# oBART : Ordinal Classification using Bayesian Additive Regression Trees

Hee Chul Jeong
Department of Statistics, Sungkyunkwan University

March 20, 2024

**Abstract**

Bayesian Additive Regression Trees (BART) has shown remarkable performances in various data settings for its flexibility. So far, BART can be used for regression problems, binary classification problems, multinomial classification, and many others. Moreover, BART has also been specifically modified for predicting ordinal categorical variables (Kindo, Wang, Pena 2013). Unlike usual categorical variables, ordinal categorical variables have distinctive features in the encoded orders so it is trivial that they require specific modifications for predictions, rather than applying the ordinary multi-class classification methods due to losses of the order information. In this paper, a modification of Kindo, Wang, Pena (2013) for ordinal classifiation is introduced. The modifications made in this paper may not look theoretically apparent, as it also uses the standard normal latent variable and the orderly cut-off points regarding to the latent variable to map the continuous outcomes of BART to ordinal categories. Nonetheless, by making those seemingly ambiguous adjustments, oBART shows remarkable improvements predictive performances.

## 1 Introduction

Classification, in terms of statistics, is designating each observation to the most probable or likely category among the available set of categories, using numerical information. There are mainly two types of classification: nominal categorical classification and ordinal categorical classification. Nominal categories are those of which the orders of numerical labels do not possess any statistical information or contextual meaning. Genders (man and woman), continents (Africa, North/South America, Asia, Europe, Oceania, Antarctica), and colors (red, blue, white, green, yellow, ...) can be examples of nominal categories. Ordinal categories, on the other hand, are those of which the orders of numerical labels convey some degree of information. Letter Grades (A, B, C, D, F), political beliefs (liberal, neutral, conservative), and sizes (small, medium, large) can be examples of ordinal categories. Since using the order information may play a critical role depending on the context, it is trivial nominal and ordinal categories need different approaches and algorithms in modeling.

In this paper, a new modeling approach for ordinal classification, oBART, is proposed. The first section and the subsections within the first section provides preliminary concepts for understanding oBART and explains the limitations of previous models. The second section explains the structures and algorithms for the proposed model. The next two sections conduct the simulation studies and real data applications and evaluate the results from them, and the last section concludes the study with the implications of oBART.

Classifying nominal categories has abundant options of predictive models available. Among the parametric models, one of the most commonly used models can be the multicategory logit regression, which was first introduced by Daniel McFadden in 1973. Let $\pi_j = P(Y = j)$, where $j = 1, ..., J$ implies the the $j^{th}$ category. Then, the structures of the multicategory logit regression is as follows.

$$log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \boldsymbol{\beta_j X}, j = 1, 2, ..., J-1 \tag{1}$$

Each $j^{th}$ set of parameters $(\alpha_j, \boldsymbol{\beta_j x})$ represents a different model so there are total of $J-1$ regression models. The category J is the baseline category, which can be an arbitrary one among all the available J categories as long as it is fixed for the entire algorithm and the interpretation. It becomes an ordinary logistics regression model when $J = 2$. As mentioned, it was first introduced in 1973 but it is still heavily used by researchers and analysts for its simple linear structure, which enables intuitive interpretations. However, as other linearly structured models, that linear structure is what hinders the multicategory logit regression from explaining, if there is any, complex relationships between the explanatory variable(s) and response variable.

As for the parametric models, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are common choices as well. LDA assumes that each predictor follows a normal distribution. Then, the $p$ predictors together form a multivariate Gaussian distribution, as it is shown below.

$$f(X_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu)\right) \tag{2}$$

Then, assuming each observation belongs to one of $K$ classes, where the $K$ different classes are differentiated by different $\mu_k$'s, each observation is designated to the class of which gives the following Bayes classifier the largest, out of $K$ classes, assuming the variance-covariance matrix $\Sigma$ remains the same for all $K$ categories.

$$\delta_k(X_i) = X_i^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + log(\pi_k), \tag{3}$$

where the true mean and variance-covariance matrix, $\mu_k$ and $\Sigma$, are replaced by $\bar{X}_k$ and $S_p$, the sample mean and pooled sample variance, in application and $\pi_k$

is replaced by the sample proportion of observation belonging to the $k^{th}$ category. Then, geometrically, several contours are formed in straight lines, as the name suggests, that differentiate one class to another.

QDA is an extension of LDA. As the name suggests, the distinguishing factor of QDA is that the contours are quadratic, and this can be derived by letting $\Sigma_k \neq \Sigma_l$ if $k \neq l$. This relaxed assumption gives the following Bayes classifier, and the remaining process stay the same as that of LDA.

$$\delta_k(X_i) = -\frac{1}{2}X_i^T\Sigma_k^{-1}X_i + X_i^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}log|\Sigma_k| + log(\pi_k) \tag{4}$$

Both the LDA and QDA can be powerful models, but they possess some critical drawbacks. First, they are inherently sensitive to outliers as the sample means and the sample variance-covariance matrix play critical roles in defining Bayes classifiers. Additionally, while the LDA has a weakness of limited flexibility, the QDA has a weakness of complex structures that could cause overfitting.

The last parametric model to be introduced is Naive Bayesian model. Naive Bayesian model assumes somewhat more constraining assumption than that of LDA and QDA, yet it leads to better results as the number of observations and explanatory variables increase in many cases. The core concept of Naive Bayes is the assumption that all $P$ explanatory variables are mutually independent, within each class, that is, given $k = 1, 2, ..., K$,

$$f_k(X_i) = f_{k,1}(X_{i,1}) \times f_{k,2}(X_{i,2}) \times \cdots f_{k,P}(X_{i,P}), \tag{5}$$

