Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics & Economics

Nonparametric Bankruptcy Prediction Using Heteroscedastic Survival BART

Niels van den Heuvel (500840)

Supervisor:	Vennes Schmidt, A.
Second assessor:	Kleen, O.
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This paper explores the application of heteroscedastic Bayesian additive regression trees (HBART) in the context of bankruptcy prediction. The HBART method is extended to accommodate classification using a probit framework. Then, by a simple data reconstruction I obtain a the survival HBART (SHBART) framework. First, I show the performance of HBART compared to BART in different regression contexts. Then, the SHBART methodology is applied and evaluated on bankruptcy data supplied with financial ratios. The results show that the SHBART method, on average, successfully assigns a higher default probability to test cases that actually went bankrupt compared to censored observations.

1 Introduction

Bankruptcy prediction is a statistical practice in financial analysis that seeks to model the plausibility of financially distressed firms filing for bankruptcy. It is of paramount importance for risk-management, financial decision-making, and regulatory compliance. Different attempts have been made to accurately classify or forecast financial distress, yet, despite the seemingly predictable nature of bankruptcy, many filings still come by surprise. Such failures can have severe consequences like financial losses for investors, potential job losses, and even large market disruptions. Therefore, robust prediction models are indispensable in financial decision-making.

There are several advantages that accurate predictions offer stakeholders. First, creditors can assess the solvency of financially exposed firms for decision-making at the intensive and extensive margin. That is, informed creditors can establish more financially favorable terms and conditions for credit loans, and at the same time disregard firms of which the expected loss of default exceeds the potential gains. Second, predictions of financial distress enable investors to tailor their portfolios in order to minimize losses. Finally, policy makers can use accurate bankruptcy predictions to implement policy to moderate the possibility of financial crises.

In the mid 20th century, researchers started to recognize the necessity of bankruptcy prediction models for these applications. One of the earliest attempts can be traced back to Beaver (1966), who expressed the importance of financial ratios as indicators for corporate distress. In his paper, the author employed discriminant analysis and derived a set of ratios, such as liquidity, profitability, and leverage ratios, that exhibit significant differences between viable and bankrupt firms. This work laid the foundation for analyzing the financial conditions of a company through key administrative ratios. Although this approach was highly critized in the following decades, Altman (1968) further defended the use of ratio analysis and applied a discriminant-ratio model, a combination of a set of financial ratios with a multivariate discriminant analysis (MDA) approach to the problem of corporate bankruptcy prediction. Despite its age and simplicity, these models are still part of the statistical toolbox of financial analysts nowadays.

Later, Ohlson (1980) identified three major problems with the MDA model used by these early papers: (i) it imposed restrictive distributional requirements for the predictors, (ii) it produced only a score with little intuitive interpretation, and (iii) it was often used in combination with "matching" procedures, which tends to lead to arbitrary choices. To avoid these problems, Ohlson (1980) employed a conditional logit analysis. Ultimately, this model yielded more

accurate forecasts in most applications (Begley, Ming & Watts, 1996).

In more recent work, Shumway (2001) argues that traditional single-period classification or static models such as those from Beaver (1966), Altman (1968), and Ohlson (1980) are not appropriate for multiple-period bankruptcy data. These models ignore the dynamic nature of corporations and its capability to adapt or change according to its economic environment and financial state. To incorporate these dynamics, Shumway (2001) exploits hazard analysis. Another problem that arises when using static models is failing to acknowledge censored data. When a firm is no longer observed, static models still consider these observations as viable. On the contrary, because it models the time until bankruptcy rather than bankruptcy itself, hazard models are able to appropriately distinguish between censored and viable cases. Using data from between 1962 and 1992, the model demonstrated that about half of the ratios previously found to be statistically significant are not.

Ever since the work of Shumway (2001), comparative analyses corroborating the superior performance of hazard models have been emerging. For example, Chava and Jarrow (2004) show that the specific hazard rate model of Shumway (2001) outperforms the traditional method of Altman (1968) in producing forecasts. They also show that hazard rate models improve considerably when monthly observation intervals are used instead of yearly data. Moreover, using data on listed firms from between 1979 and 2009, Bauer and Agarwal (2014) further demonstrate that this superiority is not exclusive to the US, but is also evident in the UK.

The application of hazard models, such as the one proposed by Shumway (2001), is really at the core of what justifies the methodology of this paper. Namely, I combine the evident merits of the hazard analysis framework with a rigorous nonparametric approach to modelling the default probability. This approach allows for an intuitive interpretation and at the same time imposes few distributional or functional restrictions on the latent default rate.

The main idea of this novel framework is modelling the location parameter of the standard normal distribution associated with the default probability of a firm at a specific time by means of heteroscedastic Bayesian additive regression trees (HBART) (Pratola, Chipman, George & McCulloch, 2020). HBART is a nonparametric Bayesian method for modelling the mean and variance processes associated with a regression model without assumptions about the functional form. It does so by fitting the mean process by a sum-of-trees ensemble and the standard deviation using the product of regression trees. In this way, it is able to capture complex, nonlinear interactions and expands upon its precursor BART (Chipman, George & McCulloch, 2010) by relaxing the assumption of homoscedsaticity. The incorporation of heteroscedasticity provides for a robust model that caputres better the dynamics of the underlying data generating process. The objective of combining this nonparametric statistical technique with survival analysis is to develop a robust bankruptcy prediction model that overcomes the limitations of traditional methods and improves on the accuracy of modelling default risk.

My framework is tested by using administrative data in the form of financial ratios for firms that are either viable (censored) or bankrupt. This paper contributes to existing literature on bankruptcy prediction through two facets: it uses completely nonparametric analysis of the latent hazard probability in combination with graphical diagnostic plots and it allows for heteroscedasticity in the probit latent variable model.

The paper is organized as follows. The survival HBART methodology is introduced in Section 2, along with some supplementary quality metrics. Section 3 discusses replications of a simulation and two empirical examples from Pratola et al. (2020) to compare BART and HBART in a regular regression context, and a bankruptcy prediction exercise. Finally, Section 4 concludes with a short summary and discussion.

