**1 Introduction**

Bayesian Additive Regression Trees (BART) arose out of earlier research on Bayesian model fitting of an outcome to a single tree (Chipman et al., 1998) which is generally called Classification and Regression Trees (CART) (Loh,

2014). In this era (circa 1998), the outstanding predictive performance of en- semble models was starting to become apparent (Breiman, 1996; Krogh and Solich, 1997; Freund and Schapire, 1997; Breiman, 2001; Friedman, 2001; Baldi and Brunak, 2001). Instead of making a single prediction from a complex model, ensemble models make a prediction which is the summary of the predictions from many simple models. Generally, ensemble models have desirable properties, e.g., on the spectrum from low bias with high variance (such as CART (Breiman,

2017)) to high bias with low variance (linear regression), ensemble models gen- erally fall somewhere in the middle which accords for their performance (Kuhn and Johnson, 2013). With some similarities to bagging (Breiman, 1996), boost- ing (Freund and Schapire, 1997; Friedman, 2001) and random forests (Breiman,

2001), BART relies on an ensemble of trees to predict the outcome.

BART is a Bayesian nonparametric, sum of trees method for continuous, dichotomous, categorical and time-to-event outcomes. Furthermore, BART is a black-box, machine learning method which fits the outcome via an arbitrary random function, *f* , of the covariates. So-called black-box models generate functions of the covariates which are so complex that interpreting the internal details of the fitted model is generally abandoned in favor of assessment via evaluations of the fitted function, *f* , at chosen values of the covariates. As shown by Chipman et al. (2010), BART’s out-of-sample predictive performance is generally equivalent to, or exceeds that, of alternatives like lasso with L1 regularization (Efron et al., 2004) or black-box models such as gradient boost- ing (Freund and Schapire, 1997; Friedman, 2001), neural nets with one hidden layer (Ripley, 2007; Venables and Ripley, 2013) and random forests (Breiman,

2001). Over-fitting is the tendency to over-do the fit of a model to an in-sample training data set’s signal, and particularly its noise, at the expense of poor pre- dictive performance for unseen out-of-sample data (Cristianini, 2014; Cook and Ranstam, 2016). Typically, BART does not over-fit to the training data due to the regularization tree-branching penalty of the BART prior, i.e., generally, each tree in the ensemble has few branches and plays a small part in the overall fit. Essentially, BART is a Bayesian nonlinear model with all the advantages of the Bayesian paradigm such as posterior inference including point and interval estimation. Conveniently, BART naturally scales to large numbers of covari- ates and facilitates variable selection; it does not require the covariates to be rescaled; neither does it require the covariate functional relationship, nor the interactions considered, to be pre-specified.

In this article, we will discuss the BART prior in the context of continuous outcomes in Section 2 along with posterior computations in Section 3. Next, we will briefly discuss BART extensions for dichotomous, categorical and survival outcomes in Section 4. We will then conclude with some recent developments and software implementations (as of this writing) in Section 5.

**2 Binary trees and the BART prior**

Here, we briefly describe binary trees and their relationship to BART; for a more detailed discussion of trees, see (Zhang et al., 2014). BART relies on an ensemble of *H* binary trees which are a type of a directed acyclic graph. We exploit the wooden tree metaphor to its fullest. Each of these trees grows from the ground up starting out as a root node. The root node is generally a branch decision rule, but it doesn’t have to be; occasionally there are trees in the ensemble which are only a root terminal node consisting of a single leaf output value. If the root is a branch decision rule, then it spawns a left and a right node which each can be either a branch decision rule or a terminal leaf value and so on. In binary tree, *T* , there are *C* nodes which are made of *B* branches and *L* leaves: *C* = *B* + *L*. And, there is an algebraic relationship between the number of branches and leaves: *B* = *L −*1. The nodes are numbered in relation to the tree’s tier level, *t*(*n*) = *l*log2 *nJ* as follows.

Tier

*t* 2*t . . .* 2*t*+1 *−*1

.