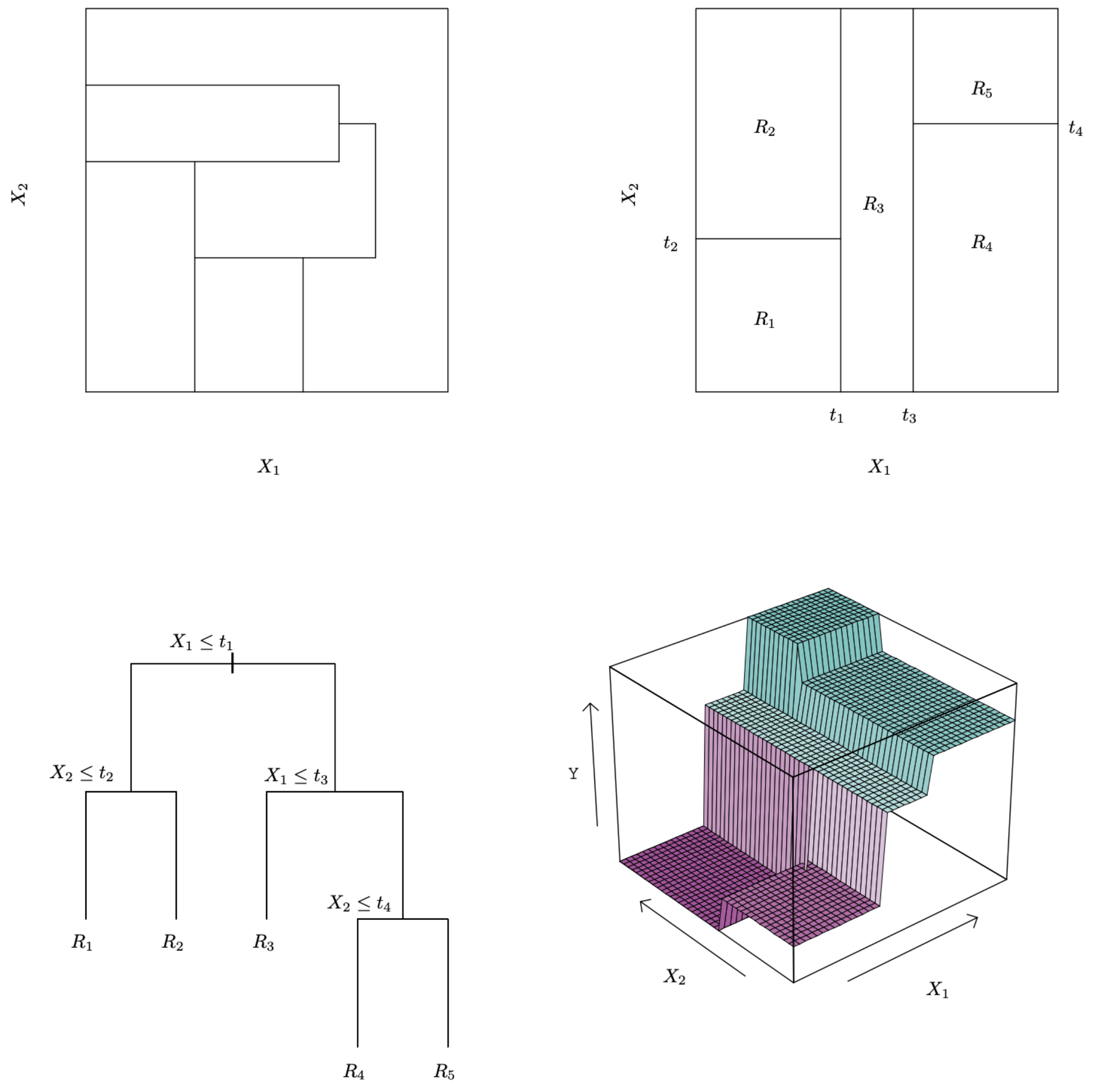
$$P(Y_i = k|X_i) = \frac{\pi_k \times f_{k,1}(X_{i,1}) \times f_{k,2}(X_{i,2}) \cdots \times f_{k,P}(X_{i,P})}{\sum_{l=1}^{K} \pi_l \times f_{k,1}(X_{i,1}) \times f_{k,1}(X_{i,1}) \times f_{k,2}(X_{i,2}) \cdots \times f_{k,P}(X_{i,P})} \tag{6}$$

where $f_{k,j}$ is the $k^{th}$ density function of the $j^{th}$ explanatory variable, and it gets trivial as for calculating the conditional probability of each $k^{th}$ category. It is not to mention that this strong assumption gets more unrealistic as $P$ increases. Nevertheless, it reduces the complexity of the model, the variance of the model, which is eventually greater than the bias of the model. Of course, this simplification can simultaneously be the strength and the drawback of the model as it is highly probable to give a poor performance if the true structure behind is intensively complex and interactive, and this is where more flexible approaches come in.

There are even greater number of choices when it comes to nonparametric approaches, and one of the most commonly used among them is the K-Nearest Neighbor (KNN). KNN was first introduced by Evelyn Fix and Joseph Hodges in 1951 and further developed by Thomas Cover in 1967. KNN, as the name suggests, uses a distance measure to classify, where the distance measure is usually the Euclidean Distance. It calculates all distances from one point to another and with the pre-defined value K, it uses the majority rule within the K nearest points around each point. With appropriate choice of K and the number of explanatory variables, KNN presents intuitive interpretations with a profound graphical representation. However, appropriately choosing K and the number of explanatory variables are what make KNN challenging.

One of the other most commonly used nonparametric models for classifying nominal categories are tree-based models. As Wei Yin L. states, modern tree-based models can be used to fit almost any type of variable (). Tree-based models are based on classification and regression trees(CART), which was first introduced by Breiman *et al* in 1984. CART splits the spaces of the explanatory variables into separate areas and print a constant value as a prediction whichever is designated to the same area. Let $R_t$ be the $t^{th}$ area and $C_t$ be the constant value given to $R_t$, which becomes a category in the case of classification. Then, the equation and graph below are typical example of CART.

$$f(\boldsymbol{X}) = \sum_{t=1}^{m} C_t \boldsymbol{I}(\boldsymbol{X} \in R_t) \tag{7}$$



The questions that follow then are which variable the model should choose as the splitting variable at each branch, which value of that splitting variable the model should choose as the splitting value, and which category the model should designate at each terminal node. At each branch, the model chooses $j \in \{1, ..., P\}$ and $s \in I\!R^{nxp}$ such that minimize misclassification rate, Gini index, and deviance, which are provided below respectively.

$$\text{Misclassification Rate} = \frac{1}{N_m} \sum_{i:x_i \in R_m} I\{y_i \neq k(m)\} = 1 - \hat{p}_{m,k(m)} \tag{8}$$

$$\text{Gini index} = \sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k}) \tag{9}$$

$$\text{Deviance} = -\sum_{k=1}^{K} \hat{p}_{m,k} log(\hat{p}_{m,k}) \tag{10}$$

On the equations above, each $R_m$, $N_m$, $\hat{p}_{m,k(m)}$, and $\hat{p}_{m,k}$ represents the region of the $m^{th}$ node, the number of observations $R_m$ has, the proportion of the majority category on $R_m$, and the proportion of $k^{th}$ category on $R_m$, respectively. There are Random Forests, XGBoost, LightGBM, and so many other powerful predictive tree-based models that were originated from CART, not to mention that there is a countless number of choices other than the ones mentioned in this paper for nominal classifications.

Classifying ordinal categories requires a different approach because the order of which the categories are labeled does contain information though applying ordinary multinomial classification methods to ordinal categories is not forbidden but is only not encouraged, whereas ordinal classification methods should not be applied to nominal categories.

The most common parametric models for ordinal classification are cumulative logit regression and cumulative probit regression. First of all, the cumulative logit regression, which is also called as the proportional odds model, was first introduced in 0000 by 00, which has the following structure. With $j = 1, ..., J - 1$,

$$log\left(\frac{P(Y \leq j)}{P(Y > J)}\right) = \alpha_j + \boldsymbol{\beta X}, j = 1, 2, ..., J - 1 \tag{11}$$

With the above cumulative logits, the probabilities that each observation belongs to a certain category is accessible. The structure of the cumulative logit regression resembles that of the multicategory logit regression in a way that it generates $J - 1$ different regressions to define each category's probability. Nevertheless, it is important to note that the rest of regression coefficients except for the $\alpha_j$'s stay the same for the entire $J - 1$ regressions. The reason behind using universal terms throughout $J - 1$ different regressions is the monotonic constraint.

The monotonic constraint refers to the assumption that the degree of effect that each individual predictor has on each category's probability is constant. The necessity of the monotonic constraint in the cumulative logit regression can be explained by the following instance. Suppose that there are $J$ categories and that there are $J-1$ different sets of $(\alpha_j, \boldsymbol{\beta_j x})$ like typical multicategory logit regression.

$$P(Y \leq j) = \frac{exp\,[\alpha_j + \boldsymbol{\beta_j X}]}{1 + exp\,[\alpha_j + \boldsymbol{\beta_j X}]} \tag{12}$$

$$P(Y \leq j+1) = \frac{exp\,[\alpha_{j+1} + \boldsymbol{\beta_{j+1} X}]}{1 + exp\,[\alpha_{j+1} + \boldsymbol{\beta_{j+1} X}]} \tag{13}$$

The above equations can be derived by the given assumptions, and by the non-decreasing(monotonic) property of the cumulative probability, the inequality $P(Y \leq j) \leq P(Y \leq j+1)$ should always hold. However, it is trivial that such inequality cannot hold in certain cases of $\beta_j$ and $\beta_{j+1}$, which is why the monotonic constraint is needed in the case of cumulative logit regression. This property also holds in the case of the cumulative probit regression, where the link function and the respective interpretation are the only differences from the cumulative logit regression.