2 Methodology

In this section, the methods used in the analysis are introduced step-by-step. First, I start with a brief overview of BART in Section 2.1, and add on to this the HBART methodology in Section 2.2. Section 2.3, discusses the Markov chain Monte Carlo (MCMC) algorithm used to compute the posterior distribution. The methodology of HBART is extended to a probit framework in Section 2.4, after which I detail the application of HBART probit to survival analysis in Section 2.5 by reconstructing the data. Lastly, a precautionary disclaimer is in order; when deriving the posterior distributions for some parameter θ and observations $y = (y_1, \dots, y_n)$, it is common to refer to priors as $\pi(\theta) \equiv p(\theta)$, likelihoods as $p(y \,|\, \theta) \equiv L(y \,|\, \theta)$, and posteriors as $p(\theta \,|\, y) \equiv \pi(\theta \,|\, y)$.

2.1 BART

For the purpose of point-estimation, Chipman et al. (2010) focus on estimating $\mathbb{E}[Y \,|\, \boldsymbol{x}]$ using an unknown mean function $f(\boldsymbol{x})$ and assume a homoscedastic process

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + \sigma Z, \tag{1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_d)$ is a vector of predictors, $\sigma = \mathbb{V}[Y \,|\, \boldsymbol{x}]$, and $Z \sim \mathcal{N}(0, 1)$ an independent random component.

An ensemble of Bayesian binary regression trees is used to model the unknown mean function $f(\boldsymbol{x})$. Let $\boldsymbol{T}$ denote the interior nodes and $\boldsymbol{M}$ the parameters corresponding to each leaf node of a regression tree. As in Chipman et al. (2010), the tree structure $\boldsymbol{T}$ is made up of (v_q, c_q) pairs, where v_q denotes what variable is used and c_q is the threshold, in the sense that $x_{v_q} < c_q$ determines the splitting rule in node q. Then, $\boldsymbol{T} = \{(v_1, c_1), (v_2, c_2), \dots\}$ encodes the tree structure of the interior nodes. Finally, let $n^g = |M|$ denote the number of terminal nodes in some tree, such that $M = \{\theta_1, \dots, \theta_{n^g}\}$. All in all, a tree defined in this way maps some input data $\boldsymbol{x}$ to a parameter through the function $g(\boldsymbol{x}; \boldsymbol{T}, \boldsymbol{M})$.

Using the ensemble of additive regression trees to model $f(\boldsymbol{x})$, the model becomes

$$Y(\boldsymbol{x}) = \sum_{j=1}^{m} g(\boldsymbol{x}; \boldsymbol{T}_j, \boldsymbol{M}_j) + \sigma Z,$$

where $Z_i \sim \mathcal{N}(0, 1)$. Since the parameters obtained from the leaf nodes correspond to the mean, let us relabel the parameters by μ. The priors defined for the parameters in the leaf nodes are $\pi(\mu_{jk}) \sim \mathcal{N}(0, \tau^2)$, where μ_{jk} denotes the k-th leaf node from the j-th tree. Note that setting the expectation of the prior to 0 prescribes the use of mean-centred data. Similarly, the

variance is described by a scaled inverse chi-squared prior, $\sigma^2 \sim \chi^{-2}(\nu, \lambda)$. Finally, trees are drawn according to a stochastic process, where a node at depth d gets assigned children with probability $\alpha(1+d)^{-\beta}$ for $\alpha \in (0,1)$ and $\beta \geq 1$.

2.2 Heteroscedastic BART

Pratola et al. (2020) propose an extension to BART that relaxes the assumption of homoscedasticity. To do so they model both the expectation and the variance, such that the response data is assumed to be generated by

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + s(\boldsymbol{x})Z,$$

where $s^2(\boldsymbol{x}) = \mathbb{V}[Y \,|\, \boldsymbol{x}]$ is an unknown variance function. The proposed variance process is in stark contrast to BART, where $s^2(\boldsymbol{x}) = \sigma$ was assumed.

The HBART method models the mean function using additive regression trees, seemingly identical to (1). The variance function, on the other hand, gets modelled by a different ensemble of Bayesian regression trees: a multiplicative regression tree model,

$$s^2(\boldsymbol{x}) = \prod_{l=1}^{m'} h(\boldsymbol{x}; \boldsymbol{T}_l', \boldsymbol{M}_l'),$$

where $\boldsymbol{T}_l'$ denotes the l-th tree and $\boldsymbol{M}_l' = \{s^2_{l,1}, s^2_{l,2}, \ldots, s^2_{l,n_l^h}\}$ the parameters in the bottom nodes, with $n_l^h = |\boldsymbol{M}_l'|$. Similar to the prior defined on the constant variance term for regular BART, the specified prior distribution for the variance terms is $s_{lk} \sim \chi^{-2}(\nu', \lambda')$, where s^2_{lk} denotes the variance of the k-th leaf node of the l-th tree. The structure of each tree is modelled by an identical stochastic process as the trees used in the mean ensemble. That is, children spawn with probability $\alpha'(1+d)^{-\beta'}$ for some node at depth d in the tree. The same parameter restrictions apply.

2.3 Bayesian Backfitting MCMC Algorithm

For the sake of computing the posterior, Pratola et al. (2020) propose a MCMC algorithm. This is essentially a Gibbs sampler, first drawing the j-th tree conditional on the other parameters and the data $(\boldsymbol{T}_j \,|\, \cdot)$, and subsequently drawing an instance of the parameters in the leaf nodes conditional on this tree and everything else $(\boldsymbol{M}_j \,|\, \boldsymbol{T}_j, \cdot)$. Given the tree and the assumption that the parameters in the leaf nodes are independent, the posterior of the parameters $(\boldsymbol{M}_j \,|\, \boldsymbol{T}_j, \cdot)$ can simply be derived from the product of the likelihood and the prior. Because of the product-of-trees model and the specified prior, the same procedure can be done for the posterior of the ensemble tree model for the variance function. That is, using the aforementioned procedures, the posterior can be calculated as follows

$$p(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}' \,|\, \boldsymbol{Y}, \boldsymbol{x}) \propto L(\boldsymbol{Y} \,|\, \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{x}) \prod_{j=1}^{m} \pi(\boldsymbol{T}_j)\pi(\boldsymbol{M}_j \,|\, \boldsymbol{T}_j) \prod_{l=1}^{m'} \pi(\boldsymbol{T}_l')\pi(\boldsymbol{M}_l' \,|\, \boldsymbol{T}_l'),$$

where $\pi(\boldsymbol{M}_j \,|\, \boldsymbol{T}_j) = \prod_{k=1}^{n_g} \pi(\mu_{jk})$ and $\pi(\boldsymbol{M}'_j \,|\, \boldsymbol{T}'_j) = \prod_{l=1}^{n_h} \pi(s^2_{lk})$ The trees $\boldsymbol{T}_j \,|\, \cdot$ and $\boldsymbol{T}'_l \,|\, \cdot$ can be drawn using a Metropolis-Hastings algorithm as proposed in (Chipman et al., 2010).

