Statistics and Computing (2024) 34:123
https://doi.org/10.1007/s11222-024-10434-4

ORIGINAL PAPER

Type I Tobit Bayesian Additive Regression Trees for censored outcome regression

Eoghan O'Neill[1]

Received: 15 October 2023 / Accepted: 30 April 2024

Abstract
Censoring occurs when an outcome is unobserved beyond some threshold value. Methods that do not account for censoring produce biased predictions of the unobserved outcome. This paper introduces Type I Tobit Bayesian Additive Regression Tree (TOBART-1) models for censored outcomes. Simulation results and real data applications demonstrate that TOBART-1 produces accurate predictions of censored outcomes. TOBART-1 provides posterior intervals for the conditional expectation and other quantities of interest. The error term distribution can have a large impact on the expectation of the censored outcome. Therefore, the error is flexibly modeled as a Dirichlet process mixture of normal distributions. An R package is available at https://github.com/EoghanONeill/TobitBART.

Keywords BART · Censored regression · Regression Trees · Bayesian nonparametrics · Machine learning

1 Introduction

Censoring occurs when, beyond some threshold value, the observed outcome is equal to the threshold instead of the true latent outcome value. For example, scientific equipment can often only make accurate measurements within a known range of outcome values, and observations outside this range are set to its limits. Often the estimand of interest is the conditional expectation or conditional average treatment effect on the outcome before censoring. Estimation of a standard regression model using data without censored values, or with censored observations set equal to threshold values, results in biased estimates. Tobit models directly model the latent outcome and censoring process (Tobin 1958).

In this paper, we combine the Bayesian Type I Tobit model (Chib 1992) with Bayesian Additive Regression Trees (Chipman et al. 2010). The latent outcome (before censoring) is modeled as a sum-of-trees, which allows for nonlinear functions of covariates. The error term is modeled as a Dirichlet process mixture of normal distributions, as in fully nonparametric BART (George et al. 2019). Smooth data generating processes with sparsity are modelled by soft trees with a Dirichlet prior on splitting variable probabilities, as introduced by Linero and Yang (2018).

In simulations and applications to real data, TOBART-1 outperforms a Tobit gradient boosted tree method, Grabit (Sigrist and Hirnschall 2019), a Tobit Gaussian Process model (Groot and Lucas 2012), standard linear Tobit, and simple hurdle models based on standard machine learning methods. Unlike other methods, TOBART-1 accounts for model uncertainty and can non-parametrically model the error term. Posterior intervals are available for censored outcomes, uncensored outcomes, conditional expectations, and probabilities of censoring. Grabit, Gaussian Processes, and other methods rely on cross-validation for parameter tuning and are sensitive to the tuned variance of the error term, whereas TOBART-1 performs well without parameter tuning and accounts for uncertainty in the variance of the error term.

TOBART-1 with a Dirichlet process mixture of normal distributions for the error term (TOBART-1-NP) removes the restrictive normality assumption often imposed in censored outcome models. We observe that this can lead to more accurate outcome predictions in simulations with non-normally distributed errors, and in real data applications, which may involve non-normally distributed outcomes.[1]

✉ Eoghan O'Neill
oneill@ese.eur.nl

[1] Econometric Institute, Erasmus University Rotterdam, Rotterdam, Netherlands

[1] A Dirichlet process mixture for the error term distribution has previously been included in a censored outcome model by Kottas and Krnjajić (2009).

Published online: 24 May 2024

A variety of methods have been proposed for nonparametric and semiparametric censored outcome models. Lewbel and Linton (2002) describe a local linear kernel estimator for the setting in which both the uncensored outcome mean function of regressors and error distribution are unknown. Fan and Gijbels (1994) describe a quantile-based local linear approximation method. Huang (2021) introduces a semiparametric method involving B-splines. Chen et al. (2005) use a local polynomial method. Other papers on the topic of semiparametric and nonparametric censored outcome regression include Cheng and Small (2021), Heuchenne and Van Keilegom (2007, 2010), Huang et al. (2019), Oganisian et al. (2021). Gaussian Process censored outcome regression methods are applied by Groot and Lucas (2012), Cao et al. (2018), Gammelli et al. (2020, 2022), Basson et al. (2023). Zhang et al. (2021), Wu et al. (2018) implement censored outcome neural network methods.

A number of recent papers have considered Tobit model selection and regularization. Zhang et al. (2012) describe Focused Information Criteria based Tobit model selection and averaging. Jacobson and Zou (2024) provide theoretical and empirical results for Tobit with a Lasso penalty and a folded concave penalty (SCAD). Müller and van de Geer (2016) and Soret et al. (2018) describe a LASSO penalized censored outcome models. Bradic and Guo (2016) study robust penalized estimators for censored outcome regression.

The Bayesian Tobit literature includes quantile regression methods (Ji et al. 2012; Yu and Stander 2007; Alhamzawi 2016), and Bayesian elastic net Tobit (Alhamzawi 2020). Ji et al. (2012) account for model uncertainty by implementing Tobit quantile regression with Stochastic Search Variable Selection. However, the outcome and latent variable are modeled as linear functions of covariates. TOBART-1 provides a competing approach to the methods referenced above that does not impose linearity.

The remainder of the paper is structured as follows: In Sect. 2 we describe the TOBART-1 model and Markov chain Monte Carlo (MCMC) implementation, Sect. 3 contains simulation studies for prediction and treatment effect estimation with censored data, Sect. 4 contains applications to real world data, and Sect. 5 concludes the paper.

2 Methods

2.1 Review of Bayesian Additive Regression Trees (BART)

Suppose there are n observations, and the $n \times p$ matrix of explanatory variables, X, has i^{th} row $x_i = [x_{i1}, ..., x_{ip}]$. Following the notation of Chipman et al. (2010), let T be a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and let $M = \{\mu_1, ..., \mu_b\}$ denote a set of parameter values associated with each of the b terminal nodes of T. The interior node decision rules are binary splits of the predictor space into the sets $\{x_{is} \leq c\}$ and $\{x_{is} > c\}$ for continuous x_s. Each observation's x_i vector is associated with a single terminal node of T, and is assigned the μ value associated with this terminal node. For a given T and M, the function $g(x_i; T, M)$ assigns a $\mu \in M$ to x_i.

For the standard BART model, the outcome is determined by a sum of trees,

$$Y_i = \sum_{j=1}^{m} g(x_i; T_j, M_j) + \varepsilon_i$$

where $g(x_i; T_j, M_j)$ is the output of a decision tree. T_j refers to a decision tree indexed by $j = 1, ..., m$, where m is the total number of trees in the model. M_j is the set of terminal node parameters of T_j, and $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

Prior independence is assumed across trees T_j and across terminal node means $M_j = (\mu_{1j} ... \mu_{b_j j})$ (where $1, ..., b_j$ indexes the terminal nodes of tree j). The form of the prior used by Chipman et al. (2010) is $p(M_1, ..., M_m, T_1, ..., T_m, \sigma) \propto \left[\prod_j \left[\prod_k p(\mu_{kj}|T_j)\right] p(T_j)\right] p(\sigma)$ where $\mu_{kj}|T_j \overset{i.i.d}{\sim} N(0, \sigma_\mu^2)$ where $\sigma_\mu = \frac{0.5}{\kappa\sqrt{m}}$ and κ is a user-specified hyperparameter.

Chipman et al. (2010) set a regularization prior on the tree size and shape $p(T_j)$. The probability that a given node within a tree T_j is split into two child nodes is $\alpha(1 + d_h)^{-\beta}$, where d_h is the depth of (internal) node h, and the parameters α and β determine the size and shape of T_j respectively. Chipman et al. (2010) use uniform priors on available splitting variables and splitting points. The model precision σ^{-2} has a conjugate prior distribution $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$ with degrees of freedom v and scale λ.

Samples from $p((T_1, M_1), ..., (T_m, M_m), \sigma|y)$ can be made by a Bayesian backfitting MCMC algorithm. This algorithm involves m successive draws from $(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y$ for $j = 1, ..., m$, where $T_{(j)}, M_{(j)}$ are the trees and parameters for all trees except the j^{th} tree, followed by a draw of σ from the full conditional $\sigma|T_1, ..., T_m, M_1, ..., M_m, y$. After burn-in, the sequence of f^* draws, $f_1^*, ..., f_Q^*$, where $f^*(.) = \sum_{j=1}^{m} g(., T_j^*, M_j^*)$, is an approximate sample of size Q from $p(f|y)$.

2.2 Soft trees and sparse splitting rules

In addition to the standard Bayesian tree model for $f(\boldsymbol{x}_i)$ described in Sect. 2.1, we also implement TOBART and TOBART-NP with soft trees and sparse splitting rules as described by Linero and Yang (2018). Predictions from soft trees are weighted linear combinations of all terminal node parameter values, with the weights being functions of dis-

tances between covariates and splitting points. The prediction from a single tree function is

$$g(\boldsymbol{x}_i; T_j, M_j) = \sum_{\ell=1}^{L_j} \mu_{j,\ell}\xi(\boldsymbol{x}_i, T_j, \ell)$$

$$\xi(\boldsymbol{x}_i, T_j, \ell) = \prod_{b\in\mathcal{A}(\ell)} \zeta\left(\frac{x_{j_b} - C_b}{\tau_b}\right)^{\mathbb{I}\{x_{j_b} > C_b\}} \times\left\{1 - \zeta\left(\frac{x_{j_b} - C_b}{\tau_b}\right)\right\}^{\mathbb{I}\{x_{j_b} \le C_b\}}$$

where L_j is the number of leaves in the j^{th} tree, $\mu_{j,\ell}$ is the ℓ^{th} terminal node parameter of the j^{th} tree, $\mathcal{A}(\ell)$ denotes the set of ancestor nodes of terminal node ℓ. The splitting variable, splitting point, and bandwidth parameter at internal node b are denoted by x_{j_b}, C_b, and τ_b respectively. The gating function ζ is the logistic function $\zeta(x) = (1 + \exp(-x))^{-1}$.

Sparse splitting rules are introduced by placing a Dirichlet prior on the splitting probabilities $(s_1, \ldots, s_p) \sim \mathcal{D}(\frac{a}{p}, \ldots, \frac{a}{p})$. The parameter a controls the level of sparsity and has the prior Beta(0.5, 1). Linero and Yang (2018) demonstrate that soft trees allow BART to model smooth functions, and the Dirichlet prior on splitting probabilities adapts to unknown levels of sparsity to provide improved predictions on high dimensional data sets.

2.3 Type I Tobit and TOBART

2.3.1 Type I Tobit model

The Type I Tobit model with censoring from below at a and censoring from above at b is:

$$Y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i \;\; \varepsilon_i \sim i.i.d.\; N(0, \sigma^2)$$

$$Y_i = \begin{cases} a \text{ if } Y_i^* \le a \\ Y_i^* \text{ if } a < Y_i^* < b \\ b \text{ if } b \le Y_i^* \end{cases}$$

where a normal prior is placed on β, and an inverse gamma prior is placed on σ^2 (Chib 1992).