In the case of using nonparametric method for ordinal classifications, the options available are severely out-numbered compared to those of nominal classifications. It is not rare to see that some researchers and analysts apply usual multinomial classification methods for ordinal classification tasks; it might not be wrong, as it is previously stated, but it is certainly not the best choice. Within those limited options, however, Ordinal Forest (Hornung 2019) might be a wise choice. Ordinal Forests (OF) is an ordinal classification variant of random forests.

(Explaning about Ordinal Forests).

# 2 Bayesian Additive Regression Trees

BART was first introduced in 2010 by Chipman, George, and McCulluch (CGM10) for estimating nonparametric function with regression trees. BART is so flexibly structured that it can reflect any, simple or complicated, relationships and interactions within the predictor space into a series of hyperrectagles.

BART has the following structural assumptions. Let $Y \in \mathrm{I\!R}^{\mathrm{n}}$ be the continuous response variable and $X = (x_1, x_2, ..., x_p) \in \mathrm{I\!R}^{\mathrm{nxp}}$ be the explanatory variables. Suppose there exists an unknown function $f$ of which explains the relation between the response variable $Y$ and the explanatory variables $X$, such that

$$Y = f(X) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \tag{14}$$

and it is $f$ that BART is aimed to approximate through the summation of $m$ trees, that is, $f(X_i) \approx h(X_i) = \sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M}_t) + \varepsilon_i$, $i = 1, 2, ..., n$, where $n$ is the total number of observations, $\mathcal{T}_t$ is the structure of $t^{th}$ tree including its depth, splitting variable(s) and splitting values, and $\mathcal{M}_t$ are the terminal nodes of the $t^{th}$ tree.

The individual tree structure $\mathcal{T}_t$ contains all splitting rules, which refers to splitting variables and splitting values. A splitting variable is a variable used to branch out at a depth $d$, and a splitting value is a specific value of the splitting variable for the division of the two regions at a depth $d$.

$\mathcal{M}_t$ is the set of terminal nodes for $t^{th}$ tree structure, where terminal nodes are the ones used as a part of the prediction directly. It is mentioned previously that BART's estimate is composed of the summation of $m$ trees, and it is the value of each tree's terminal node that the summation of $m$ trees implies.

## 2.1 Priors for BART

As BART is a Bayesian predictive model, priors and the corresponding posteriors need to be specified. There are mainly 3 different types of parameters that BART aims to optimize for approximation: the tree structures ($\mathcal{T}_t$), terminal nodes ($\mathcal{M}_t$), and variance ($\sigma^2$). With those priors, it is aimed to regularize the fit of individual trees being overly influential and further assumed that each $t^{th}$ tree component ($\mathcal{T}_t$, $\mathcal{M}_t$) and terminal nodes within the same tree are independent of the others and the variance ($\sigma^2$) ("BART : Bayesian Additive Regression Trees" page 5). The structures of the priors explained are provided below.

$$p((\mathcal{T}_1, \mathcal{M}_1), ..., (\mathcal{T}_m, \mathcal{M}_m), \sigma^2) = \left[\prod_t^m p(\mathcal{T}_t, \mathcal{M}_t)\right] p(\sigma^2) = \left[\prod_t^m p(\mathcal{M}_t|\mathcal{T}_t)p(\mathcal{T}_t)\right] p(\sigma^2) \tag{15}$$

$$p(\mathcal{M}_t|\mathcal{T}_t) = \prod_{r_t=1}^{R_t} p(\mu_{r_t}|\mathcal{T}_t) \text{ , where } \mu_{r_t} \in \mathcal{M}_t. \tag{16}$$

The prior for the tree structures ($\mathcal{T}_t$) is composed of 3 parts: the depth $d$, splitting variable(s), and splitting values. The probability of a tree branching out at the depth $d$ ($= 0, 1, 2, ...$) is defined as follows,

$$\alpha(1 + d)^\beta \text{ , where } \alpha \in (0, 1), \beta \in [0, \infty). \tag{17}$$

The above probability is called the splitting probability, and although the values of $\alpha$ and $\beta$ can be chosen arbitrarily, the default choices are 0.95 and 2, respectively. For the splitting variable(s) and values, the uniform prior on available variables and the uniform prior on the discrete set of available splitting values are used.

As the **Equation 9** suggests, $\varepsilon$ follows a normal distribution. For this reason, $N(\mu_\mu, \sigma_\mu^2)$ is used to meet the conjugacy, for the conditional prior ($\mu_{r_t}|\mathcal{T}_t$). Then, the question naturally leads to the choice of values of $\mu_\mu$ and $\sigma_\mu^2$. It is highly likely that $E(Y|\mathbf{X})$, the sum of $m$ $\mu_{ij}$'s under the sum of trees, is between the observed minimum of $Y$, $y_{min}$, and the observed maximum of $Y$, $y_{max}$. Then, by choosing $\mu_\mu$ and $\sigma_{mu}^2$ such that $m\mu_\mu - k\sqrt{m}\sigma_\mu = y_{min}$ and $m\mu_\mu + k\sqrt{m}\sigma_\mu = y_{max}$ for some predefined value $k$, we may set the probability of $E(Y|\mathbf{X})$ forms within the interval $(y_{min},\ y_{max})$. This strategy allows BART to cover the region of $Y$ substantially without overconcentration and overdispersion.

For the prior of variance $(\sigma^2)$, inverse chi-square distribution$(\sigma^2 \sim \nu\lambda/\chi^2_\nu)$ is used for the purpose of conjugacy as well. Here, $\lambda$ is a data-oriented parameter that is set where would have a substantially high prior probability mass under the root mean squared errors(RMSE) from least squares regression. In addition, the prior prevents overfitting by limiting the probability mass put upon small values of $\sigma^2$. The value of $\lambda$ can be determined by the choice of $q$ such that $p(\sigma < \hat{\sigma}) = q$ holds, and for the choice of $\nu$, picking a value between 3 and 10 is recommended, and the default setting of $(\nu$,$q)$ is (3, 0.90).