2 4 5 6 7

1 2 3

0 1

The key to discriminating between branches and leaves is via the algebraic relationship between a branch, *n*, at tree tier *t*(*n*) leading to its left, *l* = 2*n*, and right, *r* = 2*n* + 1, nodes at tier *t*(*n*) + 1, i.e., for each node, besides root, you can determine from which branch it arose and those nodes that are not a branch (since they have no leaves) are necessarily leaves.

Underlying this methodology is the BART prior. The BART prior specifies a flexible class of unknown functions, *f* , from which we can gather randomly generated fits to the given data via the posterior. Let the regression tree function *g*(***x***; *T , M*) assign a value based on the input ***x*** (LeBlanc, 2014). The binary decision tree *T* is represented by a set of ordered triples, (*n, j, k*), representing branch decision rules: *n* for the node, *j* for covariate *xj* and *k* for the cutpoint *cjk* . The branch decision rules are of the form *xj ≤ cjk* which means branch left and *xj > cjk* , branch right; or terminal leaves where it stops. *M* represents leaves and is a set of ordered pairs, (*n, µn* ): *n* for the node, *µn* for the outcome value and *n ∈ L* where *L* is the set of leaves. The function, *f* (***x***), is a sum of *H* regression tree functions:

*H*

*f* (***x***) = *g*(***x***; *Th , Mh* )

*h*=1

where *H* is “large”, let’s say, 50, 100 or 200.

For a continuous outcome, *yi* , we have the following BART regression on the vector of covariates, ***x****i* :

*yi* = *µ*0 + *f* (***x****i* ) + *Ei* where *Ei*

*∼* N (0*, w*2 *σ*2 )

iid

*i*

with *i* indexing subjects *i* = 1*, . . . , N* . The unknown random function, *f* , and the error variance, *σ*2 , follow the BART prior expressed notationally as

prior

(*f, σ*2 )

*∼* BART(*H, µ*0 *, τ, k, α, γ*; *ν, λ, q*)

where *H* is the number of trees, *µ*0 is a known constant which centers *y* and the rest of the parameters will be explained later in this section (for brevity,

we often use the simpler shorthand (*f, σ*2 ) prior

*∼*

BART). The *wi*

are known

standard deviation weight multiples (only available for continuous outcomes)

where the unit weight vector is often assumed. The centering parameter, *µ*0 , where the default is often taken to be *y* for continuous outcomes.

BART is a Bayesian nonparametric prior. Using the Gelfand-Smith generic bracket notation for the specification of random variable distributions (Gelfand and Smith, 1990), we represent the BART prior in terms of the collection of all trees, ***T*** ; collection of all leaves, ***M***; and the error variance, *σ*2 , as the follow-

ing product: (***T*** *,* ***M****, σ*2 1 = (*σ*2 1 [***T*** *,* ***M***] = (*σ*2 1 [***T*** ] [***M****|****T*** ]. Furthermore, the

individual trees themselves are independent: [***T*** *,* ***M***] = Ti*h* [*Th* ] [*Mh |Th* ]. where

[*Th* ] is the prior for the *h*th tree and [*Mh |Th* ] is the collection of leaves for the

*h*th tree. And, finally, the collection of leaves for the *h*th tree are independent: [*Mh |Th* ] = Ti*n* [*µhn |Th* ] where *n* indexes the leaf nodes.

The tree prior: [*Th* ]. There are three prior components of *Th* which govern whether the tree branches grow or are pruned. The first tree prior regularizes the probability of a branch at leaf node *n* in tree tier *t*(*n*) = *l*log2 *nJ* as

P [*Bn* = 1] = *α*(*t*(*n*) + 1)*−γ* (1)

where *Bn* = 1 represents a branch while *Bn* = 0 is a leaf, 0 *< α <* 1 and *γ ≥* 0. The following defaults are recommended: *α* = 0*.*95 and *γ* = 2; for a detailed discussion of these parameter settings, see Chipman et al. (1998). Note that this prior penalizes branch growth, i.e., in prior probability, the default number of branches will likely be 1 or 2. Next, there is a prior dictating the choice of a splitting variable *j* conditional on a branch event *Bn* which defaults to uniform probability *sj* = *P −*1 where *P* is the number of covariates (however, you can specify a Dirichlet prior which is more appropriate if the number of covariates is large (Linero, 2018)). Given a branch event, *Bn* , and a variable chosen, *xj* , the last tree prior selects a cut point, *cjk* , within the range of observed values for *xj* ; this prior is often chosen to be uniform for convenience.