2.3.2 Type I TOBART model

The Type I TOBART model replaces the linear combination $\boldsymbol{x}_i\boldsymbol{\beta}$ with the sum-of-trees function $f(\boldsymbol{x}_i)$:

$$Y_i^* = f(\boldsymbol{x}_i) + \varepsilon_i \;\; \varepsilon_i \sim i.i.d.\; N(0, \sigma^2)$$

$$Y_i = \begin{cases} a \text{ if } Y_i^* \le a \\ Y_i^* \text{ if } a < Y_i^* < b \\ b \text{ if } b \le Y_i^* \end{cases}$$

where a BART prior is placed on $f(\boldsymbol{x}_i)$ and an inverse gamma prior is placed on σ^2.[2]

2.3.3 Type I TOBART Gibbs sampler

Tobit can be implemented by MCMC with data augmentation (Chib 1992). The realization, y_i^*, of the variable Y_i^* is observed for uncensored outcomes, and is sampled from its full conditional for censored outcomes.

$$y_i^* = y_i \text{ if } y_i \in (a, b) \;\text{ and}$$

$$y_i^* \sim \begin{cases} \mathcal{TN}_{[-\infty,a]}(f(\boldsymbol{x}_i), \sigma^2) \text{ if } y_i = a \\ \mathcal{TN}_{[b,\infty]}(f(\boldsymbol{x}_i), \sigma^2) \text{ if } y_i = b \end{cases}$$

where $\mathcal{TN}_{[l,u]}$ denotes a normal distribution truncated to the interval $[l, u]$. The full conditionals for $f(\boldsymbol{x}_i)$ and σ^2 are standard full conditionals for BART with y_i^* as the dependent variable and $\boldsymbol{x}_i$ as the potential splitting variables. Appendix A contains a description of a sampler that produces draws $f^{(1)}(\boldsymbol{x}_i), \ldots, f^{(D)}(\boldsymbol{x}_i)$ and $\sigma^{(1)}, \ldots, \sigma^{(D)}$.

2.3.4 Predicting outcomes with TOBART

The conditional mean of the latent variable is $f(\boldsymbol{x}_i)$. If censoring is also applied to the test data, then the outcomes are predicted by averaging the standard Tobit expectation formula across MCMC iterations:

For all MCMC iterations $d = 1, ..., D$ calculate

$$\begin{aligned} E[Y_i|X_i = \boldsymbol{x}_i, f^{(d)}, \sigma^{(d)}] &= a\Phi\Big(\frac{a - f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\Big) \\ &+ f^{(d)}(\boldsymbol{x}_i)\Big[\Phi\Big(\frac{b - f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\Big) - \Phi\Big(\frac{a - f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\Big)\Big] \\ &+ \sigma^{(d)}\Big(\phi\Big(\frac{a - f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\Big) - \phi\Big(\frac{b - f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\Big)\Big) \\ &+ b\Big[1 - \Phi\Big(\frac{b - f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\Big)\Big] \end{aligned}$$

The predicted outcome is $\frac{1}{D}\sum_{d=1}^{D} E[Y_i|X_i = \boldsymbol{x}_i, f^{(d)}, \sigma^{(d)}]$. The expectation conditional on the outcome not being in the

[2] $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$. For standard BART, λ is set such that the q^{th} quantile of the prior distribution of σ is the sample standard deviation of the residuals from a linear model. For censored outcomes, this may give poor calibration of the σ prior. We consider four options in a simulation study in appendix E. A sample standard deviation estimate from an intercept-only Tobit model generally gives good results, although often there is little difference across λ values.

censored range is:

$$E[Y_i|a < Y_i < b, X_i = \boldsymbol{x}_i, f^{(d)}, \sigma^{(d)}]$$
$$= f^{(d)}(\boldsymbol{x}_i) + \sigma^{(d)} \frac{\phi\left(\frac{a-f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\right) - \phi\left(\frac{b-f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\right)}{\Phi\left(\frac{b-f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\right) - \Phi\left(\frac{a-f^{(d)}(\boldsymbol{x}_i)}{\sigma^{(d)}}\right)}$$

2.4 Nonparametric Type I TOBART

2.4.1 Nonparametric Type I TOBART model

The accuracy of the conditional expectation of the TOBART model depends on the validity of the assumption of normality of the errors. More general censored outcomes can be modelled by assuming a Dirichlet Process mixture distribution for the error terms.

$$y_i^* = f(\boldsymbol{x}_i) + \varepsilon_i \;\; y_i = \begin{cases} a \text{ if } y_i^* \leq a \\ y_i^* \text{ if } a < y_i^* < b \\ b \text{ if } b \leq y_i^* \end{cases}$$
$$\varepsilon_i \sim i.i.d.\ N(\gamma_i, \sigma_i^2) \;\; \vartheta_i = (\gamma_i, \sigma_i) \sim G$$
$$G \sim \mathcal{DP}(G_0, \alpha)$$

The distribution of the error term is specified similarly to George et al. (2019). The base distribution G_0 is defined as follows:

$$p(\gamma, \sigma|\nu, \lambda_1, \gamma_0, k_0) = p(\sigma|\nu, \lambda)p(\gamma|\sigma, \gamma_0, k_0)$$
$$\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2} \;\; \gamma|\sigma \sim \mathcal{N}\left(\gamma_0, \frac{\sigma^2}{k_0}\right)$$

where, in contrast to the standard BART prior of Chipman et al. (2010), ν is set to 10 instead of 3.[3] The parameter λ is set such that the q^{th} quantile of the prior distribution of σ is the sample standard deviation of the outcome, or of the residuals from a linear model. For TOBART-NP, $q = 0.9$ instead of 0.95.[4] The prior on α is the $\alpha \sim \Gamma(2, 2)$ prior introduced by Escobar and West (1995) and applied by Van Hasselt (2011).[5]

The outcome is scaled by subtracting the sample mean before applying the Gibbs sampler, therefore George et al. (2019) set $\gamma_0 = 0$.[6] The parameter k_0 is scaled with the marginal distribution of γ ($\gamma \sim \frac{\sqrt{\lambda}}{\sqrt{k_0}} t_\nu$). Given k_s (set to 10 by default), k_0 is set such that $\max_{i=1,\ldots,n} |e_i| = k_s \frac{\sqrt{\lambda}}{\sqrt{k_0}}$ where $k_s = 10.$ and $e_1, \ldots, e_n$ are the residuals from a linear model.[7] The Gibbs sampler for TOBART-NP is described in Appendix A.

For each MCMC iteration, d, and observation i, we obtain $\vartheta_i^{(d)} = (\gamma_i^{(d)}, \sigma_i^{(d)})$. The conditional expectation, $E[y_i|\boldsymbol{x}_i, f^{(d)}, \gamma_i^{(d)}, \sigma^{(d)}]$, is calculated as outlined in Sect. 2.3.4.

2.5 Treatment effect estimation for censored outcomes

Let a binary variable T_i equal 1 if unit i is assigned to treatment and 0 if i is assigned to the control group. The potential outcomes under treatment and control group allocation are denoted by $Y_i(1)$ and $Y_i(0)$ respectively. Similarly, the potential outcomes of the latent outcome are denoted by $Y_i^*(1)$, $Y_i^*(0)$. Assume the data generating process is as follows:

$$Y_i^* = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)T_i + \varepsilon_i \;\; \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \begin{cases} a \text{ if } Y_i^* \leq a \\ Y_i^* \text{ if } a < Y_i^* < b \\ b \text{ if } b \leq Y_i^* \end{cases}$$

where $\mu(\boldsymbol{x}_i)$ and $\tau(\boldsymbol{x}_i)$ are possibly nonlinear functions of covariates. Assume conditional unconfoundedness, i.e. $Y_i^*(1), Y_i^*(0) \perp T_i|X_i$. The estimand is the conditional average treatment effect on Y_i^*, i.e., $E[Y_i^*(1) - Y_i^*(0)|X_i = \boldsymbol{x}_i] = \tau(\boldsymbol{x}_i)$. However, a model naively trained on only uncensored outcomes estimates the following effects[8]

[3] George et al. (2019) recommend $\nu = 10$ as the spread of the error increases when there are many components and the spread of a single components can be reduced by increasing ν. This gives better results than $\nu = 3$ for some DGPs in a simulation study in Appendix E.

[4] This is complicated by the censoring of the outcome. Some options are: 1. Estimate the standard deviation assuming that censored outcome is normally distributed. 2. Estimate the standard deviation of a linear type I Tobit model (contains option 1 as a special case but not feasible when there are more regressors than observations). 3. Estimate the standard deviation of the censored outcome without accounting for censoring. We use option 2 for TOBART-NP.

[5] The TobitBART package also includes an option for the prior described by Rossi (2014) and George et al. (2019), $p(\alpha) \propto \left(1 - \frac{\alpha-\alpha_{min}}{\alpha_{max}-\alpha_{min}}\right)^{\psi}$, where α_{min} and α_{max} are set so that the modal numbers of components are $I_{min} = 1$ and $I_{max} = [(0.1)n]$ respectively, and $\psi = 0.5$.

[6] However, the mean cannot be estimated for censored data without making further assumptions. Options include: 1. Estimate the mean (and variance) of a censored normal distribution. 2. Calculate the sample mean of the censored outcome without accounting for censoring. We use option 1.

[7] The residuals likely underestimate the true errors for censored observations.

[8] This bias occurs if all the uncensored observations are included in one regression and differences in predictions for $T_i = 1$ and $T_i = 0$ are obtained, i.e. an S-learner approach (Künzel et al. 2019), or if the two conditional expectations are obtained from separate regressions for treated and untreated uncensored observations, i.e. a T-Learner approach. In both cases, the conditional expectations are not equal to the expectation of the latent outcome.

$$
\begin{aligned}
&E[Y_i(1)|a < y_i < b, X_i = \boldsymbol{x}_i] \\
&-E[Y_i(0)|a < y_i < b, X_i = \boldsymbol{x}_i] = \tau(\boldsymbol{x}_i) \\
&+\sigma\left(\frac{\phi\left(\frac{a-(\mu(\boldsymbol{x}_i)+\tau(\boldsymbol{x}_i)))}{\sigma}\right) - \phi\left(\frac{b-(\mu(\boldsymbol{x}_i)+\tau(\boldsymbol{x}_i))}{\sigma}\right)}{\Phi\left(\frac{b-(\mu(\boldsymbol{x}_i)+\tau(\boldsymbol{x}_i))}{\sigma}\right) - \Phi\left(\frac{a-(\mu(\boldsymbol{x}_i)+\tau(\boldsymbol{x}_i))}{\sigma}\right)}\right. \\
&\left.-\frac{\phi\left(\frac{a-\mu(\boldsymbol{x}_i)}{\sigma}\right) - \phi\left(\frac{b-\mu(\boldsymbol{x}_i)}{\sigma}\right)}{\Phi\left(\frac{b-\mu(\boldsymbol{x}_i)}{\sigma}\right) - \Phi\left(\frac{a-\mu(\boldsymbol{x}_i)}{\sigma}\right)}\right).
\end{aligned}
$$

A sufficiently flexible nonparametric method, without restrictive assumptions on the error term, will produce estimates that approximate the expression above. A model naively trained on the full data set with censoring similarly gives biased estimates (see Appendix B). By directly modelling Y_i^*, censored outcome models avoid the bias described above. Similar biases occur if the error term is not normally distributed.

3 Simulation studies

3.1 Description of prediction simulations

We adapt the data generating processes (DGPs) introduced by Friedman (1991) to a censored regression setting. This DGP has often been applied in comparisons of semiparametric regression methods. We also make use of the censored outcome simulations described by Groot and Lucas (2012), Sigrist and Hirnschall (2019), and Jacobson and Zou (2024) for fair comparison against competing methods with existing synthetic censored data.

The covariates $x_1, ..., x_p$ are independently sampled from the uniform distribution on the unit interval. The outcome before censoring is generating from one of the following functions:

- $y^* = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with censoring from below at the 15^{th} percentile of the training data y^* values (Friedman 1991).[9]
- $y^* = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with censoring from below at the 15^{th} percentile of the training data y^* values, and from above at the 85^{th} percentile of the training data y^* values (Friedman 1991).
- $y^* = 6(x_1 - 2)^2 \sin(2(6x_1 - 2)) + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with censoring from below at the 40^{th} percentile of the training data y^* values (Groot and Lucas 2012).
- $y^* = \sum_{k=1}^{5} 0.3 \max(x_k, 0) + \sum_{k=1}^{3} \sum_{j=k+1}^{4} \max(x_k x_j, 0) + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with censoring from above at the 95^{th} percentile of the training data y^* values (Sigrist and Hirnschall 2019). For this simulation, $x_1, ..., x_p$ are uniformly distributed on $[-1, 1]$ instead of $[0, 1]$.[10]
- $y^* = 3 + 5x_1 + x_2 + \frac{x_3}{2} - 2x_4 + \frac{x_5}{10} + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with censoring from below at the 25^{th} percentile of the training data y^* values (Jacobson and Zou 2024).

