $\sigma_\mu$ by $0.5(y_{max} - y_{min})$ in (5) which has much the same effect. The parameter $\sigma_0^2$ can be formally elicited or chosen using a slight adaptation of CGM's heuristic for setting the scale parameter in the prior on the error variance in (96). Larger values of $a_0$ tend to be necessary to avoid overfitting. Taking $a_0 = 1.5$ ensures that the marginal prior for the variance function $1/\sigma(\mathbf{x}_i)$ puts approximately 95% prior probability on $\sigma_0^2 f_v(\mathbf{x}) \in (\sigma_0^2/5, 5\sigma_0^2)$.

### A.5.1 MCMC

The likelihood for a single data point is

$$p(y_i) = \frac{\sigma(\mathbf{x}_i)^{-1}}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{1}{2\sigma(\mathbf{x}_i)\sigma_0^2}\left(y_i - f(\mathbf{x}_i)\right)^2\right]. \tag{98}$$

The log-linear BART prior is immediately conjugate, so no data augmentation is necessary. For updating the trees and parameters in $\sigma(\cdot)$, $f(\cdot)$ is considered fixed. MCMC in the heteroscedastic model proceeds as follows:

1. Update the mean function's trees and node parameters as in Bleich and Kapelner (2014), using $\sigma_0^2\sigma(\mathbf{x}_i)$ as the variance for each observation.

2. Update $\sigma^2(\cdot)$ using Algorithm 1 and the expressions in Section 4.2, with

$$r_{ht} = \frac{1}{2}\sum_{i=1}^{n} \mathbb{1}(\mathbf{x}_i \in \mathcal{A}_{ht}^{(v)}) \tag{99}$$

$$s_{ht} = \frac{1}{2\sigma_0^2} \sum_{i:\mathbf{x}_i \in \mathcal{A}_{ht}^{(v)}} f_{(h)}(\mathbf{x}_i)(y_i - f(\mathbf{x}_i))^2 \tag{100}$$