The leaf prior: [*µhn |Th* ]. Given a tree, *Th* , there is a prior on its leaf values,

*µhn |Th* and we denote the collection of all leaves in *Th* by *Mh* = *{*(*n, µhn* ) : *n ∈ Lh }*. Note that *yi ∈* [*y*min *, y*max ] for all *i* and denote (*µ*1(***x****i* ) *, . . . , µH* (***x****i* ) 1 as the

leaf output values from each tree corresponding to the vector of covariates, ***x****i* . If

iid 2

*µh*(***x****i* ) *|Th*

*∼* N (0*, σµ* ), then the model estimate for subject *i* is *µi* = E [*yi |****x****i* ] =

*µ*0 + 'L*h µh*(***x****i* ) where *µi ∼*N (*µ*0 *, H σµ* ). We choose a value for *σµ* which is the

2

solution to the equations *y*min = *µ*0 *− k*

*√H σµ* and *y*max = *µ*0 + *k*

*√H σµ* , i.e.,

2*k√H* . Therefore, we arrive at *µhn*

*σµ* = *y*max *−y*min

prior r

*∼* N 0*,*

r  *τ*

2*k√H*

2

where *τ* =

*y*max *− y*min . So, the prior for *µhn* is weakly informed by the data, *y*, only via the extrema, *y*min and *y*max . The parameter *k* calibrates this prior as follows.

*µi ∼*N

r

*µ*0 *,*

r *τ*  2

2*k*

P [*y*min *≤ µi ≤ y*max ] = Φ(*k*) *−* Φ(*−k*)

Since P [*µi ≤ y*max ] = P

*z ≤* 2*k*

*y*max *− µ*0 l

*τ*

*≈* P [*z ≤ k*] = Φ(*k*)

Similarly P [*µi ≤ y*min ] = Φ(*−k*)

The recommended default choice, *k* = 2, corresponds to *µi* falling within the extrema with approximately 0.95 probability. Values of *k ∈* [1*,* 3] generally yield good results. *k* is a potential candidate parameter for choice via cross-validation.

The error variance prior: (*σ*2 1. The prior for *σ*2 is the conjugate scaled

inverse Chi-square distribution, i.e., *ν λχ−*2 (*ν* ). We suggest that the degrees of freedom is in the range *ν ∈* [3*,* 10] and we recommend the default choice of 3.

Now, *λ* is based on the estimate *σ*ˆ: generally, if *P < N* , then *yi ∼*N

(***x****t****β***ˆ*, σ*ˆ2 ;

otherwise, *σ*ˆ = *sy* . Solve for *λ* such that P (*σ*2 *≤ σ*ˆ2 1 = *q*. The suggested range

*i*

for the quantity *q ∈* [0*.*75*,* 0*.*99] where the recommended default choice is 0.9.

(*ν, q*) are potential candidate parameters for cross-validation.

Other important arguments for the BART prior. We fix the number of trees at *H* : the proposed default number of trees is 200 for continuous outcomes, but, as shown by Bleich et al. (2014), 50 is also a reasonable choice: cross-validation can be considered. The number of cutpoints is an implementation detail where the standard choice is 100.

**3 Posterior computation**

In order to generate samples from the posterior for *f* , we emply Markov Chain Monte Carlo (MCMC) (Hancock, 2014) for the structure of all the trees *Th* , for *h* = 1*, . . . , H* ; the values of all leaves *µhn* for *n ∈ Lh* within tree *h*; and (for continuous outcomes) the error variance *σ*2 .