## 2.2 MCMC Algorithm for BART

With the priors and the structures mentioned previously, the posterior samples of $\mathcal{T}_t$, $\mathcal{M}_t$, and $\sigma^2$ can be drawn. It has already been implied that the posterior distribution of $\mathcal{M}_t$ and $\sigma^2$ can be of closed forms since their conjugate priors, the normal distribution and the inverse chi-squared distribution respectively, are used. $\mathcal{T}_t$'s posterior distribution, on the other hand, cannot be derived in a closed form, and this is why Metropolis-within-Gibbs sampler(Geman and Geman 1984; Hastings 1970) is employed. To add with, Bayesian Backfitting (Hastie and Tibshirani 2000) is used to define the full conditionals for $\mathcal{T}_t$ and $\mathcal{M}_t$, which is that the $j^{th}$ tree is fit iteratively while keeping the other $m-1$ trees constant but only using their information as the residuals. That is,

$$\mathcal{R}_{-j} := \mathbf{y} - \sum_{t \neq j} \mathcal{T}_t^{\mathcal{M}} \tag{18}$$

and

$$\begin{aligned} 1 &: \mathcal{T}_1|\mathcal{R}_{-1}, \sigma^2 \\ 2 &: \mathcal{M}_1|\mathcal{T}_1, \mathcal{R}_{-1}, \sigma^2 \\ 3 &: \mathcal{T}_2|\mathcal{R}_{-2}, \sigma^2 \\ 4 &: \mathcal{M}_2|\mathcal{T}_2, \mathcal{R}_{-2}, \sigma^2 \\ &\vdots \\ 2m-1 &: \mathcal{T}_m|\mathcal{R}_{-m}, \sigma^2 \\ 2m &: \mathcal{M}_m|\mathcal{T}_m, \mathcal{R}_{-m}, \sigma^2 \\ 2m+1 &: \sigma^2|\mathcal{T}_1, \mathcal{M}_1, ..., \mathcal{T}_m, \mathcal{M}_m, \boldsymbol{\varepsilon} \end{aligned} \tag{19}$$

For every tree structure's update, a Metropolis-Hastings algorithm is utilized, where the detailed explanations for the updating process is given by **Adam K. et el**. $\mathcal{R}_{-j}$ refers to the $j^{th}$ partial residual, meaning it is the difference between the observation and the sum of all trees except the $j^{th}$ tree. This strategy allows not only to simplify the full conditionals, but also to ensure that each tree or terminal node is not overly effective, which eventually prevents overfitting.

## 2.3 Binary Classification using BART

Up to now, the implementation of BART is available only if the variable of interest is continuous, having the support of $(-\infty, \infty)$. BART can easily be extended to the binary classification and multinomial classification with some modifications. Both the binary and multinomial classification inherently use the same strategy; the difference is that the multinomial classification uses the strategy multiple times. Thus, before covering the multinomial modification of BART, this section explains the methods for binary classification using BART.

### 2.3.1 Data Augmentation

Suppose the target variable $Y$ is binary, where $Y \in \{0, 1\}$. It is important to note again that BART is composed of regression trees, not classification trees, and this makes BART not suitable to be applied with the common classification approaches of tree-based predictive models. For this reason, the data augmentation technique (Albert and Chib. 1993).

Suppose there is a latent variable $Z$ such that $Z \sim N(0, 1)$, and it is $Z$ that now BART is fit to approximate, that is,

$$Z_i \approx \sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M}_t) \tag{20}$$

With the generated latent variable having the structure above, the procedures explained through **Section 1.3.1** and **Section 1.3.2** are applied as well, except that the specification of prior and posterior sampling strategy for $\sigma^2$ are not needed as it is fixed at 1. In addition, as for the Gibbs sampler where draws of $Z|y$ are obtained, it is done as follows.

$$\begin{aligned} Z_i|Y_i = 1 \quad &\sim \quad \max\left[N\left(\sum_{t=1}^{m} g(\mathbf{X}_i, ; \mathcal{T}_t, \mathcal{M}_t), 1\right), 0\right] \\ Z_i|Y_i = 0 \quad &\sim \quad \min\left[N\left(\sum_{t=1}^{m} g(\mathbf{X}_i, ; \mathcal{T}_t, \mathcal{M}_t), 1\right), 0\right] \end{aligned}$$

Thus, for the structures and the procedures explained so far, the probit link function may be used to calculate the conditional probability of $Y_i = 1$ given $\mathbf{X_i}$.

$$p(Y_i = 1|\mathbf{X_i}) \approx \Phi\left[\sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M}_t)\right]$$

Each observation is predicted to be 0 if the conditional probability is under 0.5, or, in other words, $Z_i$ is under 0.

## 2.4 Multinomial Classification using BART

As it is briefly mentioned, the multinomial extension of BART can be accomplished by simply applying multiple latent variables and "one-versus-all" strategy, using that latent variable. Currently, there are some variations when it comes to multicategorical classification using BART, but they all follow the above structure in the big picture; this paper covers the four variations of it.

### 2.4.1 U-MBACT : Unordered-Multiclass Bayesian Additive Classification Trees

U-MBACT was introduced in **Kindo, Wang, Pe (2013)**, and it is the very first extension of BART to multinomial classification. Suppose a target variable $Y_i$ such that $Y_i = \{1, 2, ..., K\}$, where the labels do not imply any ordering, and $p-$dimensional explanatory variable $X_i \in IR^{nxp}$. Then, generate a latent vector $Z_i$ with $K - 1$ elements, that is,

$$z_{i,k} \sim N_k(0, 1), \text{ for } k = 1, 2, ..., K - 1, \tag{21}$$

where each $z_{i,k}$ implies the $k^{th}$ element of the latent vector $Z_i$. Then, the binary classification explained previously is operated for each $Z_{i,k}$; the prior setting and the posterior sampling strategies stay the same, except that it operates for $K - 1$ different latent variables using "one-versus-all" technique. Lastly, the classifying procedure becomes,

$$Y_i = \begin{cases} 1, & \text{if } z_{i,k} \leq 0 \text{ for all } k \\ k+1, & \text{if } (z_{i,k} = \max_l z_{i,l}) \wedge (z_{i,k} > 0), \end{cases} \tag{22}$$

as it is explained in **Kindo, Wang, Pe (2013)**. Multinomial Probit BART (MPBART), introduced in **Kindo, Wang, Pe (2016)**, follows exactly the same format except that it classifies each observation as,

$$Y_i = \begin{cases} k, & \text{if } \max Z_i = z_{i,k} > 0 \\ K, & \text{if } \max Z_i < 0. \end{cases} \tag{23}$$

### 2.4.2 Multinomial BART and Conditional Probability

In the paper by **Sparapani R., Spanbauer C., McCulloch R. 2021**, there are two approaches proposed for multinomial classification. The first of those two resembles the structure of continuation-ratio logits(Agresti 2003) and makes use of conditional probabilities. First, let the mutually exclusive binary indicator $Y_{i,1}, ..., Y_{i,K}$ represents $K$ different categories and $p_{i,j}$ be the conditional probability defined as follows.

$$
\begin{aligned}
p_{i,1} &= P[Y_{i,1} = 1] \\
p_{i,2} &= P[Y_{i,2} = 1 | Y_{i,1} = 0] \\
p_{i,3} &= P[Y_{i,3} = 1 | Y_{i,2} = 0, Y_{i,1} = 0] \\
&\vdots \\
p_{i,K-1} &= P[Y_{i,K-1} = 1 | Y_{i,K-2} = \cdots = Y_{i,1} = 0] \\
p_{i,K} &= P[Y_{i,K-1} = 0 | Y_{i,K-2} = \cdots = Y_{i,1} = 0]
\end{aligned}
$$

, where $p_{i,K} = 1 - p_{i,K-1}$. Then, the marginal probability of an arbitrary observation belonging to a certain category can be computed using the conditional probabilities defined above.

$$
\begin{aligned}
\pi_{i,1} &= P[Y_{i,1} = 1] = p_{i,1} \\
\pi_{i,2} &= P[Y_{i,2} = 1] = p_{i,2}q_{i,1} \\
\pi_{i,3} &= P[Y_{i,3} = 1] = p_{i,3}q_{i,2}q_{i,1} \\
&\vdots \\
\pi_{i,K-1} &= P[Y_{i,K-1} = 1] = p_{i,K-1}q_{i,K-2} \cdots q_{i,1} \\
\pi_{i,K} &= P[Y_{i,K} = 1] = q_{i,K-1}q_{i,K-2} \cdots q_{i,1}
\end{aligned}
$$

Now, let $S_1 = \{1, 2, ..., N\}$ and $S_j = \{i : Y_{i,1} = \cdots = Y_{i,j-1}\}$ for $j = 2, ..., K - 1$. Then, the fitting procedure of Multinomial BART is as follows.

$$
\begin{aligned}
Y_{i,j} &\sim B(p_{i,j}), \text{ where } i \in S_j and j = 1, 2, ..., K - 1 \\
p_{i,j} &= \Phi(\mu_j + f_j(X_i)) \\
f_j(X_i) &\sim \text{ BART},
\end{aligned}
$$

where $\mu_j = \Phi^{-1}\left[\frac{\sum_i Y_{i,j}}{\sum_i I(i \in S_j)}\right]$. The instance introduced for the explanations so far uses probit link for easing the computation, but certainly the logit link may also be applied to the procedures explained above, giving up the computational efficiency.