The computation of the MCMC sampler are straightforward due to the aforementioned conditionally conjugate prior specifications on the mean and variance models. As a result of these priors, the full conditional for the mean is

$$\pi(\mu_{jk} \mid \cdot) \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n} \frac{r_i}{s^2(\boldsymbol{x}_i)}}{\frac{1}{\tau^2} + \sum_{i=1}^{n} \frac{1}{s^2(\boldsymbol{x}_i)}}, \frac{1}{\frac{1}{\tau^2} + \sum_{i=1}^{n} \frac{1}{s^2(\boldsymbol{x}_i)}}\right),$$

where $r_i = y_i - \sum_{1 \neq j} g(x_i; T_q M_q)$, and the full conditional for the variance is

$$s^2_{lk} \mid \cdot \sim \chi^{-2}\left(\nu' + n, \frac{\nu'\lambda'^2 + \sum_{i=1}^{n} \frac{e_i^2}{s^2_{-l}(\boldsymbol{x}_i)}}{\nu' + n}\right),$$

where $e_i^2 = \frac{(y_i - \sum_{j=1}^{m} g(\boldsymbol{x}_i; \boldsymbol{T}_j, \boldsymbol{M}_j)^2}{s^2_{-l}(\boldsymbol{x}_i)}$ and $s^2_{-l}(\boldsymbol{x}_i) = \prod_{q \neq l} h(\boldsymbol{x}_i; \boldsymbol{T}'_q, \boldsymbol{M}'_q)$. Combined with the likelihoods for the model parameters, there exists closed form formulations for the integrated likelihoods. For more details, please see Pratola et al. (2020).

2.4 HBART Probit for Classification

Parallel to the probit adaptation of Chipman et al. (2010), I combine the HBART model with a probit framework. Suppose that n independent binary random variables Y_i are observed, following a Bernoulli distribution with probability $p_i \equiv \mathbb{P}[Y_i = 1 \,|\, \boldsymbol{X}_i] = \Phi(\eta_i)$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq})$ is a collection of q covariates and η_i is a heteroscedastic nonlinear predictor for the probit model, for $i = 1, \ldots, n$. These probabilities can be modelled using HBART, such that the complete model is

$$\begin{aligned}
Y_i \,|\, \boldsymbol{X}_i &\sim \text{Bernoulli}(\Phi(\eta_i)), \\
\eta_i &= \frac{f(\boldsymbol{X}_i)}{s(\boldsymbol{X}_i)}, \\
f(\boldsymbol{X}_i) &= \sum_{j=1}^{m} g(\boldsymbol{X}_i; \boldsymbol{T}_j, \boldsymbol{M}_j), \\
s^2(\boldsymbol{X}_i) &= \prod_{l=1}^{m'} h(\boldsymbol{X}_i; \boldsymbol{T}'_l, \boldsymbol{M}'_l), \\
\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}' &\sim \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}').
\end{aligned} \tag{2}$$

Unfortunately, because of the nature of the probit model, there exists no conjugate prior $\pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}')$. This problem can be circumvented by using the latent variables as proposed by Albert and Chib (1993). That is, let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ be a vector of independent normal random variables such that $Z_i \,|\, \boldsymbol{X}_i \sim \mathcal{N}(f(\boldsymbol{X}_i), s^2(\boldsymbol{X}_i))$ and redefine $Y_i = \mathbb{1}[Z_i > 0]$, where $\mathbb{1}[\cdot]$ is the indicator function. Now, Y_i is deterministic conditional on Z_i. The equivalence of the

models follows from the fact that

$$\begin{aligned} p_i \equiv \mathbb{P}[Y_i = 1 \mid \boldsymbol{X}_i] &= \mathbb{P}[Z_i > 0 \mid \boldsymbol{X}_i] \\ &= \mathbb{P}\left[\varepsilon > -\frac{f(\boldsymbol{X}_i)}{s(\boldsymbol{X}_i)} \mid \boldsymbol{X}_i\right] \\ &= \Phi\left(\frac{f(\boldsymbol{X}_i)}{s(\boldsymbol{X}_i)}\right), \end{aligned}$$

where $\varepsilon \sim \mathcal{N}(0,1)$. This also explains the choice of modelling η_i as in (2). The resulting modified model is

$$\begin{aligned} Y_i &= \mathbb{1}[Z_i > 0], \\ Z_i &= f(\boldsymbol{X}_i) + s(\boldsymbol{X}_i)\varepsilon, \\ \varepsilon &\sim \mathcal{N}(0,1), \\ f(\boldsymbol{X}_i) &= \sum_{j=1}^{m} g(\boldsymbol{X}_i; \boldsymbol{T}_j, \boldsymbol{M}_j), \\ s^2(\boldsymbol{X}_i) &= \prod_{l=1}^{m'} h(\boldsymbol{X}_l; \boldsymbol{T}'_l, \boldsymbol{M}'_l), \\ \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}' &\sim \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}'). \end{aligned}$$

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Due to the slight modification of the model, the posterior distribution becomes

$$\begin{aligned} \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{X}) &\propto p(\boldsymbol{Y} \mid \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{Z}, \boldsymbol{X}) p(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{Z} \mid \boldsymbol{X}) \\ &= p(\boldsymbol{Y} \mid \boldsymbol{Z}) \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}') p(\boldsymbol{Z} \mid \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{X}) \qquad (3) \\ &= \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}') \prod_{i=1}^{n} \left[p(Y_i \mid Z_i) p(Z_i \mid \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{X}_i) \right]. \end{aligned}$$