The leaf and variance parameters are sampled from the posterior using Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990). Since the priors on these parameters are conjugate, the Gibbs conditionals are specified analyti- cally. For the leaves, each *µhn* is drawn from a Normal conditional density. The error variance, *σ*2 , is drawn from a scaled inverse Chi-square conditional.

Drawing a tree from the posterior requires a Metropolis-within-Gibbs sam- pling scheme (Mueller, 1991, 1993), i.e., a Metropolis-Hastings (MH) step (Metropo-

lis et al., 1953; Hastings, 1970) within Gibbs sampling. For single-tree models, four different proposal mechanisms are defined: the complementary BIRTH and DEATH along with CHANGE and SWAP (Chipman et al., 1998, 2013) (N.B. other MCMC tree sampling strategies have been proposed: Denison et al. (1998); Wu et al. (2007); Pratola (2016)). For the purposes of this discussion, we restrict our attention to the BIRTH and DEATH proposals each with equal probability. BIRTH selects a leaf and turns it into a branch, i.e., selects a new variable and cutpoint with two leaves “born” as its descendants. DEATH selects a branch leading to two terminal leaves and “kills” the branch by replacing it with a single leaf.

For illustration, we present the acceptance probability for a BIRTH proposal. The algorithm assumes a fixed discrete set of possible split values for each *xj* . Furthermore, the leaf values, *µhn* , are removed here (by integrating over them) so that our search in tree space is over a large, but discrete, set of possibilities. At the *m*th MCMC step, let *T m* denote the current state for the *h*th tree and *T ∗* denotes the proposed *h*th tree (subscript *h* is suppressed for convenience). *T ∗* are identical to *T m* except that one terminal leaf of *T m* is replaced by a branch of *T ∗* with two terminal leaves. The proposed tree is accepted with the following probability:

*π*BIRTH = min

r P [*T ∗* ] P [*T m |T ∗* ]

P [*T m* ] P [*T ∗ |T m* ]

1*,*

where P [*T m* ] and P [*T ∗*] are the posterior probabilities of *T m* and *T ∗* respec- tively, P [*T m |T ∗*] is the probability of proposing *T m* given current state *T ∗* (a DEATH) and P [*T ∗ |T m* ] is the probability of proposing *T ∗* given current state

*m*

*T* (a BIRTH).

First, we describe the likelihood contribution to the posterior. Let ***y****n* denote

the partition of ***y*** corresponding to the leaf node *n* given the tree *T* . Because the leaf values are a priori conditionally independent, we have [***y****|T* ] = Ti*n* [***y****n |T* ].

So, for the ratio P[*T ∗* ] after cancellation of terms in the numerator and denom-

P[ *m* ]

*T*

inator, we have the likelihood contribution:

P [***y***L *,* ***y***R *|T ∗*] = P [***y***L *|T ∗*] P [***y***R *|T ∗*]

P [***y***LR *|T m* ]

P [***y***LR *|T m* ]

where ***y***L is the partition corresponding to the newborn left leaf node; ***y***R , the

partition for the newborn right leaf node; and ***y***LR = r ***y***L

***y***R

. N.B. the terms in

the ratio are the predictive densities of a Normal mean with a known variance and a Normal prior for the mean.

Similarly, the terms that the prior contributes to the posterior ratio often cancel since there is only one “place” where the trees differ and the prior draws components independently at different “places” of the tree. Therefore, the prior

contribution to P[*T ∗* ] is

P[ *m* ]

*T*

*γ −γ* 2

P [*Bn* = 1] P [*Bl* = 0] P [*Br* = 0] *sj* = *α*(*t*(*n*) + 1)*−*

[1 *− α*(*t*(*n*) + 2)

] *sj*

P [*Bn* = 0]

1 *− α*(*t*(*n*) + 1)*−γ*

where P [*Bn* ] is the branch regularity prior (1), *sj* is the splitting variable selec- tion probability, *n* is the chosen leaf node in tree *T m* , *l* = 2*n* is the newborn left leaf node in tree *T ∗* and *r* = 2*n* + 1 is the newborn right leaf node in tree *T ∗*.