The variance of the error, σ^2, is set to 1. See the Supplementary Appendix (Online Resource 1) for the results obtained from simulations with $\sigma \in \{0.1, 2\}$. We also consider deviations from the assumption of normally distributed errors. In particular, we include results for simulations in which ε is generated from Skew-t, and Weibull(1/2, 1/5) distributions.[11] The number of covariates, p, is set to 30. We generate 500 training and 500 test observations.

3.2 Prediction simulation results

We compare the performance of TOBART-1, TOBART-1-NP, Soft TOBART-1, and Soft TOBART-1-NP against Grabit (Sigrist and Hirnschall 2019), linear Tobit (Tobin 1958), BART (Chipman et al. 2010), Random Forests (RF) (Breiman 2001), Gaussian Processes, and a Tobit Gaussian Process model (Groot and Lucas 2012).[12] The results for a Gaussian Process (GP) with only 5 variables (always including all informative variables) are included because GPs were observed to produce inaccurate predictions when applied to data with 30 variables.[13] Censored outcome predictions are evaluated using Mean Squared Error (MSE), and predicted probabilities of censoring are evaluated using the Brier Score.[14] [15] All results are averaged over 5 repetitions. [16]

The results for simulations with normally distributed errors are presented in Tables 1 and 2. The TOBART algorithms generally outperform competing methods across all DGPs, except unsurprisingly for the linear Jacobson and Zou (2024) simulations linear Tobit is outperformed only by Soft

[9] The original Friedman simulations did not involve censoring.

[10] This simulation differs somewhat from the original simulation of Sigrist and Hirnschall (2019) for which the variable determining censoring was not perfectly correlated with the observed outcome before censoring.

[11] Bradic and Guo (2016) considered Weibull(1/2, 1/5) errors in a simulation study.

[12] Standard BART for continuous outcomes is trained on censored outcomes. Probit BART is trained on a binary variable indicating censorship. Similarly, Random Forests are separately trained on continuous censored outcomes and a binary censorship indicator.

[13] The GP Matlab code was obtained from https://www.cs.ru.nl/~perry/software/tobit1.html.

[14] See the Supplementary Appendix (Online Resource 1) for implementation details and parameter settings.

[15] Latent outcome predictions similarly demonstrate that TOBART outperforms other methods, and these results are available on request. However, it is unsurprising that Tobit based latent outcome predictions outperform naive approaches due to the aforementioned censoring bias.

[16] Computational times are included in Appendix D.

TOBART. TOBART-NP can slightly improve on TOBART in some cases, but generally the results are similar when errors are normally distributed. The differences in criteria across methods are small for the more linear DGPs from Sigrist and Hirnschall (2019) and Jacobson and Zou (2024), as linear Tobit is designed for a linear DGP, and the nonlinear methods BART and RF can model the relatively simple response surface well. It is worth noting that TOBART outperforms Grabit even though the true standard deviation, $\sigma = 1$, is included as one of five possible Grabit hyperparameter values in cross-validation. The same pattern of results can be observed for simulations with $\sigma = 0.1$ and $\sigma = 2$ in the Supplementary Appendix. The Supplementary Appendix contains comparisons of Area Under the Curve for all methods and DGPs, from which similar conclusions can be drawn.

The results for Skew-t and Weibull distributed errors are also presented in Tables 1 and 2.[17] The TOBART models outperform all other methods for almost all DGPs and criteria. The results for the Weibull distribution generally favour TOBART-NP and Soft TOBART-NP, indicating that there is some improvement from the Dirichlet Process model when the errors are sufficiently non-Gaussian.

The average coverage and length of 95% prediction intervals for the latent outcomes and the observed outcomes are given in the Supplementary Appendix (Online Resource 1). For most DGPs and error distributions, TOBART and Soft TOBART provides the closest to 95% coverage of prediction intervals for both latent and observed outcomes. For some DGPs with non-normal errors, the more conservative intervals produced by TOBART-NP and Soft TOBART-NP provide better coverage.

3.3 Description of treatment effect simulations

A number of recent simulation studies have demonstrated that BART is among the most accurate treatment effect estimation methods (Wendling et al. 2018; McConnell and Lindner 2019; Dorie et al. 2019; Hahn et al. 2019). However, in practice many data sets, including randomized trial data sets, contain censored outcomes. For example, antibody concentrations or environmental levels of chemicals can only be measured accurately within a certain range as a result of limitations of measuring equipment. Often economic data is censored due to privacy considerations, for example income might be censored above a certain threshold. TOBART provides a machine learning treatment effect estimation method with uncertainty quantification that can be applied to this data while still making use of the information provided by censored observations. We demonstrate the effectiveness of TOBART by censoring the outcomes of DGPs from published studies of machine learning methods for treatment effect estimation. The chosen data generating processes contain linear and non-linear functions of covariates, constant and heterogeneous effects, and various degrees of confounding.

3.3.1 Censored Caron et al. (2022) simulations

$P = 10$ covariates are generated from a multivariate Gaussian distribution, $X_1, \ldots, X_{10} \sim \mathcal{MVN}(\mathbf{0}, \Sigma)$, with $\Sigma_{jk} = 0.6^{|j-k|} + 0.1\mathbb{I}(j \neq k)$. The binary treatment variable is Bernoulli distributed, $Z_i \sim \text{Bern}(\pi(\boldsymbol{x}_i))$, where

$$\pi(\boldsymbol{x}_i) = \Phi(-0.4 + 0.3X_{i,1} + 0.2X_{i,2})$$

and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.
The prognostic score function, $\mu(\boldsymbol{x}_i)$, and CATE function, $\tau(\boldsymbol{x}_i)$, are defined as

$$\mu(\boldsymbol{x}_i) = 3 + X_{i,1} + 0.8\sin(X_{i,2}) + 0.7X_{i,3}X_{i,4} - X_{i,5}$$
$$\tau(\boldsymbol{x}_i) = 2 + 0.8X_{i,1} - 0.3X_{i,2}^2$$

The outcome before censoring is generated as:

$$Y_i^* = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)Z_i + \varepsilon_i \text{ where } \varepsilon_i \sim \mathcal{N}(0, 1)$$

The number of sampled observations is 200. The observed outcome Y_i is censored from below at the 15^{th} percentile of the generated Y_i^* values, and from above at the 85^{th} percentile.

3.3.2 Censored Friedberg et al. (2020) simulations

$P = 20$ covariates are generated from independent standard uniform distributions $X_1, ..., X_{20} \sim \mathcal{U}[0, 1]$. There is no confounding as $\pi(\boldsymbol{x}_i) = 0.5$ and $Z_i \sim \text{Bern}(\pi(\boldsymbol{x}_i))$. The prognostic score function, $\mu(\boldsymbol{x}_i)$, and CATE function, $\tau(\boldsymbol{x}_i)$, are defined as $\mu(\boldsymbol{x}_i) = 0$ and

$$\tau(\boldsymbol{x}_i) = \left(1 + \frac{1}{1 + \exp\left(-20\left(X_{i,1} - \frac{1}{3}\right)\right)}\right) \times \left(1 + \frac{1}{1 + \exp\left(-20\left(X_{i,2} - \frac{1}{3}\right)\right)}\right).$$

The outcome before censoring is generated as:

$$Y_i^* = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)Z_i + \varepsilon_i \text{ where } \varepsilon_i \sim \mathcal{N}(0, 1)$$

The number of sampled observations is 200. The observed outcome Y_i is censored from below at the 15^{th} percentile of the generated Y_i^* values, and from above at the 85^{th} percentile.

[17] Results for t-distributed errors with $\nu = 3$ are in the Supplementary Appendix.

Table 1 Simulation study, mean squared error

Data	*Tobit*	BART	RF	Grabit	TOBART	TOBART NP
Normal distribution, $\sigma = 1$						
Friedman (1991)	4.764	1.765	4.559	2.291	1.162	1.154
Friedman (1991) 1 side	6.444	1.768	6.004	2.743	1.457	1.509
Groot and Lucas (2012)	12.886	2.247	4.612	0.702	0.631	0.617
Jacobson and Zou (2024)	0.694	0.722	0.855	0.739	0.718	0.720
Sigrist and Hirnschall (2019)	1.353	1.142	1.170	1.146	1.072	1.074
Skew-t distribution, $location = 1$, $scale = 1$, $\nu = 4$						
Friedman (1991)	5.075	2.172	4.790	2.495	1.648	1.597
Friedman (1991) 1 side	6.815	2.352	6.190	3.197	2.163	2.060
Groot and Lucas (2012)	14.700	2.999	5.563	1.204	1.149	1.071
Jacobson and Zou (2024)	1.195	1.264	1.301	1.271	1.238	1.184
Sigrist and Hirnschall (2019)	1.519	1.333	1.337	1.340	1.279	1.267
Weibull distribution, $shape = 0.5$, $scale = 0.2$						
Friedman (1991)	4.599	1.502	4.576	2.099	0.871	0.811
Friedman (1991) 1 side	6.221	1.624	5.962	2.650	1.344	1.154
Groot and Lucas (2012)	11.297	1.746	4.071	0.760	0.787	0.648
Jacobson and Zou (2024)	0.721	0.814	0.873	0.783	0.960	0.696
Sigrist and Hirnschall (2019)	0.739	0.473	0.536	0.447	0.426	0.362

Data	*SoftBART*	GP	GP 5 vars	GP Tobit	Soft TOBART	Soft TOBART NP
normal distribution, $\sigma = 1$						
Friedman (1991)	0.942	50.248	0.985	50.331	**0.743**	0.745
Friedman (1991) 1 side	1.101	85.594	1.076	85.532	0.929	**0.927**
Groot and Lucas (2012)	2.187	16.045	2.728	16.042	**0.546**	**0.546**
Jacobson and Zou (2024)	0.684	7.109	0.681	7.053	**0.675**	**0.675**
Sigrist and Hirnschall (2019)	1.002	2.460	0.949	2.487	**0.973**	**0.973**
Skew-t distribution, $location = 1$, $scale = 1$, $\nu = 4$						
Friedman (1991)	1.250	47.049	1.316	47.099	1.066	**1.027**
Friedman (1991) 1 side	1.579	81.482	1.604	81.415	1.405	**1.367**
Groot and Lucas (2012)	2.769	19.210	3.348	19.210	0.947	**0.933**
Jacobson and Zou (2024)	1.170	8.877	1.138	8.823	1.142	**1.133**
Sigrist and Hirnschall (2019)	1.146	4.487	1.100	4.556	1.135	**1.127**
Weibull distribution, $shape = 0.5$, $scale = 0.2$						
Friedman (1991)	0.721	46.692	0.758	46.730	0.492	**0.436**
Friedman (1991) 1 side	1.054	80.639	1.001	80.573	0.787	**0.733**
Groot and Lucas (2012)	1.712	15.435	1.903	15.436	0.610	**0.598**
Jacobson and Zou (2024)	0.718	7.447	0.716	7.392	0.705	**0.691**
Sigrist and Hirnschall (2019)	0.338	2.708	0.338	2.744	0.340	**0.305**

Minimum values, excluding GP trained with only the 5 relevant variables, are in bold

3.3.3 Censored Nie and Wager (2021) simulations

The covariates are generated as follows across scenarios A to D. In simulation A, $X_1, ..., X_{12} \sim \mathcal{U}[0, 1]$. In simulations B to D, $X_1, ..., X_{12} \sim \mathcal{N}(0, 1)$.

$\pi(\boldsymbol{x}_i)$ is defined as follows across scenarios A to D: (A) $\text{trim}_{0.1}\{\sin(\pi X_{i,1} X_{i,2})\}$, (B) constant equal to 0.5, (C) $1/\{1+\exp(X_{i,2} + X_{i,3})\}$, (D) $1/\{1 + \exp(-X_{i,1}) + \exp(-X_{i,2})\}$.