### 2.4.3 Multinomial BART and Logit Transformation

While the ”Multinomial BART and Conditional Probability” is not theoretically limited to a certain link function, the method introduced now is specifically made for logit link. According to **Sparapani R., Spanbauer C., McCulloch R. 2021**, the probit link can also be applied but ”would appear to contradict the development of” logit transformation approach.

The approach basically resembles the framework of the multi-categorical logit regression, where the differences are that it uses BART on the optimizations and that it fits $K$ models instead of $K - 1$ models.

$$P[Y_i = j] = \frac{exp\left[(\mu_j) + f_j(X_i)\right]}{\sum_{j'=1}^{K} exp\left[(\mu_{j'}) + f_{j'}(X_i)\right]} = \pi_{i,j}$$

$$\text{, where } f_j(X_i) \sim \text{ BART, } j = 1, ..., K$$

The $\mu_j$'s represent the constants for quasi-centering the probabilities $P[Y_i = j]$'s. Then, the question of identifiability naturally arises since $\pi_{i,j} = \frac{exp[(\mu_j)+f_j(X_i)]}{\sum_{j'=1}^{K} exp\left[(\mu_{j'})+f_{j'}(X_i)\right]} = \frac{exp[(\mu_j)+f_j(X_i)+c]}{\sum_{j'=1}^{K} exp\left[(\mu_{j'})+f_{j'}(X_i)+c\right]}$. However, this is not necessarily the center of interest and the core concept in achieving an above-standard performance since $\pi_{i,j}$ is identified anyways. After calculating those probabilities for each observation, the prediction method, which is now the remaining procedure, is the same as all the others.

## 2.5 O-MBACT : Ordered-Multiclass Bayesian Additive Classification Trees

Despite their predictive performances under various data settings, the above 4 uses of BART for performing multinomial classifications have 2 major limitations. First, they all utilize "one-versus-all" strategy to differentiate one category to another, which seems reasonable as the encoded orders of the categories do not convey any source of information. However, this strategy is not suitable for predicting ordinal categorical variables as it does not incorporate orderly information, which leads to a low precision. Second, as the number of categories increases, the amount of computation required increases linearly. Suppose there are $K$ ordinal categories, then the 3 variations of BART mentioned previously would need to grow $K$ or $K-1$ sets of $m$ trees for predictions. If growing those sets of $m$ trees give extraordinary predictive performance, then the computational power would not be expensive to endure. However, as it is noted on the first limitation, this strategy does not guarantee a promising result, but only high expanses on computation. Those limitations had inspired O-MBACT by **Kindo, Wang, Pe (2013)**, and the primary ideas of O-MBACT are the use of standard normal latent variable and cut-off points.

Let $\eta_{i,k} = \sum_{j=1}^{k} = p_j(X_i)$ be the $k^{th}$ cumulative probability. The probability will be identified using the probit link under the following structure.

$$\eta_{i,k} = \Phi\left(\gamma_k - \sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M}_t)\right) \text{ for, } k = 1, 2, ..., K-1, \tag{24}$$

where the $g(\cdot)$ represents the individual tree of BART and $\gamma_k$ represents the $k^{th}$ cut-off points defined upon the standard normal distribution. It is the cumulative probability that the latent variable and BART aim to optimize, and it is achieved through the following posterior sampling process.

The sampling concept for BART and the latent variable remain to be the same as those in preceding chapter. Suppose BART and the latent variable is updated, then by setting the uninformative uniform priors on the cut-off points

$\gamma_k$'s, the posterior sampling can be done with the uniform distribution with the lower bound $a$ and upper bound $b$, which are defined as

$$b = \min\left[\min_i \left[Z_i : Y_i = k+1\right], \gamma_{k+1}\right]$$

$$a = \max\left[\max_i \left[Z_i : Y_i = k\right], \gamma_{k-1}\right]$$

where $\gamma_0 = -\infty$ and $\gamma_K = \infty$. In addition, one of the cut-off points from $\gamma_1$ to $\gamma_{K-1}$ must be fixed to achieve the identifiability on the process of posterior sampling, which is set to be $\gamma_1 = 0$ in the paper. Then, with the cut-off points drawn by the standard above, the cumulative probabilities of categories are computed using the cut-off points and BART's prediction of latent variable, as shown in **equation 24**. The remaining work for each observation is then obviously calculating the probabilities of categories and using the category with the highest probability as the prediction.

$$\hat{Y}_i = argmax_k \; p_k(X_i) = \eta_{i,k} - \eta_{i,k-1}$$

[each $Y_i$ is categorized based on which cut-off points the value of its corresponding value of the standard normal latent variable $Z_i$ lies between.]

While O-MBACT proposed an intriguing breakthrough of applying ordinal classification using BART, there seems to be room for improvement. First of all, the purpose and the use of latent variable is primarily focused on deriving categorical probabilities, and this perspective may be persuasive since this is the perspective that many methods for ordinal classification models take; however, this may not always be the case. Consider the different BART modifications preceded. The central use of latent variable $Z_i$ or $Z_{i,k}$'s was to compare it with the only cut-off point, 0, or with the rest of $Z_{i,k'}$'s ($k' \neq k$). In consequence, this perspective of which BART is sought to be optimized for the latent variable, which does directly affects both the cut-off points and the BART predictions, may cause uncertainty. Secondly, it is stated that $\gamma_1 = 0$ is set for the entire MCMC procedure for the identifiability. Nevertheless, although this deliberate fixation of a cut-off point may help to attain the identifiability, it certainly is not the most ideal remedy since this directly affects the entire MCMC procedure and changes the samples drawn from different posterior distributions.

# 3 Proposed Model : oBART

The proposing model of the paper, oBART, prevents the 2 possible concerns of O-MBACT. First, the latent variable $Z_i$ similarly, the latent variable is segmented by $K-1$ cut-off points similarly, but it is directly used for classifying each observation rather than being the middle passage to the classification. Second, it does not deliberately fix one or more cut-off points thereby affects the entire MCMC chains. It is not that the identifiability is not important. It is rather suggesting that it can be indirectly achieved through an educated and reasonable choice of initial values. Before proceeding, the following notations must be familiar. Let $H$ be the total number of MCMC iterations, $j$ be the index of each MCMC iteration, and $t$ be the index of all $m$ trees.