*m*

Finally, the ratio P[*T |T ∗*

] is

P[*T ∗ |T m* ]

P [DEATH*|T ∗*] P [*n|T ∗*]

P [BIRTH*|T m* ] P [*n|T m* ] *sj*

where P [*n|T* ] is the probability of choosing node *n* given tree *T* . *sj* appears in both the numerator and denominator of the acceptance probability *π*BIRTH , therefore, cancelling which is mathematically convenient.

**4 Dichotomous, categorical and survival outcomes with BART**

Dichotomous outcomes can be handled with either a probit or a logit link func- tion as in other Bayesian regression models. For probit, typically, the Albert and Chib (1993) technique is employed while for the logit link there are a few common alternatives (Holmes and Held, 2006; Fru¨hwirth-Schnatter and Fru¨h- wirth, 2010; Gramacy and Polson, 2012). There are several techniques for cat- egorical outcomes as well; see (Albert and Chib, 1993; McCulloch and Rossi,

1994; McCulloch et al., 2000; Fru¨hwirth-Schnatter and Fru¨hwirth, 2010; Kindo et al., 2016; Murray, 2017). For survival analysis, the discrete time approach (Fahrmeir, 2014) can be used to turn the problem into a dichotomous outcome with either the probit or logit link (Sparapani et al., 2016). And, similarly, recurrent events (Sparapani et al., 2018) and competing risks (Sparapani et al.,

2019) can be handled. Also, for survival analysis, if the accelerated failure time assumption holds, then you can pair BART with Dirichlet Process Mixtures (Dunson, 2017; Henderson et al., 2017).

**5 Recent developments and software implemen- tations for BART**

BART is still evolving as new variants emerge. Zhang et al. (2007) paired BART with a conditional autoregressive (CAR) model for spatial data. Se- quential BART is an adaptation to missing value imputation (Xu et al., 2016). Pratola et al. (2017) adapt BART to heteroscedastic data with two ensembles of trees: one that is BART and another that is multiplicative (rather than ad- ditive like BART). Hahn et al. (2017) have extended BART to causal inference. Linero and Yang (2018) present what they call SoftBART which creates smooth BART functions. George et al. (2018) marry BART with Dirichlet Processes to avoid the Normally distributed errors assumption. And, finally, there are some developments which are helpful in using BART with high perfomance comput- ing frameworks such as the Message Passing Interface (Walker and Dongarra,

1996; Gabriel et al., 2004): two such approaches are Consensus MCMC (Pratola et al., 2014) and Modified Likelihood Inflating sampling Algorithm (Entezari et al., 2018).

Since BART is essentially a computational approach, software is necessary for its use. As of this writing, several R package implementations of BART are available on the Comprehensive R Archive Network (CRAN): **BayesTree** (Chipman and McCulloch, 2016), **bartMachine** (Bleich et al., 2014), **dbarts** (Dorie et al., 2016) and **BART** (McCulloch et al., 2019).

**References**

J Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–79, 1993.

P Baldi and S Brunak. *Bioinformatics: the machine learning approach*. MIT Press, Cambridge, MA, 2001.

Justin Bleich, Adam Kapelner, Edward I George, and Shane T Jensen. Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, 8(3):1750–1781, 2014.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman. *Classification and regression trees*. Routledge, New York, NY,

2017.

H Chipman and RE McCulloch. *BayesTree: Bayesian Additive Regression Trees*,

2016. [[https://CRAN.R-project.org/package=BayesTree](https://CRAN.R-project.org/package%3DBayesTree)].

H A Chipman, E I George, and R E McCulloch. BART: Bayesian Additive

Regression Trees. *Annals of Applied Statistics*, 4:266–98, 2010.

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

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian regres- sion structure discovery. *Bayesian Theory and Applications,(Eds, P. Damien, P. Dellaportas, N. Polson, D. Stephens), Oxford University Press, USA*, 2013.

JA Cook and J Ranstam. Overfitting. *British Journal of Surgery*, 103(13):

1814–1814, 2016.