The latter terms in (3) are then

$$\begin{aligned} p(Y_i \mid Z_i) &= \mathbb{1}[Y_i = 1]\mathbb{1}[Z_i > 0] + \mathbb{1}[Y_i = 0]\mathbb{1}[Z_i \leq 0] \\ p(Z_i \mid \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{X}_i) &= \phi(Z_i; f(\boldsymbol{X}_i), s^2(\boldsymbol{X}_i)). \end{aligned}$$

The function $\phi(\cdot; \mu, \sigma^2)$ here denotes the probability density of $\mathcal{N}(\mu, \sigma^2)$. Now all tools have been established to formulate the Gibbs sampling scheme. Using the fact that, given $\boldsymbol{X}$, $(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}')$ is conditionally independent of $\boldsymbol{Y}$, the full conditional posterior of the HBART

probit parameters is given by

$$\begin{aligned}\pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}' \,|\, \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{X}) &= p(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}' \,|\, \boldsymbol{Z}, \boldsymbol{X}) \\ &\propto \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}') p(\boldsymbol{Z} \,|\, \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{X}) \\ &= \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}') \prod_{i=1}^{n} p(Z_i \,|\, \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', \boldsymbol{X}_i) \\ &= \pi(\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}') \prod_{i=1}^{n} \phi(Z_i; f(\boldsymbol{X}_i), s^2(\boldsymbol{X}_i)).\end{aligned} \tag{4}$$

The former term in the final decomposition of (4) can be computed as in Pratola et al. (2020) and the latter by standard sampling methods. Moreover, the full conditional posterior of the latent variables $\boldsymbol{Z}$ is given by either half of the normal distribution, depending on the value of Y_i. That is,

$$Z_i \,|\, \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}', Y_i, \boldsymbol{X}_i \sim \begin{cases} \mathcal{N}(G(X_i), H(X_i)) \mathbb{1}[Z_i \leq 0], & \text{if } Y_i = 0 \\ \mathcal{N}(G(X_i), H(X_i)) \mathbb{1}[Z_i > 0], & \text{if } Y_i = 1 \end{cases}$$

A normal distribution of this sort is referred to as the truncated normal distribution. A possible solution to generating draws from this distribution is simply by rejection sampling, but there are several ways to do this faster. In my application, I employ the sampling scheme used in the code of the R package *bart*. In sum, to adapt HBART to a probit setting, the Gibbs sampler requires n subsequent draws from $Z_i \,|\, \cdot$ followed by m subsequent draws from $\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{T}', \boldsymbol{M}' \,|\, \cdot$ as in Section 2.3.

2.5 Heteroscedastic Survival BART

2.5.1 Data Reconstruction

The adaptation of probit models to survival analysis is one of (re)constructing the data, similarly to Sparapani, Logan, McCulloch and Laud (2016). Suppose n triplets $(\delta_i, t_i, \boldsymbol{X}_i)$ are observed, where δ_i indicates whether an event happened, $t_i \in \{0 = t_{(0)} < t_{(1)} < \cdots < t_{(k)}\}$ denotes the event time, and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq})$ is a vector of q individual specific covariates. Using these definitions, define

$$Y_{ij} = \delta_i \mathbb{1}[t_j = t_{(j)}], \text{ for } j = 1, \ldots, k_i, \tag{5}$$

where $k_i = \#\{j : t_{(j)} \leq t_i\}$. Let $\boldsymbol{Y}^*$ be the vectorization of the collection of observed instances of Y_{ij}. That is, the $\sum_{i=1}^{n} k_i \times 1$ vector

$$\boldsymbol{Y}^* = \begin{pmatrix} Y_{11} & \cdots & Y_{1k_1} & \cdots & Y_{n1} & \cdots & Y_{nk_n} \end{pmatrix}'.$$

To match this data structure, the covariates are repeated to match the first index and time stamps are added to the covariate matrix.

$$\boldsymbol{X}^* = \begin{pmatrix} t_{(1)} & \cdots & t_{(k_1)} & \cdots & t_{(1)} & \cdots & t_{(k_n)} \\ X_{11} & \cdots & X_{11} & \cdots & X_{n1} & \cdots & X_{n1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{1p} & \cdots & X_{1p} & \cdots & X_{np} & \cdots & X_{np} \end{pmatrix}'.$$

Furthermore, let $p_{ij} \equiv \mathbb{P}[Y_{ij} = 1 \,|\, Y_{i1} = 0, \ldots, Y_{i,j-1} = 0]$ and $\boldsymbol{p}$ the collection all these probabilities. The likelihood of the response variable given the aforementioned probabilities is

$$L(\boldsymbol{Y} \,|\, \boldsymbol{p}) = \prod_{i=1}^{n} \prod_{j=1}^{k_i} p_{ij}^{Y_{ij}} (1 - p_{ij})^{1 - Y_{ij}}$$

2.5.2 Induced Quantities

Since the aim of this paper is to analyze bankruptcy through survival analysis, the computation of the sheer probabilities p_{ij} will not provide a complete picture. Therefore, it is interesting to consider some other fundamental quantities in traditional survival analysis: the hazard (λ) and survival (S) function. These are

$$S(t_{(j)} \,|\, \boldsymbol{X}_i) \equiv \mathbb{P}[Y_{i1} = \ldots = Y_{ij} = 0 \,|\, \boldsymbol{X}_i] = \prod_{l=1}^{j} (1 - p_{ij})$$

$$\lambda(t_{(j)} \,|\, \boldsymbol{X}_i) \equiv \mathbb{P}[Y_{ij} = 1 \,|\, Y_{i1} = \ldots = Y_{i,j-1} = 0] = p_{ij}$$

The survival and hazard functions are closely related. The hazard rate can be interpreted as the instantaneous default probability. On the other hand, the survival probability is defined as the probability that the firm survives beyond time $t_(j)$. In other words, the firm does not default before $t_(j)$.

2.6 Quality Metrics

For the sake of model comparison, this section proposes several quality metrics and visual displays. I will adhere to the recommendations of Pratola et al. (2020) with regard to the quality metrics. These include predictive qq-plots, H-evidence plots, the root mean squared error (RMSE), and the e-statistic. For the extension, the performance of the SHBART model will be evaluated using time-dependent Brier score and a receiver operating characteristic (ROC) curve.