$\mu(\boldsymbol{x}_i)$ is defined as follows across scenarios A to D: (A) $\sin(\pi X_{i,1} X_{i,2}) + 2(X_{i,3} - 0.5)^2 + X_{i,4} + 0.5X_{i,5}$, (B) $\max\{X_{i,1} + X_{i,2}, X_{i,3}, 0\}$, (C) $2\log\{1 + \exp(X_{i,1} + X_{i,2} + X_{i,3})\}$, (D) $\frac{1}{2}[\max\{X_{i,1} + X_{i,2} + X_{i,3}, 0\} + \max\{X_{i,4} + X_{i,5}, 0\}]$.

$\tau(\boldsymbol{x}_i)$ is defined as follows across scenarios A to D: (A) $(X_{i,1} + X_{i,2})/2$, (B) $X_{i,1} + \log\{1 + \exp(X_{i,2})\}$, (C) constant

Table 2 Simulation study, Brier score

Data	Tobit	BART	RF	Grabit	TOBART	TOBART NP
Normal distribution, $\sigma = 1$						
Friedman (1991)	0.140	0.165	0.195	0.135	0.069	0.070
Friedman (1991) 1 side	0.061	0.058	0.077	0.067	0.032	0.033
Groot and Lucas (2012)	0.287	0.121	0.171	0.158	0.116	0.115
Jacobson and Zou (2024)	0.099	0.104	0.123	0.105	0.102	0.102
Sigrist and Hirnschall (2019)	0.052	0.052	0.053	0.050	0.047	0.047
Skew-t distribution, $location = 1$, $scale = 1$, $\nu = 4$						
Friedman (1991)	0.152	0.173	0.197	0.143	0.083	0.081
Friedman (1991) 1 side	0.074	0.071	0.094	0.070	0.047	0.044
Groot and Lucas (2012)	0.281	0.122	0.172	0.177	0.119	0.119
Jacobson and Zou (2024)	0.098	0.103	0.121	0.141	0.106	0.101
Sigrist and Hirnschall (2019)	0.049	0.049	0.050	0.049	0.047	0.046
Weibull distribution, $shape = 0.5$, $scale = 0.2$						
Friedman (1991)	0.137	0.163	0.197	0.170	0.058	0.053
Friedman (1991) 1 side	0.060	0.059	0.081	0.061	0.028	0.024
Groot and Lucas (2012)	0.290	0.058	0.144	0.116	0.077	0.073
Jacobson and Zou (2024)	0.056	0.061	0.088	0.072	0.072	0.046
Sigrist and Hirnschall (2019)	0.046	0.045	0.046	0.040	0.041	0.038

Data	Soft BART	GP	GP 5 vars	GP Tobit	Soft TOBART	Soft TOBART NP
Normal distribution, $\sigma = 1$						
Friedman (1991)	0.116	0.477	0.120	0.477	**0.055**	**0.055**
Friedman (1991) 1 side	0.033	0.455	0.055	0.456	**0.025**	**0.025**
Groot and Lucas (2012)	0.107	0.241	0.233	0.241	**0.106**	0.107
Jacobson and Zou (2024)	0.100	0.440	0.168	0.444	**0.099**	**0.099**
Sigrist and Hirnschall (2019)	0.050	0.054	0.050	0.055	**0.042**	**0.042**
Skew-t distribution, $location = 1$, $scale = 1$, $\nu = 4$						
Friedman (1991)	0.135	0.489	0.144	0.489	0.064	**0.062**
Friedman (1991) 1 side	0.047	0.459	0.078	0.459	0.034	**0.033**
Groot and Lucas (2012)	0.110	0.251	0.258	0.251	**0.111**	**0.111**
Jacobson and Zou (2024)	0.098	0.444	0.171	0.447	0.098	**0.097**
Sigrist and Hirnschall (2019)	0.048	0.049	0.046	0.050	**0.044**	**0.044**
Weibull distribution, $shape = 0.5$, $scale = 0.2$						
Friedman (1991)	0.120	0.488	0.120	0.488	0.037	**0.032**
Friedman (1991) 1 side	0.031	0.463	0.060	0.463	0.016	**0.012**
Groot and Lucas (2012)	0.050	0.247	0.206	0.247	0.067	**0.065**
Jacobson and Zou (2024)	0.049	0.440	0.144	0.444	0.054	**0.039**
Sigrist and Hirnschall (2019)	0.044	0.045	0.037	0.046	0.036	**0.034**

Minimum values, excluding GP trained with only the 5 relevant variables, are in bold

equal to 1, (D) $\max\{X_{i,1} + X_{i,2} + X_{i,3}, 0\} - \max\{X_{i,4} + X_{i,5}, 0\}$.
The outcome before censoring is generated as:

$$Y_i^* = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)(Z_i - 0.5) + \varepsilon_i \text{ where } \varepsilon_i \sim \mathcal{N}(0, 1)$$

The number of sampled observations is 200. The observed outcome Y_i is censored from below at the 15^{th} percentile of the generated Y_i^* values, and from above at the 85^{th} percentile.

3.4 Treatment effect simulation results

All methods are evaluated in terms of Precision in Estimation of Heterogeneous Effects (PEHE), which is defined as $\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i))^2$. Confidence intervals are evaluated

Table 3 Treatment effect simulation results

Method	Caron et al. (2022) DGP			Friedberg et al. (2020) DGP			Nie and Wager (2021) DGP A		
	PEHE	Cov	Len	PEHE	Cov	Len	PEHE	Cov	Len
BART all	0.943	0.311	0.873	1.301	0.378	0.814	0.058	0.638	0.453
BART uncens	1.345	0.235	0.931	2.772	0.122	0.770	0.122	0.385	0.436
Soft BART all	0.609	0.640	1.491	0.538	0.558	1.325	0.060	0.778	0.616
Soft BART uncens	1.048	0.428	1.412	2.067	0.294	1.139	0.145	0.387	0.445
CF	0.863	0.415	0.891	0.953	0.365	0.784	0.259	0.767	0.683
LLF	0.850	0.462	1.069	0.876	0.443	1.037	0.281	0.791	0.776
Grabit	0.722	–	–	0.512	–	–	0.128	–	–
TOBART	0.536	0.863	1.917	0.329	0.900	1.905	**0.043**	0.993	1.121
TOBART-NP	0.528	**0.935**	2.470	0.314	**0.970**	2.318	**0.043**	1	1.655
Soft TOBART	**0.390**	**0.935**	1.970	0.218	0.905	1.738	**0.043**	**0.968**	0.867
Soft TOBART-NP	0.394	0.972	2.657	**0.211**	0.9855	2.261	**0.043**	1	1.471

Method	Nie and Wager (2021) DGP B			Nie and Wager (2021) DGP C			Nie and Wager (2021) DGP D		
	PEHE	Cov	Len	PEHE	Cov	Len	PEHE	Cov	Len
BART all	1.041	0.2912	0.661	0.126	0.283	0.560	1.628	0.360	0.502
BART uncens	1.435	0.1972	0.579	0.319	0.048	0.591	1.699	0.346	0.461
Soft BART all	0.548	0.577	0.976	0.102	0.516	0.700	1.102	0.565	1.181
Soft BART uncens	0.966	0.398	0.929	0.265	0.291	0.804	1.674	0.340	0.423
CF	0.650	0.686	1.148	0.555	0.013	0.543	1.165	0.391	0.722
LLF	0.550	0.846	1.570	0.543	0.038	0.563	0.933	0.535	1.096
Grabit	0.650	–	–	0.242	–	–	1.500	–	–
TOBART	0.305	0.902	1.694	0.032	**0.994**	1.192	0.948	0.759	1.913
TOBART-NP	0.316	**0.954**	2.307	0.032	1	2.114	0.937	0.818	2.304
Soft TOBART	**0.145**	0.968	1.480	**0.025**	0.996	0.975	**0.679**	0.807	1.897
Soft TOBART-NP	0.153	0.986	2.231	**0.025**	1	1.975	0.684	**0.864**	2.252

Caron et al. (2022), Friedberg et al. (2020), and Nie and Wager (2021) simulations with censoring from below at 15^{th} percentile and from above at 85^{th} percentile. *PEHE* Precision in Estimation of Heterogeneous Effects, *Cov* average coverage of 95% intervals, *Len* average length of 95% intervals. Minimum PEHE values are in bold. Coverage values closest to 0.95 are in bold.

in terms of average coverage of 95% intervals and average length of intervals.

The results are presented in Table 3. For all DGPs, at least one TOBART method attains lower PEHE than all other methods, often by a large margin. Local Linear Forests (Friedberg et al. 2020) attain similar PEHE to TOBART and TOBART-NP for Nie and Wager (2021) DGP D, which involves partly linear prognostic and treatment effect functions, although soft TOBART is notably more accurate. The average coverages of TOBART and soft TOBART credible intervals for $\tau(\boldsymbol{x}_i)$ are generally much closer to 95% than the coverages of intervals produced by competing methods. TOBART-NP produces very wide credible intervals relative to TOBART. TOBART-NP produces better coverage than TOBART for four DGPs.

4 Data application

For the data application, we consider the same methods as in Sect. 3.2, excluding Gaussian Processes and adding a hurdle model combining linear regression and probit. For each data set, we average results over 10 training-test splits. Each split is defined by taking a random sample of floor($0.7n$) training observations stratified by censorship status. Categorical variables were encoded as sets of dummy variables. The numbers of observations, covariates, and proportions of censored observations are given in Table 4. Appendix C contains data descriptions with references to original sources.

Table 4 Data application: number of observations (n), number of covariates (p), and proportions censored from below and above

Data Set	n	p	% cens. below	% cens. above
Real censoring				
antibody	330	3	26.1	0
Recon	423	108	11.1	0
Atrazine	48	2	29.2	0
SedPb	42	2	3.6	0
Pollen_Thia	204	4	42.6	0
Missouri	127	3	26.8	0
BostonHousing	506	108	0	3.2
Fake censoring				
ozone	330	8	15	15
ankara	1609	9	15	15
bbb	208	133	15	15
tri	185	58	15	15
ais	202	10	15	15
hatco	100	13	15	15
servo	167	11	15	15
cpu	209	6	15	15
diamonds	308	14	15	15
tec	240	122	15	15

4.1 Data application results

The data application results are presented in Table 5. For most data sets the results are similar across methods, particularly when methods are evaluated in terms of Brier score for predicted probabilities of censoring. Soft TOBART-NP performs best in terms of area under the receiver operating characteristics curve (AUROC). TOBART can give notably lower MSE of outcome predictions relative to other methods for some data sets. Prediction interval coverage is generally similar across methods, although prediction interval length for TOBART-based methods can be notably smaller than for standard BART.

In contrast to the simulation studies above, there is not a clear winning method in Table 5. Although censored outcome models have been applied to these data sets in previous work, perhaps other models are more suitable for some data sets. This is evidenced by the fact that for many data sets the combination of probit and a linear model outperforms Tobit. Therefore for some data sets zero inflated, hurdle, or sample selection models might be more appropriate. For the data sets on which Tobit outperforms probit and a linear model in terms of MSE, namely `Recon` and `Atrazine`, the best method is Soft TOBART. The TOBART models also notably outperform other methods when applied to the `BostonHousing` and `Missouri` data sets.

A lesson from this study is that it is important to select the appropriate model for the data set. The TOBART and Grabit methods are designed for the same form of DGPs, therefore it is arguably fairer to compare these two methods. Soft TOBART produces lower MSE predictions than Grabit across almost all data sets.[18] Nonetheless, the results are less impressive than those observed in the simulation study. Possible explanations for this include slow mixing of the TOBART Markov Chain, small sample sizes for some data sets, and very small or very large proportions of censored outcomes.