Nello Cristianini. Overfitting (overtraining). *Wiley StatsRef: Statistics Ref- erence Online*, 2014. doi: https://doi.org/10.1002/9780471650126.dob0513. pub2.

David GT Denison, Bani K Mallick, and Adrian FM Smith. A Bayesian CART

algorithm. *Biometrika*, 85(2):363–377, 1998.

V Dorie, H Chipman, and RE McCulloch. *dbarts: Discrete Bayesian Additive Regression Trees Sampler*, 2016. [[https://CRAN.R-project.org/package= dbarts](https://CRAN.R-project.org/package%3Ddbarts)].

David B. Dunson. Nonparametric Bayes. *Wiley StatsRef: Statistics Reference*

*Online*, 2017. doi: https://doi.org/10.1002/9781118445112.stat07905.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Reihaneh Entezari, Radu V Craiu, and Jeffrey S Rosenthal. Likelihood inflating sampling algorithm. *Canadian Journal of Statistics*, 46(1):147–175, 2018.

Ludwig Fahrmeir. Discrete survival-time models. *Wiley StatsRef: Statistics Ref- erence Online*, 2014. doi: https://doi.org/10.1002/9781118445112.stat06012.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on- line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

J H Friedman. Greedy function approximation: a gradient boosting machine.

*Annals of Statistics*, 29:1189–1232, 2001.

Sylvia Fru¨hwirth-Schnatter and Rudolf Fru¨hwirth. Data augmentation and MCMC for binary and multinomial logit models. In *Statistical modelling and regression structures*, pages 111–132. Springer, 2010.

Edgar Gabriel, Graham E Fagg, George Bosilca, Thara Angskun, Jack J Don- garra, Jeffrey M Squyres, Vishal Sahay, Prabhanjan Kambadur, Brian Bar- rett, Andrew Lumsdaine, Ralph Castain, David Daniel, Richard Graham, and Timothy Woodall. Open MPI: Goals, concept, and design of a next gener- ation MPI implementation. In *European Parallel Virtual Machine/Message Passing Interface Users’ Group Meeting*, pages 97–104. Springer, 2004.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calcu- lating marginal densities. *Journal of the American Statistical Association*, 85 (410):398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Edward George, Prakash Laud, Brent Logan, Robert McCulloch, and Rodney Sparapani. Fully nonparametric Bayesian Additive Regression Trees. *arXiv preprint arXiv:1807.00068*, 2018.

Robert B Gramacy and Nicholas G Polson. Simulation-based regularized logistic regression. *Bayesian Analysis*, 7(3):567–590, 2012.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heteroge- neous effects. *arXiv preprint arXiv:1706.09523*, 2017.

John M. Hancock. Markov Chain Monte Carlo (MCMC, Metropolis-Hastings, Gibbs Sampling). *Dictionary of Bioinformatics and Computational Biology*,

2014. doi: https://doi.org/10.1002/9780471650126.dob0973.

WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Nicholas C Henderson, Thomas A Louis, Gary L Rosner, and Ravi Varadhan. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *arXiv preprint arXiv:1706.06611*,

2017.

C Holmes and L Held. Bayesian auxiliary variable models for binary and multi- nomial regression. *Bayesian Analysis*, 1:145–68, 2006.

Bereket P Kindo, Hao Wang, and Edsel A Pen˜a. Multinomial probit Bayesian

Additive Regression Trees. *Stat*, 5(1):119–131, 2016.

A Krogh and P Solich. Statistical mechanics of ensemble learning. *Physical*

*Review E*, 55:811–25, 1997.

M Kuhn and K Johnson. *Applied Predictive Modeling*. Springer, New York, NY,

2013.

Michael LeBlanc. Regression trees. *Wiley StatsRef: Statistics Reference Online*,

2014. doi: https://doi.org/10.1002/9781118445112.stat07514.pub2.

A. Linero. Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–36,

2018.

Antonio R Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.