An H-evidence plot is an original graphical device by Pratola et al. (2020) that depicts the degree of heteroscedasticity detected by the HBART model. It displays the credible sets of different mean posterior estimates of the standard deviation. These credible sets are sorted on their corresponding posterior means. If heteroscedasticity exists, one could expect to see a steep slope for the mean of the posterior distributions of the standard deviations. To statistically substantiate this, the credible sets are plotted to rule out large overlaps. That is, if the credible

sets are small and the slope is steep, it is likely that there is heteroscedasticity present in the underlying data generating process.

Another graphical tool borrowed from Pratola et al. (2020) is the predictive qq-plot. In such a plot, the quantiles of the predictive posterior distribution are compared to those of a uniform distribution. When the model is correctly specified, this plot should depict a unit slope ray.

The final evaluation technique for the replication examples is the e-statistic (Székely, Rizzo et al., 2004). The statistic is defined as

$$e = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|U_i - V_j\| - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|U_i - U_j\| - \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \|V_i - V_j\| \right),$$

where $U_i \in \mathbb{R}^d$ are drawn according to distribution F_1 $\forall i = 1, \ldots, n_1$, $V_j \in \mathbb{R}^d$ are drawn according to distribution F_2 $\forall j = 1, \ldots, n_2$, and $\|\cdot\|$ denotes the Euclidian norm of its argument. This statistic offers to test the null hypothesis $H_0 : F_1 = F_2$. Although power comparison would be an informative decision, it is rather difficult to implement, and therefore beyond the scope of this paper. In this application simple comparison will suffice.

Brier scores (Brier et al., 1950) are score functions to assess the performance of probabilistic predictions. It is closely tied to the RMSE in the sense that the two are equivalent in a unidimensional setting. In the context of survival analysis, time-dependent Brier scores can be computed as

$$BS(j) = \frac{1}{N_j} \sum_{i=1}^{N_j} (p_{ij} - y_{ij})^2, \tag{6}$$

where N_j denotes the number of individuals at risk at time $t_{(j)}$. Brier scores range between 0 and 1, and lower Brier scores correspond to better probabilistic predictions.

ROC curves is a method of performance evaluation well-entrenched in survival analysis literature. To compute the curve, first the estimated event probabilities are sorted and discretized in 100 intervals corresponding to $x\%$ of the individuals at risk at some $t_{(j)}$. Then, for the $x\%$ highest risk individuals the fraction of individuals that experienced the event is calculated by dividing the number of high risk individuals by the total number of events over the whole sample period. Theses fractions are then aggregated over all sample periods and plotted against their corresponding fraction x.

3 Empirical Exercises

To evaluate the performance of the HBART method, this section explores one simulated- and three empirical exercises. Firstly, a univariate simulation is discussed. The reason for opting for a univariate simulation is that it is difficult to visualize the relation between the dependent and the independent variables through all interactions in the context of multivariate analysis. In a univariate setting, other traditional statistical methods may perform equally well, while retaining its interpretability. Therefore, two more empirical exercises from Pratola et al. (2020) are selected for evaluation. The second exercise treats HBART applied to alcohol consumption data, which is supplied with 35 potential covariates. As a third example, a dataset on fishing

catch production is used, which contains 25 potential covariates. Finally, the SHBART probit framework is employed to model firm bankruptcy.

3.1 Simulated Example

The simulated data in this example consists of 500 in-sample and 500 out-of-sample observations, all generated by $Y_i = 4X_i^2 + 0.2e^{2X_i}Z_i$, where $X_i \sim U(0,1)$ and $Z_i \sim \mathcal{N}(0,1)$, $\forall i$. This process clearly produces heteroscedastic data, which will be cardinal in showing the performance of HBART in contrast to BART. The data generating process (DGP) can also be written as

$$Y_i = f(X_i) + s(X_i)Z_i, \tag{7}$$

where $f(X_i) = 4X_i^2$ and $s(X_i) = 0.2e^{2X_i}$. The generated prediction data along with some estimated quantities can be seen in Figure 16 of the Appendix.

Figure 1 displays the quantiles of the estimation methods as compared to those of a uniform distribution. A perfect fit would result in a 45° degree line. It is evident from the left panel that BART is inferior to HBART in capturing the dynamics of the complete distribution of the simulated data. This result is supported by the value of the e-statistics for each method, which is 3.45 and 0.661 for BART and HBART, respectively. Interestingly enough, the root mean squared error (RMSE) of the two methods is similar for BART (0.764) and HBART (0.772). This demonstrates that the assumption of homoscedasticity can be relaxed without much loss of performance in point estimation, at least in this specific example.

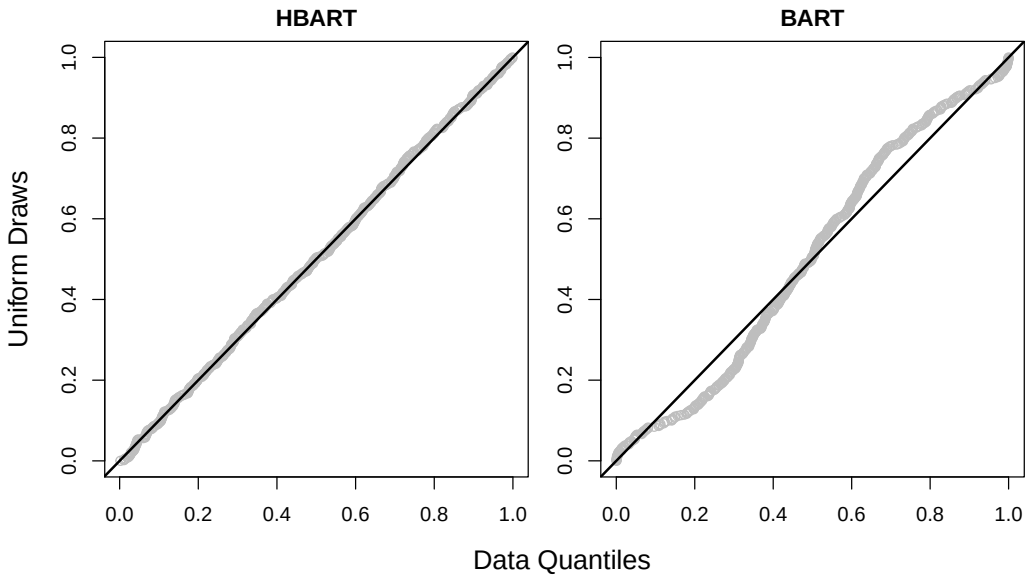

Figure 1: Simulated Example; Predictive QQ-Plots.