## 3.1 Structure of oBART

As BART uses the probit link to bisect two categories by 0, the only and fixed cut-off point in this case, oBART as well uses the probit link, except that it is divided into multiple segments. To be specific, the latent variable $Z$ of oBART follows truncated normal distributions,

$$Z_i = \begin{cases} N(0,1)I(\gamma_{K-1} \leq Z_i), & \text{for } Y_i = K \\ N(0,1)I(\gamma_{K-2} \leq Z_i < \gamma_{K-1}), & \text{for } Y_i = K-1 \\ \vdots & \\ N(0,1)I(\gamma_1 \leq Z_i < \gamma_2), & \text{for } Y_i = 2 \\ N(0,1)I(Z_i < \gamma_1), & \text{for } Y_i = 1 \end{cases} \tag{25}$$

and with this structure, the probability of the $i^{th}$ observation belonging to the $k^{th}$ category can be defined as follows.

$$\begin{gathered} P(Z_i < \gamma_1) = P(Y_i = 1) \\ P(\gamma_1 \leq Z_i < \gamma_2) = P(Y_i = 2) \\ \vdots \\ P(\gamma_{K-2} \leq Z_i < \gamma_{K-1}) = P(Y_i = K-1) \\ P(\gamma_{K-1} \leq Z_i) = P(Y_i = K) \end{gathered}$$

As a result, the augmented joint density function $Y$ and $Z$ is

$$P(Y, Z|\gamma_1, \gamma_2, ..., \gamma_{K-1}) = \prod_{i=1}^{n} \phi(Z_i) \left[I(Z_i < \gamma_1)\right]^{I(Y_i=1)} \left[I(\gamma_1 \leq Z_i < \gamma_2)\right]^{I(Y_i=2)} \cdots \left[I(\gamma_{K-1} \leq Z_i)\right]^{I(Y_i=K)} \tag{26}$$

Despite the intricacy of the appearance, the set of cut-off points $\{\gamma_k : k = 1, 2, ..., K-1\}$ is the only hyperparameter in the equation above and is the only parameter used to make predictions, which naturally leads to the concern of properly estimating them since it directly affects the performance of oBART.

In most cases, no prior information to the cut-off points would be available or accessible except the fact that they are parts of a normal distribution and their orders ($\gamma_1 < \gamma_2 < \cdots < \gamma_{K-1}$). For this reason, a truncated multivariate standard normal prior is used. Specifically,

$$\boldsymbol{\gamma} \sim N_{K-1}(0, \boldsymbol{I}_{K-1}) I\left[-\infty < \gamma_1 < \cdots < \gamma_{K-1} < \infty\right] \tag{27}$$

With this prior and the joint likelihood defined in **(7)**, the following posterior is derived.

$$P(\boldsymbol{\gamma}|Y, Z) \propto P(Y, Z|\boldsymbol{\gamma}) P(\boldsymbol{\gamma})$$
$$\sim N_{K-1}(0, \boldsymbol{I}_{K-1}) \prod_{k=1}^{K-1} I\left[\max_{i:Y_i=k} Z_i < \gamma_k \leq \min_{i:Y_i=k+1} Z_i\right]$$

Since the derived posterior is still a truncated multivariate normal, we may use a block-Gibbs sampler.

The key components of oBART, as well as O-MBACT, are the cut-off points. Though the induced posterior distribution of each cut-off point may seem reasonable, the posterior samples drawn via the above structure cannot converge since the upper and lower bound of each cut-off point will only get narrower as the number of observations increases, which would not allow the sampler to fully traverse the parameter space(paper reference). This lack of mixing hence diverging cut-off points was once a concern so the modification of (keun baik lee's paper) was applied to resolve such issue but it only succeeded in making the sampler explore profoundly, leaving the convergence and performance also not desirable. (**might not be used :** It is shown later in the simulation studies and the real data application(s) that the predictive performances are still remarkable even though the cut-off points have not converged.)

## 3.2 Algorithm

With the implementation of the ideas provided above, this subsection provides a detailed process of each iteration of the MCMC algorithm. The initial settings of the tree structures are equivalent to those of CGM10; the base parameter $\alpha$ and the power parameter $\beta$ are set as 0.95 and 2, respectively. The probability of grow, prune, and change are 0.28, 0.28, and 0.44, respectively.

The procedure is provided below.

**Step1.** Setting the initial values

Theoretically, the choice of initial values do not affect the convergence of the MCMC samples, as long as the order ($\gamma_1 < \gamma_2 < \cdots < \gamma_{K-1}$) is maintained. However, for the sake of computation, it is better to give rather informative or reasonable initial values. Let $\gamma_k^{(0)}$'s be the initial values of the cut-off points $\gamma_k$'s. Then, $\gamma_k^{(0)}$'s can be given by calculating the number of observations for each ordinal category and drawing the quantile value of the cumulative proportion of each category from the standard normal distribution.

$$\gamma_k^{(0)} = \Phi^{-1}(\tau_k), \text{ where } \tau_k = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i \le k\}, \text{ for } k = 1, 2, ..., K-1 \quad (28)$$

**Step2.** Setting the priors of BART

The prior settings for the tree structures ($\mathcal{T}$) and the mean parameters ($\mathcal{M}$) are the same as CGM10, except that the mean parameters are of the latent variables.

**Step3.** Updating $Z_i^{(j)}$'s

Each $Z_i^{(j)}$ needs to be updated according to the procedure explained in Section 2.1.1. In the case of $j = 1$, the values of cut-off points used for the updates are the initial values set in **Step1**. For $j > 1$, the values of cut-off points used for the updates will be those of $j - 1^{th}$.

$$Z_i = \begin{cases} N(\sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t^{(j-1)}, \mathcal{M}_t^{(j-1)}), 1) I(\gamma_{K-1}^{(j-1)} \le Z_i), & \text{for } Y_i = K \\ N(\sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t^{(j-1)}, \mathcal{M}_t^{(j-1)}), 1) I(\gamma_{K-2}^{(j-1)} \le Z_i < \gamma_{K-1}^{(j-1)}), & \text{for } Y_i = K-1 \\ \vdots \\ N(\sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t^{(j-1)}, \mathcal{M}_t^{(j-1)}), 1) I(\gamma_1^{(j-1)} \le Z_i < \gamma_2^{(j-1)}), & \text{for } Y_i = 2 \\ N(\sum_{t=1}^{m} g(\mathbf{X}_i; \mathcal{T}_t^{(j-1)}, \mathcal{M}_t^{(j-1)}), 1) I(Z_i < \gamma_1^{(j-1)}), & \text{for } Y_i = 1 \end{cases} \quad (29)$$

**Step4.** Updating $L_k^{(j)}$'s and $U_k^{(j)}$'s

At each iteration, the values of the latent variable used for the updates will be those of $j^{th}$.

$$L_k^{(j)} = \max_{i:Y_i=k}(Z_i^{(j)}) \quad \text{and} \quad U_k^{(j)} = \min_{i:Y_i=k+1}(Z_i^{(j)}) \quad (30)$$

**Step5.** Updating $\gamma_k^{(j)}$'s

Each $\gamma_k^{(j)}$ needs to be updated according to the procedure explained in Section 2.1.2. For every iteration, the values of lower bounds and upper bounds used for the updates will be those of $(j)$-th, updated in **Step5** and**Step6**.

$$\boldsymbol{\gamma}^{(j)} \sim N_{K-1}(0, I_{K-1}) \prod_{k=1}^{K-1} I\left[L_k^{(j)} < \gamma^{(j)} \le U_k^{(j)}\right], \text{ for } k = 1, 2, ..., K-1 \quad (31)$$

**Step6.** Updating $\mathcal{T}_t$'s and $\mathcal{M}_t$'s

The updating process of $\mathcal{T}_t$'s and $\mathcal{M}_t$'s are the same as CGM10, except that they are updated for the latent variable.

**Step7.** Repeating the iterations

Repeat the process **Step3** to **Step6** for the entire MCMC iterations.