Censored outcome models are intended for prediction of latent outcomes, and it is not possible to evaluate these predictions by only using censored outcome data. In order to demonstrate the usefulness of TOBART for modelling of latent outcomes using real data, we artificially censor outcomes from some real datasets. The data sets summarized in Table 4 were previously studied by Kapelner and Bleich (2016) (`Ozone` and `Ankara`) and Linero and Yang (2018) (all other data sets).[19] We introduce fake censoring from below and above at the 15^{th} and 85^{th} percentiles respectively. Therefore the true values of "censored" outcomes are known.

The results in Table 6 suggest that TOBART-based methods can produce much more accurate predictions of latent outcomes than competing methods even if differences in

[18] Potentially Grabit could produce better results with more hyperparameter tuning, although this would be computationally costly.

[19] The outcomes for all data except `Ozone` and `Ankara` were transformed to resemble normally distributed using code provided by Linero and Yang (2018).

Table 5 Data application results: mean squared error of outcome predictions relative to TOBART; Brier score and AUROC for predicted probabilities of censoring; 95% posterior predictive interval average coverage and length. Average over 10 random splits into 70% training data 30% test data. Minimum MSE, minimum Brier score, maximum AUROC, and coverage values closest to 0.95 are in bold

Data Set	TOBART	TOBART NP	Soft TOBART	Soft TOBART NP	Grabit	BART	RF	Probit+LM	Tobit
MSE relative to TOBART									
antibody	1.00	0.98	1.01	0.97	**0.96**	1.00	0.99	0.99	0.99
Recon	1.00	0.74	**0.63**	0.77	0.66	0.74	0.68	0.85	0.81
Atrazine	1.00	0.98	**0.97**	0.98	0.99	1.01	1.01	1.02	1.00
SedPb	1.00	0.97	0.94	0.95	1.03	**0.93**	1.08	1.02	1.02
Pollen_Thia	1.00	**0.99**	1.00	1.00	**0.99**	**0.99**	**0.99**	**0.99**	1.00
Missouri	1.00	0.66	**0.58**	0.73	0.88	0.76	0.77	0.93	0.93
BostonHousing	**1.00**	1.49	1.02	1.06	1.38	1.18	1.31	113.37	125.10
Brier score									
antibody	0.22	0.20	0.22	0.20	0.23	0.19	0.19	0.19	0.22
Recon	0.17	0.12	0.17	0.11	0.15	**0.09**	**0.09**	0.22	0.24
Atrazine	0.11	0.12	0.19	0.17	0.22	0.01	0.02	**0.00**	0.10
SedPb	**0.04**	**0.04**	0.05	0.05	0.07	**0.04**	0.05	0.05	0.05
Pollen_Thia	0.17	0.16	0.18	0.16	0.22	**0.13**	**0.13**	**0.13**	0.17
Missouri	0.17	0.15	0.17	0.14	0.18	**0.13**	**0.13**	0.18	0.22
BostonHousing	0.02	0.03	**0.01**	0.02	0.02	0.03	0.02	0.13	0.03
AUROC									
antibody	0.49	0.57	0.49	**0.58**	0.44	**0.58**	**0.58**	**0.58**	0.50
Recon	0.71	0.83	0.83	**0.85**	0.76	**0.85**	0.75	0.57	0.74
Atrazine	**1.00**	**1.00**	**1.00**	**1.00**	0.50	**1.00**	**1.00**	**1.00**	**1.00**
SedPb	0.77	0.77	0.77	**0.78**	0.74	0.60	0.52	0.39	0.77
Pollen_Thia	0.86	**0.87**	0.86	**0.87**	**0.87**	0.86	0.86	0.86	0.86
Missouri	0.90	0.90	**0.91**	**0.91**	0.86	0.90	0.87	0.70	0.73
BostonHousing	0.97	0.91	**0.99**	**0.99**	0.96	0.96	**0.99**	0.74	0.52
95% prediction interval coverage									
antibody	0.96	0.98	0.96	0.98		**0.95**			
Recon	0.74	0.98	0.97	0.98		**0.96**			
Atrazine	0.97	**0.96**	0.97	**0.96**		**0.96**			
SedPb	0.90	0.90	**0.94**	0.92		0.92			
Pollen_Thia	0.98	**0.97**	0.98	**0.97**		**0.97**			
Missouri	**0.97**	**0.97**	0.99	0.98		0.98			
BostonHousing	**0.95**	0.96	0.96	0.96		0.94			
95% prediction interval length									
antibody	5.91	7.92	5.74	7.91		8.56			
Recon	8.63	31.48	21.50	33.55		33.47			
Atrazine	18.51	12.85	22.63	13.51		39.60			
SedPb	13.89	13.61	15.28	14.67		16.06			
Pollen_Thia	11.27	8.80	11.38	8.97		20.02			
Missouri	11.95	10.79	12.00	13.07		19.89			
BostonHousing	10.08	13.94	10.68	11.33		10.93			

Table 6 Fake censoring data application results: MSE of outcome and latent outcome predictions relative to TOBART; latent outcome 95% posterior predictive interval average coverage and length. Average over 5 random splits into 70% training data 30% test data. Minimum MSE, and coverage values closest to 0.95 are in bold

Data Set	TOBART	TOBART NP	Soft TOBART	Soft TOBART NP	Grabit	BART	RF
Observed censored outcome MSE relative to TOBART							
ozone	**1.00**	1.02	1.04	1.04	1.77	1.03	**1.00**
ankara	1.00	1.04	**0.79**	**0.79**	1.27	1.03	1.29
bbb	1.00	1.02	0.98	1.01	1.09	0.90	**0.84**
tri	1.00	0.94	1.00	0.93	1.39	0.26	**0.23**
ais	**1.00**	**1.00**	**1.00**	**1.00**	1.01	1.01	1.09
hatco	1.00	1.14	**0.84**	1.08	1.56	1.22	1.05
servo	1.00	1.11	**0.74**	0.98	1.90	0.97	1.42
cpu	1.00	1.02	0.92	0.94	1.35	0.80	**0.78**
diamonds	1.00	1.75	0.39	1.89	2.64	**0.15**	2.77
tec	1.00	1.29	**0.42**	0.66	1.61	1.39	3.14
Latent outcome MSE relative to TOBART							
ozone	**1.00**	**1.00**	1.03	1.03	1.91	1.13	1.09
ankara	1.00	1.06	**0.84**	0.87	2.39	3.46	4.14
bbb	**1.00**	**1.00**	1.01	1.02	1.26	1.17	1.10
tri	1.00	1.00	1.02	1.03	1.00	1.07	**0.96**
ais	1.00	1.00	1.01	1.02	1.00	**0.97**	0.98
hatco	**1.00**	1.13	1.13	1.24	2.29	1.31	1.30
servo	**1.00**	1.06	1.29	1.17	3.15	1.47	2.32
cpu	1.00	1.02	**0.83**	0.84	1.54	1.42	1.55
diamonds	**1.00**	1.76	2.39	5.17	5.71	5.29	24.91
tec	1.00	1.16	**0.62**	0.91	2.13	1.77	4.38
Latent outcome 95% prediction interval coverage							
ozone	0.97	**0.94**	**0.96**	**0.94**		0.85	
ankara	0.99	0.98	**0.95**	0.94		0.74	
bbb	0.91	**0.93**	0.91	0.92		0.79	
tri	**0.92**	**0.92**	**0.92**	**0.92**		0.84	
ais	**0.95**	**0.95**	0.96	**0.95**		0.83	
hatco	**0.95**	0.97	0.92	0.97		0.84	
servo	**0.95**	0.97	0.90	0.97		0.86	
cpu	**0.96**	**0.96**	0.97	0.98		0.80	
diamonds	1.00	1.00	**0.97**	1.00		0.71	
tec	0.98	0.99	**0.93**	0.99		0.73	
Latent outcome 95% prediction interval length							
ozone	20.86	18.28	20.57	18.77		13.79	
ankara	11.47	9.00	6.71	6.49		5.33	
bbb	1.85	2.09	1.84	2.04		1.46	
tri	0.45	0.46	0.49	0.49		0.36	
ais	2.28	2.25	2.35	2.29		1.64	
hatco	1.70	2.40	1.48	2.43		1.39	
servo	1.36	1.78	1.24	1.82		1.22	
cpu	1.82	2.06	1.69	1.97		1.24	
diamonds	0.41	0.82	0.38	0.93		0.14	
tec	2.07	2.88	1.19	2.20		1.17	

MSE of observed outcome predictions are relatively small. Latent outcome posterior predictive interval coverage is generally much better for TOBART than for BART. This is unsurprising, as the BART posterior predictive intervals are not designed for latent outcomes.

5 Conclusion

Type I TOBART produces accurate predictive probabilities of censoring, predictions of outcomes, and treatment effect estimates. TOBART-NP, gives better uncertainty quantification for some simulated DGPs. Advantages of TOBART over competing methods include the fact that hyperparameter tuning is not required, and the straightforward combination of the method with other variations on BART to allow for smooth DGPs and sparsity (Linero and Yang 2018).

Supplementary information

The online supplementary appendix contains (A) additional simulation study results, (B) additional data application results, and (C) implementation details and parameter settings.

Appendix A TOBART-1 Gibbs sampler

A.1 Gibbs sampler algorithms

For completeness of exposition, we describe here the full conditional samples from $p((T_k, M_k)|\{(T_j, M_j)\}_{j\neq k}, \sigma, \boldsymbol{y}^*)\, k = 1, \ldots, m$ introduced by Chipman et al. (2010) in Algorithm 1. This sample is separated into a Metropolis-Hastings draw of $p(T_k|\{(T_j, M_j)\}_{j\neq k}, \sigma, \boldsymbol{y}^*)\ k = 1, \ldots, m$ following by a closed form (multivariate normal) draw from $p(M_k|T_k, \{(T_j, M_j)\}_{j\neq k}, \sigma, \boldsymbol{y}^*)\, k = 1, \ldots, m$).

The TOBART and TOBART-NP Gibbs samplers are described in algorithm 2.

A.2 TOBART-NP out of sample distribution of the error

For test data predictive intervals, we may sample TOBART-NP error term values for out of sample observations. Van Hasselt (2011) describes the sampling method as follows. Let $\tilde{n}$ denote the index of an out of sample observation. At iteration $t \in \{1, ..., T\}$ of the Markov chain, given $\{\vartheta_{i,t}\}_{i=1}^n$, generate an out-of-sample value $\vartheta_{\tilde{n},t}$ according to:

$$\vartheta_{\tilde{n},t} \begin{cases} = \vartheta_{i,t} \text{ with probability } \frac{1}{\alpha+n} \text{ for } i = 1, ..., n \\ \sim G_0 \text{ with probability } \frac{\alpha}{\alpha+n} \end{cases}$$

An estimate of the posterior predictive distribution of the error is

$$\hat{f}(u|y, s) = \frac{1}{T}\sum_{t=1}^{T} f(u|\mu_{\tilde{n},t}, \sigma^2_{\tilde{n},t})$$

Also, samples $u_{\tilde{n}}^{(t)}$ can be made from $\mathcal{N}\left(\mu_{\tilde{n},t}, \sigma^2_{\tilde{n},t}\right)$ for each iteration t of the MCMC sampler.