The H-evidence plot for the simulation can be found in Figure 2. By construction, this specific plot can easily be applied in higher-dimensional cases, because the estimate $\widehat{s}(x)$ incorporates the information for each covariate in one-dimensional space. That is, if the DGP would be homoscedastic, then the $\widehat{s}(x)$ are draws from identical distributions and the posterior intervals should align horizontally. In the simulated analysis, consistent with the model, the H-evidence

plot demonstrates that HBART detects and incorporates heteroscedasticity, as it clearly deviates from the constant standard deviation inferred by BART. For more details on this example, Figure 16 of the Appendix contains plots for the true and estimated data, and Figures 17 and 18 depict the estimated parameters against the true ones. Furthermore, the Appendix also contains Figure 19, which suggests the convergence of the MCMC draws by means of traceplots.

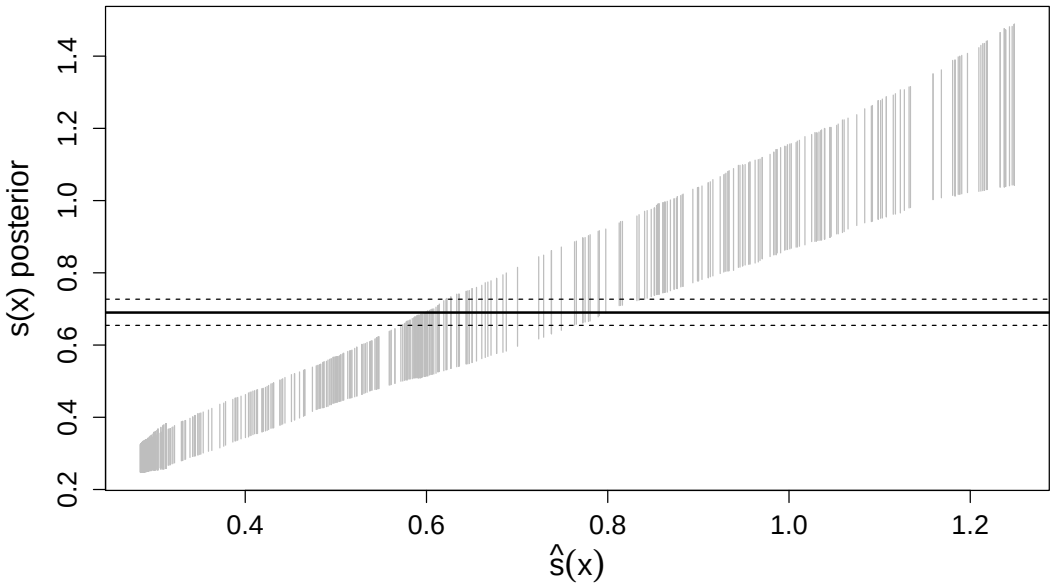

Figure 2: Simulated Example; H-Evidence Plot. The gray vertical lines represent 90% credible sets sorted by the point estimates of the standard deviation. The horizontal, solid line indicates the standard deviation as estimated by BART and the dashed lines showcase the associated 90 % credible set.

3.2 Alcohol Consumption

In the first of three empirical exercises, the performance of HBART is evaluated in the context of alcohol consumption. The data is originally used in (Kenkel & Terza, 2001) to study the effect of advice from a physician on individual alcohol consumption. In this case, the response variable is the (self-reported) alcohol consumption in the last two weeks. Moreover, the dataset consists of 35 predictors that portray the individual's demographics. In total 2,462 individuals are available for the analysis. Before analyzing the data using the HBART framework, some characteristics of the data will be explored in order to detect early signs of heteroscedasticity or other statistical properties.

Figure 3 displays the boxplots for some categorical covariates. The cardinal covariate in predicting alcohol consumption in this specific context is the advice indicator. This variable is determined by the respondent's answer to the question 'Have you ever been told by a physician to drink less?'. The alcohol consumption distributions, as shown by the box plots, for different answers to this question differ. Alcohol consumption for the respondents that reported not having ever received such advice is more concentrated at the lower end of the distribution. The 25-50% quantile range of the respondents that got no advice is more densely concentrated than those that did; the bottom 25% of respondents, in either case, seems to be roughly equivalent.

Another interesting characteristic is the age of the respondent at the time of measurement. As age increases, average alcohol consumption declines. Over 25% of the respondents over the age of 70 even reported not consuming alcohol in the past two weeks at all. This is in contrast to other age groups, where alcohol consumption is more common. Another interesting observation, which applied to the advice covariate as well, is the consistency of the location of the bottom 25% quantile in each subgroup. This indicates that those respondents that have consumed little alcohol in the past two weeks are equally distributed among the subsamples, with the exception of the oldest age group. In particular, about 22% of the respondents reported not having consumed alcohol in the past two weeks. It may be reassuring to see that the subsamples are at least homogeneous in this respect.