**Step9.** Making predictions

With $S$ posteriors samples, use the MAP estimator for predictions.

$$\hat{Y}_i^{MAP} = \text{argmax}_{k \in \{1,...,K\}} \left[ \sum_{j=1}^{S} I\{\hat{Y}_i^{(j)} = k\} \right]$$

# 4 Applications

Prior to comparing performances of different models, it is vital to specify the metrics for comparing and evaluating performances. As it is stated, ordinal classification is unique and differentiated from classic classification for its ordered nature so it needs metrics that can reflect not only the precision, but also the degree of the precision a model has. That is, measuring how far a model predicts is just as important as measuring how well a model predicts.

The most classic metric when it comes to classification is accuracy, and because measuring corrections is the very first priority of any classification, it is needless to emphasize the need of accuracy as one of the metrics. The accuracy is defined as below, and as anyone could notice, the higher accuracy a model has, the better. Let $n_{test}$ represents the number of test observations, and $\hat{Y}_b$ and $Y_b$ represents the predicted category and the actual category of the $b^{th}$ test observation. Then,

$$\frac{1}{n_{test}} \sum_{b=1}^{n_{test}} I\{\hat{Y}_b = Y_b\}. \tag{32}$$

In certain cases like ordinal classification, the order of categories does represent how far a certain category is from another so it is trivial that the metrics that could measure how far or correlated each prediction is from each actual category. Firstly, the Mean Absolute Deviation (MAD) is used, which is defined as,

$$\frac{1}{n_{test}} \sum_{b=1}^{n_{test}} |\hat{Y}_b - Y_b|, \tag{33}$$

As consistently mentioned, the key idea is that the cost of a prediction error is higher as the prediction is further from the true category. This metric is commonly used in ordinal classification, and it is the metric used from the simulation and real data studies by **Kindo, Wang, Pe (2013)**. Secondly, the last metric used is the Quadratic Weighted Kappa. It is also one of the most commonly used metric for ordinal classification since it also considers the closeness of each prediction and gives credits as it gets closer; the equation provided below illustrates this trait.

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \tag{34}$$

where the weight matrix $w_{i,j}$ is a $K \times K$ matrix such that $w_{i,j} = \frac{(i-j)^2}{(K-1)^2}$, $O_{i,j}$ represents the number of observed agreement on $i^{th}$ and $j^{th}$ category, and $E_{i,j}$ represents the expected agreement on $i^{th}$ and $j^{th}$ category by chance, that is,

$$E_{i,j} = \frac{\text{row sum of the } i^{th}\text{category} \times \text{column sum of the } j^{th}\text{category}}{\text{total number of categories}}$$

.

The models being compared throughout **Section 4** are Ordinal Forests(OF), Ordered-Multiclass Bayesian Additive Classification Trees(OMBACT), and ordinal classification using Bayesian Additive Regression Trees(oBART). In addition, throughout the remaining of this section, the following notations will be used. Let $Y_i = k, i = 1, 2, ..., n$ and $k = 1, 2, ..., K$, where n is the total number of observations and K is the total number of ordinal categories. $X_{i,j}$ refers to the $i$-th value of the $j$-th predictor, out of $P$ predictors.

## 4.1 Simulation Studies

This section states and compares the performance of oBART to other models using numerous simulated datasets. Each dataset has an unique structure and generating process. There are 4 simulation settings, where two of them are the replications of previous papers and one of them is uniquely structured in this paper.

### 4.1.1 Simulation 1

The first simulation setting is the one used by **Janitza et el 2015** and **Hornung 2019**. It is first used by **Janitza et el 2015** and was later referenced by **Hornung 2019** with a slight modification because of its mixture nature within the generating process. In this study, it is again slightly modified for time efficiency. In both of the papers, there are 65 predictors, only 15 of which have nonzero regression coefficients and the other 50 of which have zero regression coefficients, in the purpose of assessing the variable selection capability of Random Forests and Ordinal Forests. The proposed model of this paper, oBART, does have the same working structure and capability of variable selection as the original BART; nonetheless, the predictors with zero regression coefficients are excluded since the variable selection is not the main purpose of the applied studies. In addition, the number of predictors having nonzero regression coefficients are reduced to 9, reducing two predictors for each of the three unique regression coefficients, 1, 0.75, and 0.5). Therefore, the data structure of the first simulation setting is as follows.

Let $\beta_{0,k}$ be the intercept term of the $k^{th}$ cumulative logit regression and $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_9) = (1, 1, 1, 0.75, 0.75, 0.75, 0.5, 0.5, 0.5)$ be the regression coefficients for the eight cumulative logit regressions. Then, having that $X_{i,j} \sim N(0,1)$, where $j = 1, ..., 9$ and that $(\beta_{0,1}, \beta_{0,2}, \beta_{0,3}, \beta_{0,4}, \beta_{0,5}, \beta_{0,6}, \beta_{0,7}, \beta_{0,8}) = (-5.9, -3.41, -1.55, -0.31$ we may define the probabilities of each observation belonging to different categories using the below structure.

$$log\left(\frac{P(Y_i \leq k)}{P(Y_i > k)}\right) = \beta_{0,k} + \boldsymbol{X_i}\boldsymbol{\beta}^T, k = 1, 2, ..., 8 \tag{35}$$

With those probabilities, the multinomial experiment is conducted for defining each observation to one of 9 ordered categories.

The table given below shows the summary of 100 repetitions of the simulation setting on the 3 models previously explained. In all cases, oBART clearly shows the best performance metrics.

| Model | $\mu_A(\sigma_A)$ | $\mu_{MAD}(\sigma_{MAD})$ | $\mu_{QWK}(\sigma_{QWK})$ |
|---|---|---|---|
| oBART | **0.5712(0.268)** | **0.6399(0.465)** | **0.6992(0.251)** |
| O-MBACT | 0.3316(0.03) | 1.1111(0.083) | 0.4465(0.094) |
| OF | 0.2884(0.034) | 1.2904(0.09) | 0.3335(0.094) |

### 4.1.2 Simulation 2

The second simulation setting is the one used by **Kindo, Wang, Pe (2013)**. To be specific, the setting introduced on "Nonlinear Ordered Simulated Data" of **Section 6** is used. However, it is also slightly simplified that only the 20 valid predictors out of the 50 predictors are included for the same reason as the first simulation setting.

Let $X_{i,j} \sim N(0,1)$, where $j = 1, ..., 20$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.78, 0.93, 1.08, 1.25)$, then $Z_i$ is calculated through the following structure and is divided into 5 categories according to $\boldsymbol{\gamma}$.

$$X^*_{i,1} = \max\{\min\{X_{i,1}, X_{i,2}, X_{i,3}\}, \min\{X_{i,4}, X_{i,5}, X_{i,6}\}, \min\{X_{i,7}, X_{i,8}, X_{i,9}\}\}$$

$$X^*_{i,2} = \min\{\max\{X_{i,11}, X_{i,12}, X_{i,13}\}, \max\{X_{i,14}, X_{i,15}, X_{i,16}\}, \max\{X_{i,17}, X_{i,18}, X_{i,19}\}\}$$

$$Z_i = 0.25X^*_{i,1} + 0.3X^*_{i,2} + \sin(0.2X_{i,10}) + \cos(0.1X_{i,20})$$

$$Y_i = \begin{cases} 1, & \text{if } Z_i < 0.78 \\ 2, & \text{if } 0.78 \leq Z_i < 0.93 \\ 3, & \text{if } 0.93 \leq Z_i < 1.08 \\ 4, & \text{if } 1.08 \leq Z_i < 1.25 \\ 5, & \text{if } 1.25 < Z_i \end{cases}$$