Wei-Yin Loh. Classification and regression tree methods. *Wiley StatsRef: Statis- tics Reference Online*, 2014. doi: https://doi.org/10.1002/9781118445112. stat03886.

RE McCulloch, RA Sparapani, R Gramacy, C Spanbauer, and M Pra- tola. *BART: Bayesian Additive Regression Trees*, 2019. [[https://CRAN. R-project.org/package=BART](https://CRAN.R-project.org/package%3DBART)].

Robert McCulloch and Peter E Rossi. An exact likelihood analysis of the multi- nomial probit model. *Journal of Econometrics*, 64(1-2):207–240, 1994.

Robert E McCulloch, Nicholas G Polson, and Peter E Rossi. A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, 99(1):173–193, 2000.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Au- gusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

P Mueller. A generic approach to posterior integration and Gibbs sam- pling. Technical Report 91-09, Purdue University, West Lafayette, Indi- ana, 1991. [[http://www.stat.purdue.edu/research/technical\_reports/ pdfs/1991/tr91-09.pdf](http://www.stat.purdue.edu/research/technical_reports/pdfs/1991/tr91-09.pdf)].

P Mueller. Alternatives to the Gibbs sampling scheme. Technical report, In- stitute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, 1993. [[http://citeseerx.ist.psu.edu/viewdoc/download?doi=](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.5613&amp;rep=rep1&amp;type=pdf)

[10.1.1.48.5613&rep=rep1&type=pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.5613&amp;rep=rep1&amp;type=pdf)].

Jared S Murray. Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:1701.01503*, 2017.

Matthew Pratola, Hugh Chipman, Edward George, and Robert McCulloch. Heteroscedastic BART using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*, 2017.

Matthew T Pratola. Efficient Metropolis–Hastings proposal mechanisms for

Bayesian regression tree models. *Bayesian analysis*, 11(3):885–911, 2016.

Matthew T Pratola, Hugh A Chipman, James R Gattiker, David M Higdon, Robert McCulloch, and William N Rust. Parallel Bayesian Additive Regres- sion Trees. *Journal of Computational and Graphical Statistics*, 23(3):830–52,

2014.

Brian D Ripley. *Pattern recognition and neural networks*. Cambridge University press, 2007.

R Sparapani, L Rein, S Tarima, T Jackson, and J Meurer. Non- parametric recurrent events analysis with BART and an applica- tion to the hospital admissions of patients with diabetes. *Bio- statistics*, (in press), 2018. doi: https://doi.org/10.1093/biostatistics/ kxy032. [[https://academic.oup.com/biostatistics/advance-article/ doi/10.1093/biostatistics/kxy032/5061112](https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy032/5061112)].

Rodney Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric competing risks analysis using Bayesian Additive Re- gression Trees (BART). *Statistical Methods in Medical Research*, (in press),

2019. doi: https://doi.org/10.1177/0962280218822140.

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in medicine*, 35(16):2741–53, 2016.

William N Venables and Brian D Ripley. *Modern applied statistics with S- PLUS*. Springer Science & Business Media, New York, 2013. [[https://cran. r-project.org/package=nnet](https://cran.r-project.org/package%3Dnnet)].

David W Walker and Jack J Dongarra. MPI: a standard message passing inter- face. *Supercomputer*, 12:56–68, 1996.

Yuhong Wu, H˚akon Tjelmeland, and Mike West. Bayesian CART: Prior spec- ification and posterior simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66, 2007.

Dandan Xu, Michael J Daniels, and Almut G Winterstein. Sequential BART

for imputation of missing covariates. *Biostatistics*, 17(3):589–602, 2016.

Heping Zhang, John Crowley, Harold C. Sox, and Richard A. Olshen Jr. Tree- structured statistical methods. *Wiley StatsRef: Statistics Reference Online*,

2014. doi: https://doi.org/10.1002/9781118445112.stat05678.

Song Zhang, Ya-Chen Tina Shih, and Peter Mu¨ller. A spatially-adjusted Bayesian Additive Regression Tree model to merge two datasets. *Bayesian Analysis*, 2(3):611–633, 2007.