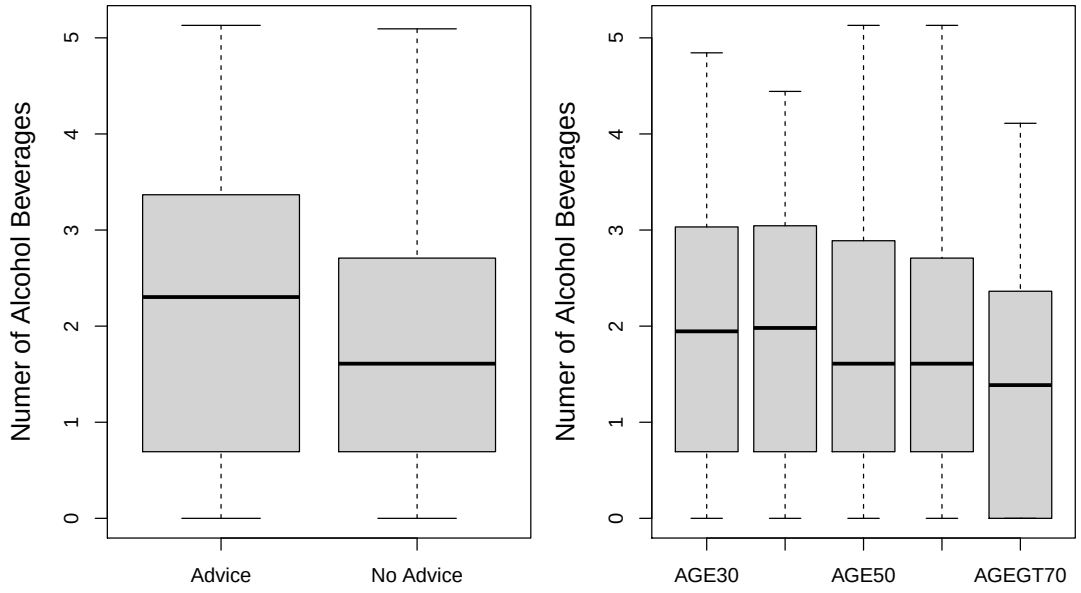

Figure 3: Alcohol Example; Box Plots for Response and Age.

Finally, an important predictor may be the years of education of the respondents. Figure 4 plots alcohol consumption for the years of education and distinguishes respondents that have received a physician's advice. There seem to be no clear discrepancies in alcohol consumption when comparing education years. There is, of course, the possibility that this is due to a lack of observations at the lower end of the years of education. Moreover, Figure 4 also does not unveil an explanation for the heterogeneity among the respondents that reported having received advice from a physician and those that did not. All in all, the statistical properties of alcohol consumption seem to not be completely heterogeneous in some dimensions, but there is no clear conclusion to be drawn from these results.

Hereafter, 2,000 draws are used for the convergence of the Gibbs sampler. Then, 1,000 initial draws for the MCMC are discarded and the following 2,000 are kept. Moreover, the mean model uses 200 regression trees and the model for the standard deviation uses 40; each predictor is allowed 1,000 cutpoints. The prior hyperparameter κ is set to 5 for the heteroscedastic model and 2 for BART. For the variance model ν is set to 3 and λ is set according to the approach of Chipman et al. (2010). Finally, the estimation set consists of 1,477, or 60%, randomly sampled rows from the entire dataset.

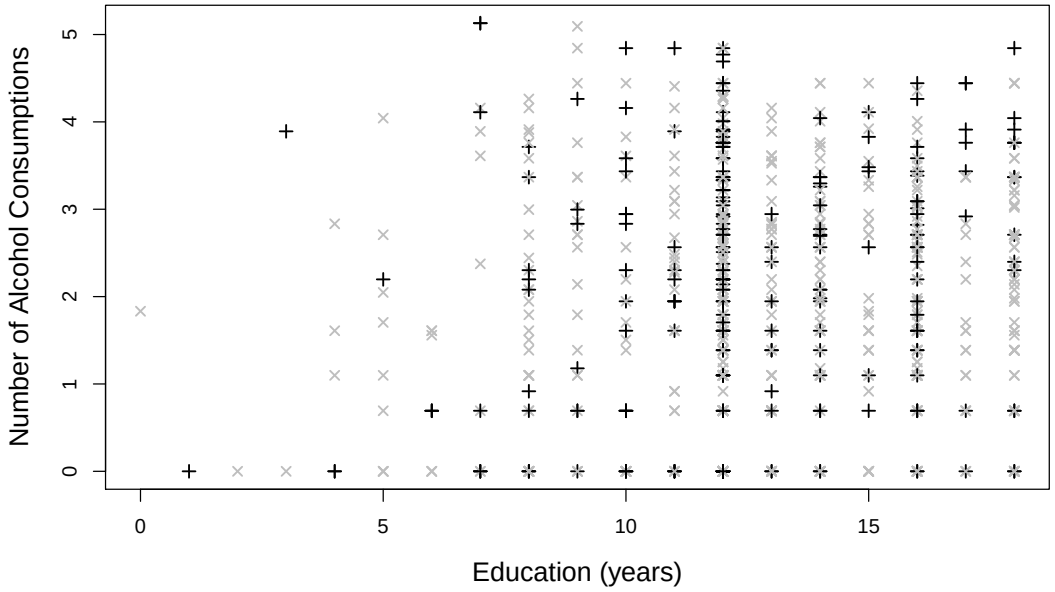

Figure 4: Alcohol Consumption Based on Years of Education. The grey crosses correspond to respondents that received no advice from a physician.

In contrast to the simulation example, when predicting alcohol consumption in this context, HBART does not drastically improve the estimation of the sample distribution as measured by the e-statistic. Both methods obtain a similar RMSE of 1.34, reinforcing the supposition in the simulated example that relaxing the assumption of homoscedasticity does not largely affect the performance in point prediction. On the other hand, the value of the e-statistic in this example is 3.00 for BART and 2.77 for HBART. This result suggests that quantiles obtained from both BART and HBART differ from their theoretical values. This is not to say the difference is statistically significant. Power comparison would be useful in this scenario, but is a complicated procedure, as can be seen in Székely et al. (2004). Figure 20 of the Appendix confirms that the quantiles of BART and HBART compare to their expected values in a similar way. The lack of relative improvement to homoscedastic BART is an early suggestion that heteroscedasticity may not play a major role in this empirical exercise.

The H-evidence plot in Figure 5 corroborates the aforementioned lack of improvement by HBART. That is, the slope for the credible sets for the estimates of the standard deviation by HBART is not as steep as one would expect if heteroscedasticity was a major problem for the data. Almost all credible sets intersect that of the estimated variance in the homoscedastic model. All in all, it seems to be the case that adapting BART to incorporate heteroscedasticity in this example is not necessary, but at the same time, it does no harm to its predictive performance.