Appendix B Treatment effect with censored outcomes; additional details

A model naively trained on the full dataset with censoring estimates the following:

$$\begin{aligned}
&E[Y_i(1) - Y_i(0)|\boldsymbol{x}_i] = \tau(\boldsymbol{x}_i) \\
&\times\left(\Phi\left(\frac{b-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)}{\sigma}\right) - \Phi\left(\frac{a-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)}{\sigma}\right)\right) \\
&+\mu(\boldsymbol{x}_i)\left(\Phi\left(\frac{b-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)}{\sigma}\right)\right. \\
&-\Phi\left(\frac{a-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)}{\sigma}\right) - \Phi\left(\frac{b-\mu(\boldsymbol{x}_i)}{\sigma}\right) \\
&\left.+\Phi\left(\frac{a-\mu(\boldsymbol{x}_i)}{\sigma}\right)\right) \\
&+\sigma\left(\phi\left(\frac{a-(\mu(\boldsymbol{x}_i)+\tau(\boldsymbol{x}_i)))}{\sigma}\right)\right. \\
&-\phi\left(\frac{b-(\mu(\boldsymbol{x}_i)+\tau(\boldsymbol{x}_i))}{\sigma}\right) - \phi\left(\frac{a-\mu(\boldsymbol{x}_i)}{\sigma}\right) \\
&\left.+\phi\left(\frac{b-\mu(\boldsymbol{x}_i)}{\sigma}\right)\right) \\
&+a\left(\Phi\left(\frac{a-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)}{\sigma}\right) - \Phi\left(\frac{a-\mu(\boldsymbol{x}_i)}{\sigma}\right)\right) \\
&+b\left(\Phi\left(\frac{b-\mu(\boldsymbol{x}_i)}{\sigma}\right) - \Phi\left(\frac{b-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)}{\sigma}\right)\right)
\end{aligned}$$

Appendix C Description of data sets

- `antibody`: Measles vaccine response data set obtained from Moulton and Halsey (1995), originally from Job et al. (1991). The outcome is an antibody measurement censored from below at 0.1, $n = 330$ and $p = 3$.
- `Recon`: Atrazine concentrations in streams throughout the Midwestern United States. Data available in the `R` package `NADA` (Helsel 2005) sourced from Mueller et al.

Algorithm 1 Full conditional sampler for Bayesian trees (Chipman et al. 2010)

Input: (Latent) outcome values $\boldsymbol{y}^*$, covariates $\boldsymbol{X}$, constant error variance for TOBART σ^2, trees and terminal node parameters $\{(T_k, M_k)\}_{k=1}^m$ or observation-specific error means and variances for TOBART-NP $\{\gamma_i, \sigma_i^2\}_{i=1}^n$.

for $k = 1, \ldots m$ **do**

1. Create partial residuals. For TOBART $R_{k,i} = y_i^* - \sum_{s \neq k} g_k(x_i)$. For TOBART-NP $R_{k,i} = y_i^* - \gamma_i - \sum_{s \neq k} g_k(x_i)$.

2. Draw from $T_k | \{(T_j, M_j)\}_{j \neq k}, \sigma, \boldsymbol{y}^*$ using a Metropolis-Hastings Sampler. Propose T_k' using a `PRUNE`, `CHANGE`, or `GROW` proposal.

The `PRUNE` proposal removes (uniformly at random) a split that results in two terminal nodes from T_k. i.e. remove the children nodes from a random internal node without grandchildren.

The `CHANGE` proposal uniformly at random selects an internal node without grandchildren from tree T_k and randomly samples a new splitting variable (uniformly) and splitting point (uniformly).

The `GROW` proposal uniformly at random selects a terminal node of tree T_k (with some minimum number of observations) and uniformly at random samples a new splitting variable and splitting point to create tree T_k'.

3. The log-likelihoods of trees T_k and T_k', marginalizing out the terminal node parameters M_k (or M_k'), are standard weighted linear regression log-likelihoods $\log\left(p(\mathbf{R}_k | T_k, \{\sigma_i^2\}_{i=1}^n)\right) =$

$$\sum_{l=1}^{b_k} \left[-|n_{kl}| \log\left(\sqrt{2\pi}\right) - \sum_{i \in n_{kl}} \log(\sigma_i) - \frac{1}{2} \log\left(1 + \sum_{i \in n_{kl}} \frac{\sigma_\mu^2}{\sigma_i^2}\right) - \sum_{i \in n_{kl}} \frac{R_{ik}^2}{2\sigma_i^2} + \frac{\left(\sum_{i \in n_{kl}} \frac{R_{ik}}{\sigma_i^2}\right)^2}{2\left(\frac{1}{\sigma_\mu^2} + \sum_{i \in n_{kl}} \frac{1}{\sigma_i^2}\right)} \right]$$

b_k is the number of terminal nodes in T_k. There are n_{kl} observations in the l^{th} terminal node.

4. The Metropolis-Hastings step accepts the new tree proposal T_k' with probability equal to $\frac{p(T_k' \to T_k)}{p(T_k \to T_k')} \frac{p(\mathbf{R}_k | T_k', \{\sigma_i^2\}_{i=1}^n)}{p(\mathbf{R}_k | T_k, \{\sigma_i^2\}_{i=1}^n)} \frac{p(T_k')}{p(T_k)}$, where $p(T_k \to T_k')$ is the probability of proposing tree T_k' given the current tree is T_k.

The terminal node parameters $M_k = (\mu_{1k}, \ldots, \mu_{b_k k})'$ are drawn from the full conditional $p(M_k | T_k, \{(T_j, M_j)\}_{j \neq k}, \sigma, \boldsymbol{y}^*)$ $k = 1, \ldots, m$), which separates into independent univariate normal draws. For $\ell = 1, \ldots, b_k$ sample from

$$p(\mu_{\ell k} | \{R_{k,i}\}_{i \in \ell}, \{\sigma_i^2\}_{i=1}^n, \sigma_\mu^2) = \mathcal{N}(\mu_\ell | \tilde{\mu}_\ell, \tilde{\sigma}_\ell^2) \ , \ \tilde{\sigma}_\ell^2 = \frac{1}{\frac{1}{\sigma_\mu^2} + \sum_{i \in \ell} \frac{1}{\sigma_i^2}}, \tilde{\mu}_\ell = \tilde{\sigma}_\ell^2 \left(\sum_{i \in \ell} \frac{R_{k,i}}{\sigma_i^2} \right)$$

end for

(1997). The outcome is Atrazine concentration, censored from below at 0.05, $n = 423$ and $p = 108$.

- `Atrazine`: Atrazine concentrations in Nebraska ground water. Data available in the `R` package `NADA` (Helsel 2005) sourced from Junk et al. (1980). The outcome is Atrazine concentration, censored from below at 0.01, $n = 48$ and $p = 2$.
- `SedPb`: Lead concentrations in stream sediments before and after wildfires. Data available in the `R` package `NADA` (Helsel 2005). The outcome is Lead concentration, censored from below at 4, $n = 82$ and $p = 2$.
- `Pollen_Thia`: Thiamethoxam concentrations in pollen from the Ontario Pollen Monitoring Network. Data available in the `R` package `NADA2` (Helsel 2005) sourced from Junk et al. (1980). The outcome is Thiamethoxam concentration, censored from below at 0.05, $n = 204$ and $p = 4$.
- `Missouri`: TCDD concentrations used by Zirschky and Harris (1986) in a geostatistical analysis of Hazardous waste data in Missouri. Data available in the `R` package `CensSpatial` (Helsel 2005). The outcome is censored from below at 0.1, $n = 127$ and $p = 3$.
- `BostonHousing`: Housing data for 506 census tracts of Boston from the 1970 census available in the `R` package `mlbench` (Leisch and Dimitriadou 2021), sourced from Harrison Jr and Rubinfeld (1978), Pace and Gilley (1997). Outcome is median value of owner-occupied homes in USD 1000's censored from above at 50. $n = 506$, $p = 108$.
- `ankara`: Mean temperature and other weather data for Ankara from 1994 to 1998.
- `ozone`: Ozone concentrations and weather data for Los Angeles in 1976.
- `bbb`: Log of the ratio of the concentration of a compound in the brain and in the blood (Mente and Lombardo 2005). Availeble in `R` package `caret`.
- `tri`, `ais`, `hatco`, `servo`, `cpu`, `diamonds`, `tec`: See Kim et al. (2007) for data sources.

Appendix D Comparison of simulation study computational times

This appendix contains a comparison of average computational times, in minutes, across iterations for each DGP of the simulation study. The times for BART, RF, and Soft BART do not contain the time taken to train separate models for estimation of binary censoring probabilities. The Grabit time does

Algorithm 2 TOBART and TOBART-NP Gibbs Sampler

Input: Number of MCMC iterations, B. Outcome values $\boldsymbol{y}$, covariates $\boldsymbol{X}$, censoring limits a and b, hyperparameter values α, β, κ, ν, λ (or q). For TOBART-NP, hyperparameters γ_0 and k_0.

0. Set initial values of f_y, intialize all m trees as stumps with no splits. For TOBART, initialize σ. For TOBART-NP, set initial values of γ_i, σ_i.

for $b = 1, \ldots B$ **do**

1. Draw latent variable from $\boldsymbol{y}^*|\gamma_i, \sigma_i, f_y, \boldsymbol{y}$

$$y_i^* \sim \begin{cases} \mathcal{TN}_{(-\infty,a]}\left(f(x_i) + \gamma_i, \sigma_i^2\right) \text{ if } y_i = a \\ y_i \text{ if } a < y_i^* < b \\ \mathcal{TN}_{(b,\infty]}\left(f(x_i) + \gamma_i, \sigma_i^2\right) \text{ if } y_i = b \end{cases}$$

2. Draw the sum of trees $f_y|\boldsymbol{y}^*, \gamma_i, \sigma_i$, by applying Algorithm 1.

3. [TOBART] Draw σ^2 from an inverse gamma distribution $IG\left(\frac{n+\nu}{2}, \frac{\sum_{i=1}^n (y_i^* - \hat{y}_i)^2 + \nu\lambda}{2}\right)$.

3. [TOBART-NP]

for $i = 1, ..., n$ **do**

Sample from $(\gamma_i, \sigma_i)|\boldsymbol{y}^*, \{(\gamma_k, \sigma_k), k \neq i\}, f_y$. This follows the procedure described in Escobar (1994), Escobar and West (1995), Escobar and West (1998). For a similar context see step 3 of algorithm 1 of Chib and Greenberg (2010). Define $\vartheta_i = (\gamma_i, \sigma_i)$. Let $\vartheta_{-i} = \{(\gamma_{-i}, \sigma_{-i})\} = \{(\gamma_k, \sigma_k), k \neq i\}$ be the set of pairs of parameters excluding (γ_i, σ_i). Let $(\gamma_{-i,r}^*, \sigma_{-i,r}^*)$ with $r = 1, ..., k_{-i}$ be the set of k_{-i} unique pairs of parameters in the set $\{(\gamma_{-i}, \sigma_{-i})\}$.

(i) Calculate $q_{i,0} = \alpha t_\nu\left(y_i^* | f_y(x_i), \lambda(1 + \frac{1}{k_0})\right)$ where t_ν denotes the probability density function of a t distribution with ν degrees of freedom.

(ii) For $r = 1, ..., k_{-i}$ calculate $q_{-i,r} = n_{-i,r}\mathcal{N}\left(y_i^* | \gamma_{-i,r}^* + f_y(x_i), (\sigma_{-i,r}^*)^2\right)$.

(iii) Scale $q_{i,0}$ and $q_{-i,r}$ to $\tilde{q}_{i,0} = \frac{q_{i,0}}{q_{i,0} + \sum_{r=1}^{k_{-i}} q_{-i,r}}$, and , $\tilde{q}_{-i,r} = \frac{q_{-i,r}}{q_{i,0} + \sum_{s=1}^{k_{-i}} q_{-i,s}}$ for $r = 1, ..., k_{-i}$.

(iv) Draw $r' \in \{0, 1, ..., k_{-i}\}$ from a categorical distribution with probabilities $\{\tilde{q}_{i,0}, \tilde{q}_{i,1}, ..., \tilde{q}_{i,k_{-i}}\}$

(v) $$\begin{cases} \text{If } r' \neq 0, \text{ set } (\gamma_i, \sigma_i) = (\mu_{r'}, \sigma_{r'}) \\ \text{If } r' = 0, \text{ draw } \sigma_i^2 \sim \mathcal{IG}\left(\frac{\nu+1}{2}, \frac{\nu\lambda}{2} + \frac{(y_i^* - f_y(x_i))^2}{2\left(1+\frac{1}{k_0}\right)}\right) \text{ then } \gamma_i|\sigma_i^2 \sim \mathcal{N}\left(\frac{1}{k_0+1}(y_i^* - f_y(x_i)), \frac{\sigma_i^2}{k_0+1}\right) \end{cases}$$

end for

4. [TOBART-NP] The following mixing step speeds up convergence of the Markov chain. Steps of this form were introduced by Bush and MacEachern (1996) and West et al. (1994).