Given the setting explained above, the simulation was replicated for 100 times, and the table shown below displays and compares the overall performances of the 3 models on the tests. The model with the best performance is bold-faced for each metric. As it is shown, oBART by far exceeds the performances of the other models in all 3 standards.

| Model | $\mu_A(\sigma_A)$ | $\mu_{MAD}(\sigma_{MAD})$ | $\mu_{QWK}(\sigma_{QWK})$ |
|---|---|---|---|
| oBART | **0.5234(0.028)** | **0.5645(0.041)** | **0.8068(0.022)** |
| O-MBACT | 0.4348(0.025) | 0.6487(0.038) | 0.7578(0.023) |
| OF | 0.456(0.035) | 0.7956(0.065) | 0.7109(0.037) |

### 4.1.3 Simulation 3

The last simulation setting is structured to exhibit the distinguishing trait of oBART mentioned previously. oBART provides not only accurate predictions, but also interpretable predictions. (oBART provides posterior samples of both cutoff values $\gamma$ and latent variable $Z$ as well). Comparing predictive performances of the models is not just the sole purpose of the last simulation setting, as it was for the last two settings, but it is to verify the capability of providing information that would enhance interpretation, other than the variable selection. Suppose there are 9 ordinal categories, and the simulation is structured to distribute observations to each category as balan 5% of the observations to the first and last category, and the remaining 7 categories were designated the remaining 90% of the observations as identically as possible. designating categories is as follows.

$$\begin{aligned}
X_j &\sim N(0,2)\text{ , for } j = 1,2,3,7 \\
X_4 &\sim Poisson(5) \\
X_5 &\sim Bernoulli(0.3) \\
X_6 &\sim Bernoulli(0.5) \\
X_8 &= X_7^2 \\
X_9 &= X_2 \times X_3 \\
X_{10} &= X_2 \times X_3 \\
X_{11} &= X_2 \times X_3
\end{aligned}$$

$$Z_i = -2X_{i,1} + 4X_{i,2} - 2X_{i,3} + X_{i,4} + 5X_{i,5} - 5X_{i,6} + 4X_{i,7} + 2X_{i,8} - 3X_{i,10} - 4X_{i,11} + \epsilon_i, \text{where } \epsilon_i \sim N(0,1)$$

$$
Y_i = \begin{cases} 1, \text{ if } Z_i < -31 \\ 2, \text{ if } -31 \leq Z_i < -15 \\ 3, \text{ if } -15 \leq Z_i < -5.5 \\ 4, \text{ if } -5.5 \leq Z_i < 3 \\ 5, \text{ if } 3 \leq Z_i < 13 \\ 6, \text{ if } 13 \leq Z_i < 25 \\ 7, \text{ if } 25 \leq Z_i < 40 \\ 8, \text{ if } 40 \leq Z_i < 62 \\ 9, \text{ if } 62 \leq Z_i \end{cases}
$$

| Model | $\mu_A(\sigma_A)$ | $\mu_{MAD}(\sigma_{MAD})$ | $\mu_{QWK}(\sigma_{QWK})$ |
|---|---|---|---|
| oBART | **0.8785(0.026)** | **0.1253(0.027)** | 0.9682(0.011) |
| O-MBACT | 0.7859(0.031) | 0.2176(0.032) | **0.9685(0.011)** |
| OF | 0.6213(0.036) | 0.443(0.047) | 0.8701(0.036) |

The above table shows the overall performances of the models, and oBART peaks the performance in terms of accuracy and MAD once again in this setting while it is only barely behind the QWK of O-MBACT.

It was briefly mentioned that oBART can not only perform better than the other models in terms of accuracy, but also in terms of precision; if you are wrong, then it is absolutely important to know how far or close you are from the right answer. The oBART's superiority in precision can be noticed through the MAD's. In all three settings, oBART shows the undoubtedly best metrics.

## 4.2 Real Data Applications

The purpose of this experiment was to measure the thermal diffusivity of the copper metal. Hence that purpose is achieved and the experimental value is close enough to the theoretical value [4]. This experiment provides an opportunity to get acquainted with heat conduction in a way that is essentially different from that of classical experiments on stationary heat transmission. This experiment also allows one to learn thermal diffusivity measuring techniques in a simple and pedagogical way.

### 4.2.1 Real Data 1 : (title of the data)

The purpose of this experiment was to measure the thermal diffusivity of the copper metal. Hence that purpose is achieved and the experimental value is close enough to the theoretical value [4]. This experiment provides an opportunity to get acquainted with heat conduction in a way that is essentially different from that of classical experiments on stationary heat transmission. This experiment also allows one to learn thermal diffusivity measuring techniques in a simple and pedagogical way.

### 4.2.2 Real Data 2 : (title of the data)

The purpose of this experiment was to measure the thermal diffusivity of the copper metal. Hence that purpose is achieved and the experimental value is close enough to the theoretical value [4]. This experiment provides an opportunity to get acquainted with heat conduction in a way that is essentially different from that of classical experiments on stationary heat transmission. This experiment also allows one to learn thermal diffusivity measuring techniques in a simple and pedagogical way.

### 4.2.3 Real Data 3 : (title of the data)

The purpose of this experiment was to measure the thermal diffusivity of the copper metal. Hence that purpose is achieved and the experimental value is close enough to the theoretical value [4]. This experiment provides an opportunity to get acquainted with heat conduction in a way that is essentially different from that of classical experiments on stationary heat transmission. This experiment also allows one to learn thermal diffusivity measuring techniques in a simple and pedagogical way.

## 5 Conclusion

Although there is a countless number of cases to be applied, there has not been a great abundance of options for an ordinal classification. For the time being, it has been quite customary to simply treat ordinal categorical variables as nominal categorical variables and apply models as such. Apparently, this rather convenient approach cannot ensure promising results after all so the demands and studies for nonparametric ordinal classification approaches have risen consistently. Among those, the model that showed great improvement in predictive performance was Ordered-Multiclass Bayesian Additive Classification Trees, the variation of Bayesian Additive Regression Trees for ordinal classification. Nevertheless, there were some adjustments that could be made, and oBART, the proposed model of this paper with those adjustments, shows a leap of predictive power in both the simulation settings and the real data. Clearly, oBART also has room for improvement for the ones mentioned in the paper, which could be studied further in the future studies. However, it is hard to deny that oBART has an innegligibly advanced predictive power as of now.

## References

[1] A. Mandelis, L. Nicolaides, Y. Chen, "*Structure and the reflectionless/refractionless nature of parabolic diffusion-wave fields*", Phys. Rev.Lett. **87**, 020801-1—020801-4 (2001).

[2] A. Bodas, V. Gandia, E. Lopez-Baeza, "*An undergraduate experiment on the propagation of thermal waves*", Am. J. Phys. **66**, 528-1-533 (1998).

[3] L. Verdini and A. Santucci, "*Propagation properties of thermal waves and thermal diffusivity in metals*", Nuovo Cimento B **62**, 399-421 (1981).

[4] M. S. Anwar, J. Alam, M. Wasif, R. Ullah, S. Shamim, W. Zia, "*Fourier analysis of thermal diffusive waves*", J. Phy **82**, 928 (2014).

[5] George B. Arfken, Hans J. Weber, "*Mathematical methods for physicists*", 6th Ed, Chapter 14.

[6] K. Etori, "*Remarks on the temperature propagation and the thermal diffusivity of a solid*", Jpn. J. Appl. Phys. **11**, 955-957 (1972).