Let n_j denote the number of observations in cluster j, $N_j = \{i : \varrho_i = j\}$ where the variable ϱ_i equals the index of the cluster to which observation i belongs. Let $u_i = y_i^* - f_y(x_i)$ and let $\bar{u}^{(j)} = \frac{\sum_{i \in N_j} u_i}{n_j}$ be the mean of u_i values for all observations in cluster N_j.

Note that $p(\gamma_j^*, \sigma_j^* | \boldsymbol{y}^*, f_y) \propto \prod_{i=1}^{n_j} \mathcal{N}\left(y_i^* | f_y(x_i) + \gamma_j, \sigma_j^2\right) \mathcal{N}\left(\gamma_j | 0, \frac{\sigma_j^2}{k_0}\right) \mathcal{IG}\left(\sigma_j^2 | \frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$.

A standard conjugacy result implies that we can sample $\sigma_j^{*2} \sim \mathcal{IG}\left(\frac{\nu+n_j}{2}, \frac{\nu\lambda}{2} + \frac{1}{2}\sum_{\forall i \in N_j}(u_i - \bar{u}^{(j)})^2 + \frac{n_j k_0}{k_0+n_j}\frac{(\bar{u}^{(j)})^2}{2}\right)$, $\gamma_j^* \sim \mathcal{N}\left(\frac{n_j \bar{u}^{(j)}}{k_0+n_j}, \frac{(\sigma_j^*)^2}{k_0+n_j}\right)$

5. [TOBART-NP] Sample an auxiliary variable $\kappa \sim Beta(\alpha + 1, n)$ and sample α from the mixture distribution

$\alpha|k \sim p_\kappa \text{Gamma}(c_1 + k, c_2 - \log\kappa) + (1 - p_\kappa)\text{Gamma}(c_1 + k - 1, c_2 - \log\kappa)$

where k is the current number of mixture components, i.e. unique elements of $\{\vartheta_i\}_{i=1}^n = \{\gamma_i, \sigma_i\}_{i=1}^n$. p_κ is the mixing probability, set so that $\frac{p_\kappa}{1-p_\kappa} = \frac{c_1+k-1}{n(c_2-\log\kappa)}$.

If the prior on α is the prior applied by George et al. (2019), McCulloch et al. (2021) and Conley et al. (2008), then samples are obtained from $\alpha|k$ by noting that $p(\alpha|k) \propto p(k|\alpha)p(\alpha) \propto \alpha^k \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \times (1 - \frac{\alpha-\alpha_{min}}{\alpha_{max}-\alpha_{min}})^\psi$ (Antoniak 1974). A sample can be obtained by discretizing the support and making a multinomial draw. McCulloch et al. (2021) use an equally spaced grid of 100 values from α_{min} to α_{max}.

end for

not contain the considerable time required for parameter tuning by 5-fold cross-validation (the model was re-trained 135 times in each fold for different parameter settings).

The Gaussian Process `MATLAB` code written by Groot and Lucas (2012) was called in `R` via the `R` package `R.matlab`. All other functions were implemented in `R`. Therefore the Gaussian Process times are omitted for fair speed comparison. The Gaussian Process functions were fast, and ran for at most a few minutes per iteration. Tables 7, 8 and 9 contain the computational times for simulations with normal, skew-t, and Weibull distributed errors respectively.

Appendix E Simulation study - TOBART prior settings

This appendix contains a comparison of simulation study results for different prior parameter settings.

For standard TOBART, we present results for different λ parameter settings. Recall that $\sigma^{-2} \sim Ga(\frac{\nu}{2}, \frac{\nu\lambda}{2})$ and λ is set such that the q^{th} quantile of the prior distribution of σ is equal to some estimate $\hat{\sigma}$. For standard BART, $\hat{\sigma}$ is the sample standard deviation of the residuals from a linear

Table 7 Simulation study, normal distribution, $\sigma = 1$, computational times, in minutes

Data	Tobit	BART	RF	Grabit	TOBART	TOBART NP
Friedman (1991)	0.028	6.800	9.912	0.675	3.142	27.762
Friedman (1991) 1 side	0.035	5.713	10.269	1.747	1.106	56.761
Groot and Lucas (2012)	0.167	6.046	40.916	1.153	33.681	26.193
Jacobson and Zou (2024)	0.030	7.508	13.267	0.364	3.563	11.993
Sigrist and Hirnschall (2019)	0.030	6.262	11.670	0.265	13.910	41.027
		Soft BART		Soft TOBART		Soft TOBART NP
Friedman (1991)		22.214		24.416		109.241
Friedman (1991) 1 side		16.938		34.236		89.566
Groot and Lucas (2012)		30.325		89.028		103.573
Jacobson and Zou (2024)		33.373		9.905		150.921
Sigrist and Hirnschall (2019)		17.918		44.855		97.271

Table 8 Simulation study, Skew-t distribution, computational times, in minutes

	Tobit	BART	RF	Grabit	TOBART	TOBART NP
Friedman (1991)	0.011	6.543	9.505	0.461	2.912	48.585
Friedman (1991) 1 side	0.086	6.802	28.386	0.560	16.317	29.928
Groot and Lucas (2012)	0.026	6.438	11.567	0.352	3.527	20.167
Jacobson and Zou (2024)	0.054	6.813	19.126	0.404	8.625	19.915
Sigrist and Hirnschall (2019)	0.068	7.536	22.446	0.353	10.291	79.636
		Soft BART		Soft TOBART		Soft TOBART NP
Friedman (1991)		18.040		40.479		88.207
Friedman (1991) 1 side		32.417		31.089		117.085
Groot and Lucas (2012)		22.372		17.976		132.344
Jacobson and Zou (2024)		29.291		20.369		131.629
Sigrist and Hirnschall (2019)		36.055		100.532		166.530

Table 9 Simulation study, Weibull distribution, computational times, in minutes

	Tobit	BART	RF	Grabit	TOBART	TOBART NP
Friedman (1991)	0.011	6.705	17.705	0.489	9.170	39.305
Friedman (1991) 1 side	0.094	7.075	30.974	0.634	21.699	21.012
Groot and Lucas (2012)	0.083	6.531	22.959	0.469	16.355	11.189
Jacobson and Zou (2024)	0.123	6.695	39.388	0.657	28.227	30.822
Sigrist and Hirnschall (2019)	0.074	7.479	26.137	0.386	13.289	80.480
		Soft BART		Soft TOBART		Soft TOBART NP
Friedman (1991)		25.684		38.256		103.480
Friedman (1991) 1 side		35.920		22.301		136.298
Groot and Lucas (2012)		33.716		10.948		130.280
Jacobson and Zou (2024)		33.999		12.964		120.808
Sigrist and Hirnschall (2019)		39.141		12.418		151.876

Table 10 Simulation study, mean squared error. Different Prior calibration settings for error term distribution

Data	TOBART naive sd	TOBART Tobit sd	TOBART cens sd	TOBART lm sd	TOBART NP, $\nu = 3$, Tobit sd	TOBART NP, $\nu = 10$ Tobit sd
normal distribution, $\sigma = 1$						
Friedman (1991)	1.196	1.269	1.162	1.180	1.176	1.154
Friedman (1991) 1 side	1.478	1.510	1.457	1.462	1.518	1.509
Groot and Lucas (2012)	0.638	0.634	0.631	0.644	0.625	0.617
Jacobson and Zou (2024)	0.717	0.717	0.718	0.716	0.720	0.720
Sigrist and Hirnschall (2019)	1.072	1.072	1.072	1.075	1.075	1.074
Skew-t distribution, $location = 1, scale = 1, \nu = 4$						
Friedman (1991)	1.715	1.725	1.648	1.696	1.677	1.597
Friedman (1991) 1 side	2.220	2.307	2.163	2.200	2.079	2.060
Groot and Lucas (2012)	1.170	1.189	1.149	1.170	1.068	1.071
Jacobson and Zou (2024)	1.236	1.242	1.238	1.240	1.186	1.184
Sigrist and Hirnschall (2019)	1.281	1.280	1.279	1.283	1.268	1.267
Weibull distribution, $shape = 0.5, scale = 0.2$						
Friedman (1991)	0.933	0.910	0.871	0.928	0.807	0.811
Friedman (1991) 1 side	1.447	1.396	1.344	1.371	1.146	1.154
Groot and Lucas (2012)	0.771	0.782	0.787	0.846	0.649	0.648
Jacobson and Zou (2024)	0.942	0.920	0.960	0.906	0.697	0.696
Sigrist and Hirnschall (2019)	0.417	0.424	0.426	0.424	0.365	0.362
t distribution, $\nu = 3$						
Friedman (1991)	2.406	2.498	2.459	2.420	2.339	2.369
Friedman (1991) 1 side	3.761	3.786	3.667	3.679	3.629	3.538
Groot and Lucas (2012)	2.073	2.113	2.028	2.105	1.963	1.964
Jacobson and Zou (2024)	2.427	2.401	2.394	2.454	2.301	2.306
Sigrist and Hirnschall (2019)	2.635	2.603	2.615	2.599	2.493	2.498

model. However, a standard linear model does not account for censoring, and therefore may give poor prior calibration.

We consider the following options:

- naive sd: The sample standard deviation of the outcomes without accounting for censoring.
- Tobit sd: The maximum likelihood estimate of the standard deviation of the error term from a linear Tobit model (with covariates).
- cens sd: The maximum likelihood estimate of the standard deviation of the error term from an intercept-only linear Tobit model. This is an estimate of the standard deviation of y^* that adjusts for censoring, assuming normality and no effects of covariates.
- lm sd: The default BART setting. The sample standard deviation of residuals from a linear model.

A limitation of the options that account for censoring is that the estimates rely on the assumption of normality. unsurprisingly, we observe that no setting provides the best results for all DGPs in Table 10. The $\hat{\sigma}$ estimate from an intercept-only Tobit model gives good results. It is generally larger than the estimates from other options and results in a less informative prior. Therefore we apply this option for our main results.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-024-10434-4.

Acknowledgements The author gratefully acknowledges helpful comments from Mikhail Zhelonkin, Chen Zhou, and participants at the Econometric Institute internal seminar.

Author Contributions E.O'N is the sole author of this paper.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Alhamzawi, A.: A new Bayesian elastic net for Tobit regression. J. Phys. Conf. Ser. **1664**(1), 012047 (2020). https://doi.org/10.1088/1742-6596/1664/1/012047

Alhamzawi, R.: Bayesian elastic net Tobit quantile regression. Commun. Stat. Simul. Comput. **45**(7), 2409–2427 (2016). https://doi.org/10.1080/03610918.2014.904341

Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. **2**(6), 1152–1174 (1974). https://doi.org/10.1214/aos/1176342871

Basson, M., Louw, T.M., Smith, T.R.: Variational Tobit Gaussian process regression. Stat. Comput. **33**(3), 64 (2023). https://doi.org/10.1007/s11222-023-10225-3

Bradic, J., Guo, J.: Robust confidence intervals in high-dimensional left-censored regression. (2016) arXiv:1609.07165 [math.ST]

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Bush, C.A., MacEachern, S.N.: A semiparametric Bayesian model for randomised block designs. Biometrika **83**(2), 275–285 (1996). https://doi.org/10.1093/biomet/83.2.275

Cao, F., Ba, S., Brenneman, W.A., Joseph, V.R.: Model calibration with censored data. Technometrics **60**(2), 255–262 (2018). https://doi.org/10.1080/00401706.2017.1345704

Caron, A., Baio, G., Manolopoulou, I.: Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation. J. Comput. Graph. Stat. **31**(4), 1202–1214 (2022). https://doi.org/10.1080/10618600.2022.2067549

Chen, S., Dahl, G.B., Khan, S.: Nonparametric identification and estimation of a censored location-scale regression model. J. Am. Stat. Assoc. **100**(469), 212–221 (2005). https://doi.org/10.1198/016214504000000836

Cheng, J., Small, D.S.: Semiparametric models and inference for the effect of a treatment when the outcome is nonnegative with clumping at zero. Biometrics **77**(4), 1187–1201 (2021). https://doi.org/10.1111/biom.13368

Chib, S.: Bayes inference in the Tobit censored regression model. J. Econom. **51**(1–2), 79–99 (1992). https://doi.org/10.1016/0304-4076(92)90030-U

Chib, S., Greenberg, E.: Additive cubic spline regression with Dirichlet process mixture errors. J. Econom. **156**(2), 322–336 (2010). https://doi.org/10.1016/j.jeconom.2009.11.002

Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees. Ann. Appl. Stat. **4**(1), 266–298 (2010). https://doi.org/10.1214/09-AOAS285

Conley, T.G., Hansen, C.B., McCulloch, R.E., Rossi, P.E.: A semiparametric Bayesian approach to the instrumental variable problem. J. Econom. **144**(1), 276–305 (2008). https://doi.org/10.1016/j.jeconom.2008.01.007

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D.: Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Stat. Sci. **34**(1), 43–68 (2019). https://doi.org/10.1214/18-STS667

Escobar, M.D.: Estimating normal means with a Dirichlet process prior. J. Am. Stat. Assoc. **89**(425), 268–277 (1994). https://doi.org/10.1080/01621459.1994.10476468

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**(430), 577–588 (1995). https://doi.org/10.1080/01621459.1995.10476550

Escobar, M.D., West, M.: Computing nonparametric hierarchical models. In: Dey, D., Müller, P., Sinha, D. (eds.)Practical nonparametric and semiparametric Bayesian statistics, 1–22. New York, NY: Springer New York. (1998) https://doi.org/10.1007/978-1-4612-1732-9_1

Fan, J., Gijbels, I.: Censored regression: local linear approximations and their applications. J. Am. Stat. Assoc. **89**(426), 560–570 (1994). https://doi.org/10.1080/01621459.1994.10476781

Friedberg, R., Tibshirani, J., Athey, S., Wager, S.: Local linear forests. J. Comput. Graph. Stat. **30**(2), 503–517 (2020). https://doi.org/10.1080/10618600.2020.1831930

Friedman, J.H.: Multivariate Adaptive Regression Splines. Ann. Stat. **19**(1), 1–67 (1991). https://doi.org/10.1214/aos/1176347963

Gammelli, D., Peled, I., Rodrigues, F., Pacino, D., Kurtaran, H.A., Pereira, F.C.: Estimating latent demand of shared mobility through censored Gaussian processes. Transp. Res. Part C Emerg. Technol. **120**, 102775 (2020). https://doi.org/10.1016/j.trc.2020.102775

Gammelli, D., Rolsted, K.P., Pacino, D., Rodrigues, F.: Generalized multi-output Gaussian process censored regression. Pattern Recogn. **129**, 108751 (2022). https://doi.org/10.1016/j.patcog.2022.108751

George, E., Laud, P., Logan, B., McCulloch, R., Sparapani, R.: Fully nonparametric Bayesian additive regression trees. In: Jeliazkov, I., Tobias, J.L. (eds.) Topics in identification, limited dependent variables, partial observability, experimentation, and flexible modeling: part B, Volume 40, 89–110. Emerald Publishing Limited. (2019) https://doi.org/10.1108/S0731-90532019000040B006

Groot, P., Lucas, P.J.: Gaussian process regression with censored data using expectation propagation. In: Proceedings of the 6th European workshop on probabilistic graphical models, 115–122 (2012)

Hahn, P.R., Dorie, V., Murray, J.S.: Atlantic causal inference conference (ACIC) data analysis challenge 2017. (2019) arXiv:1905.09515 [stat.ME]

Harrison, D., Jr., Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air. J. Environ. Econ. Manag. **5**(1), 81–102 (1978). https://doi.org/10.1016/0095-0696(78)90006-2

Helsel, D.R.: Nondetects and data analysis. Statistics for censored environmental data Statistics in Practice. Wiley-Interscience (2005)

Heuchenne, C., Van Keilegom, I.: Location estimation in nonparametric regression with censored data. J. Multivar. Anal. **98**(8), 1558–1582 (2007). https://doi.org/10.1016/j.jmva.2007.03.008

Heuchenne, C., Van Keilegom, I.: Estimation in nonparametric location-scale regression models with censored data. Ann. Inst. Stat. Math. **62**(3), 439–463 (2010). https://doi.org/10.1007/s10463-009-0219-3

Huang, H.: Novel semi-parametric Tobit additive regression models. (2021) arXiv:2107.01497 [stat.ME]

Huang, H., Tang, Y., Li, Y., Liang, H.: Estimation in additive models with fixed censored responses. J. Nonparametr. Stat. **31**(1), 131–143 (2019). https://doi.org/10.1080/10485252.2018.1537441

Jacobson, T., Zou, H.: High-dimensional censored regression via the penalized Tobit likelihood. J. Bus. Econ. Stat. **42**(1), 286–297 (2024). https://doi.org/10.1080/07350015.2023.2182309

Ji, Y., Lin, N., Zhang, B.: Model selection in binary and Tobit quantile regression using the Gibbs sampler. Comput. Stat. Data Anal. **56**(4), 827–839 (2012). https://doi.org/10.1016/j.csda.2011.10.003

Job, J.S., Halsey, N.A., Boulos, R., Holt, E., Farrell, D., Albrecht, P., Brutus, J.R., Adrien, M., Andre, J., Chan, E., Kissinger, P., Boulos, C.: Successful immunization of infants at 6 months of age with

high dose Edmonston-Zagreb measles vaccine. Pediatr. Infect. Dis. J. **10**(4), 303–311 (1991). https://doi.org/10.1097/00006454-199104000-00008

Junk, G.A., Spalding, R.F., Richard, J.J.: Areal, vertical, and temporal differences in ground water chemistry: II. Organic constituents. J. Environ. Qual. **9**(3), 479–483 (1980). https://doi.org/10.2134/jeq1980.00472425000900030031x

Kapelner, A., Bleich, J.: bartMachine: machine learning with Bayesian additive regression trees. J. Stat. Softw. **70**(4), 1–40 (2016). https://doi.org/10.18637/jss.v070.i04

Kim, H., Loh, W.Y., Shih, Y.S., Chaudhuri, P.: Visualizable and interpretable regression models with good prediction power. IIE Trans. **39**(6), 565–579 (2007). https://doi.org/10.1080/07408170600897502

Kottas, A., Krnjajić, M.: Bayesian semiparametric modelling in quantile regression. Scand. J. Stat. **36**(2), 297–319 (2009). https://doi.org/10.1111/j.1467-9469.2008.00626.x

Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proc. Natl. Acad. Sci. **116**(10), 4156–4165 (2019). https://doi.org/10.1073/pnas.1804597116

Leisch, F., Dimitriadou, E.: MLbench: machine learning benchmark problems. R foundation for statistical computing. R package version 2.1-3.1 (2021)

Lewbel, A., Linton, O.: Nonparametric censored and truncated regression. Econometrica **70**(2), 765–779 (2002). https://doi.org/10.1111/1468-0262.00304

Linero, A.R., Yang, Y.: Bayesian regression tree ensembles that adapt to smoothness and sparsity. J. R. Stat. Soc. Ser. B Stat Methodol. **80**(5), 1087–1110 (2018). https://doi.org/10.1111/rssb.12293

McConnell, K.J., Lindner, S.: Estimating treatment effects with machine learning. Health Serv. Res. **54**(6), 1273–1282 (2019). https://doi.org/10.1111/1475-6773.13212

McCulloch, R.E., Sparapani, R.A., Logan, B.R., Laud, P.W.: Causal inference with the instrumental variable approach and Bayesian nonparametric machine learning. (2021) arXiv:2102.01199 [stat.ML]

Mente, S.R., Lombardo, F.: A recursive-partitioning model for blood-brain barrier permeation. J. Comput. Aided Mol. Des. **19**, 465–481 (2005). https://doi.org/10.1007/s10822-005-9001-7

Moulton, L.H., Halsey, N.A.: A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics **51**(4), 1570–1578 (1995). https://doi.org/10.2307/2533289

Mueller, D., Ruddy, B., Battaglin, W.: Logistic model of nitrate in streams of the upper-midwestern United States. J. Environ. Qual. **26**, 1223–1230 (1997). https://doi.org/10.2134/jeq1997.00472425002600050005x

Müller, P., van de Geer, S.: Censored linear model in high dimensions. TEST **25**(1), 75–92 (2016). https://doi.org/10.1007/s11749-015-0441-7

Nie, X., Wager, S.: Quasi-oracle estimation of heterogeneous treatment effects. Biometrika **108**(2), 299–319 (2021). https://doi.org/10.1093/biomet/asaa076

Oganisian, A., Mitra, N., Roy, J.A.: A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. Biometrics **77**(1), 125–135 (2021). https://doi.org/10.1111/biom.13244

Pace, R.K., Gilley, O.W.: Using the spatial configuration of the data to improve estimation. J. Real Estate Financ. Econ. **14**(3), 333–340 (1997). https://doi.org/10.1023/A:1007762613901

Rossi, P.: Bayesian non-and semi-parametric methods and applications. Princeton University Press, The Econometric and Tinbergen Institutes Lectures (2014)

Sigrist, F., Hirnschall, C.: Grabit: gradient tree-boosted Tobit models for default prediction. J. Bank. Financ. **102**, 177–192 (2019). https://doi.org/10.1016/j.jbankfin.2019.03.004

Soret, P., Avalos, M., Wittkop, L., Commenges, D., Thiébaut, R.: Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors. BMC Med. Res. Methodol. **18**(1), 1–13 (2018). https://doi.org/10.1186/s12874-018-0609-4

Tobin, J.: Estimation of relationships for limited dependent variables. Econometrica **26**(1), 24–36 (1958). https://doi.org/10.2307/1907382

Van Hasselt, M.: Bayesian inference in a sample selection model. J. Econom. **165**(2), 221–232 (2011). https://doi.org/10.1016/j.jeconom.2011.08.003

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N.H., Gallego, B.: Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Stat. Med. **37**(23), 3309–3324 (2018). https://doi.org/10.1002/sim.7820

West, M., Müller, P., Escobar, M.D.: Hierarchical priors and mixture models, with application in regression and density estimation. In: Smith, A.F.M., Freeman, P.R. (eds.) Aspects of uncertainty: a tribute to D. V. Lindley, 363–386. Chichester; New York: Wiley (1994)

Wu, W., Yeh, M.Y., Chen, M.S.: Deep censored learning of the winning price in the real time bidding. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2526–2535 (2018)

Yu, K., Stander, J.: Bayesian analysis of a Tobit quantile regression model. J. Econom. **137**(1), 260–276 (2007). https://doi.org/10.1016/j.jeconom.2005.10.002

Zhang, J., Li, Z., Song, X., Ning, H.: Deep Tobit networks: A novel machine learning approach to microeconometrics. Neural Netw. **144**, 279–296 (2021). https://doi.org/10.1016/j.neunet.2021.09.003

Zhang, X., Wan, A.T.K., Zhou, S.Z.: Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. J. Bus. Econ. Stat. **30**(1), 132–142 (2012). https://doi.org/10.1198/jbes.2011.10075

Zirschky, J.H., Harris, D.J.: Geostatistical analysis of hazardous waste site data. J. Environ. Eng. **112**(4), 770–784 (1986). https://doi.org/10.1061/(ASCE)0733-9372(1986)112:4(770)