3.3 Fishing Catch Production

A second empirical exercise uses data from fishing catch production (Fernández, Ley & Steel, 2002). The data contains 6,806 observations for 56 fisher boats over 1993 and the first half of 1994. The response variable in this example is the daily catch of fishing boats. Moreover, after converting some categorical variables to dummies, 25 covariates become available. These covariates contain information about the ship, the time of the year, and the location.

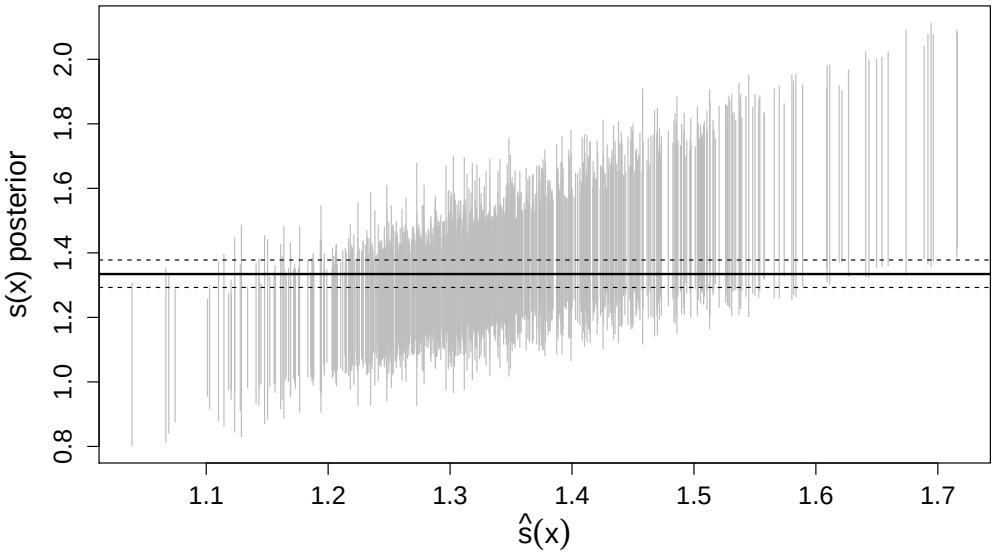

Figure 5: Alcohol Example; H-Evidence Plot. The gray vertical lines represent 90% credible sets sorted by the point estimates of the standard deviation. The horizontal, solid line indicates the standard deviation as estimated by BART and the dashed lines showcase the associated 90 % credible set.

Figure 6 depicts the box plots corresponding to the zone that the observed fisher boat was active on that day. The complete span of the distribution of observations in zone 3O is much larger. More specifically, zone 3O seems to have more spread in observations that exceed the median[1]. Nevertheless, the bottom 50% data points seem to be reasonably homogeneous among the four categories. In sum, the location of fishing is a source of heterogeneity which may contribute to the fact that heteroscedasticity is present.

Another source of possible heterogeneity in the covariates is the mesh size of the nets used in fishing. The box plots corresponding to the different mesh sizes are displayed in Figure 7. Again, ignoring the outliers, the response values seem to differ substantially when plotted on different mesh sizes. In particular, the distribution of observations that used nets with a mesh size of 129mm stands out. However, this category contains only 16 observations, too few to draw any sound conclusion. On the other hand, large mesh sizes seem to be concentrated at lower values. This may suggest that the mesh size is too large for catching lots of fish, mesh sizes that are larger are intentionally chosen (for example, some ships may aim for a specific species of fish and do not want to bother with smaller fish), or the yield of fish as a consequence of mesh size could be confounded by other covariates (maybe larger mesh sizes are chosen in zones or periods of less fish). A brief attempt to explore the third possibility for different yields of fish as a result of the mesh size at different locations is provided in the Appendix in Figure 22. Looking globally at the differences in response values for these categories, however, it may be wise to equip the prediction model with the ability to handle heteroscedastic data.

[1]Note that these boxplots do not include outliers for the sake of clear visualization. Observations that are not contained in the interval between the first and third quartile expanded with 1.5 times the interquartile range between the first and third quartile in each direction are considered outliers. The other zones contained many outliers, which implies that they have a higher concentration in a smaller region of response values. This was confirmed by reviewing the histograms.

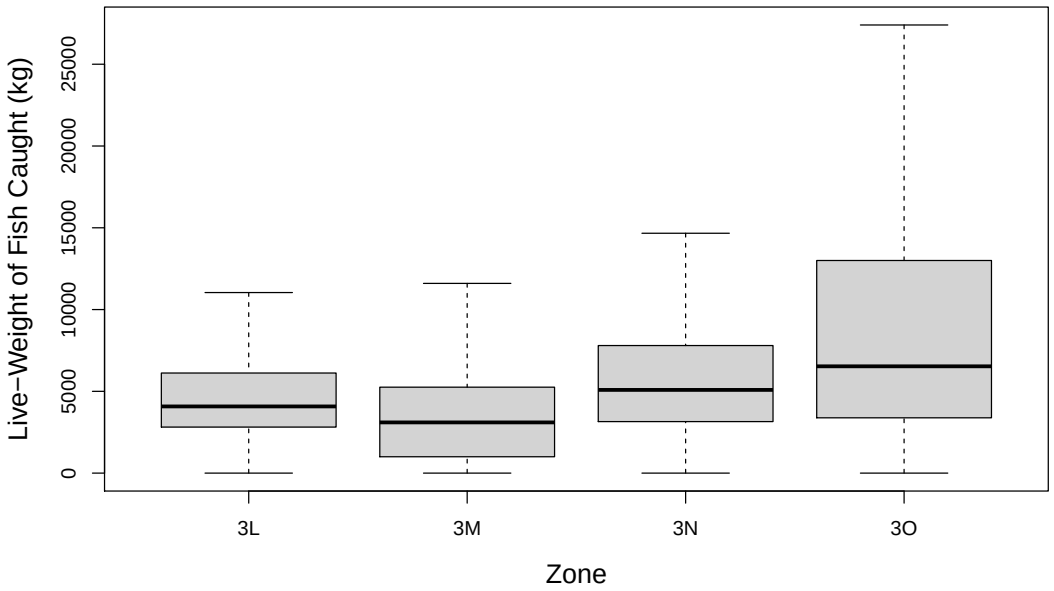

Figure 6: Fishery Example; Box Plots Zone.

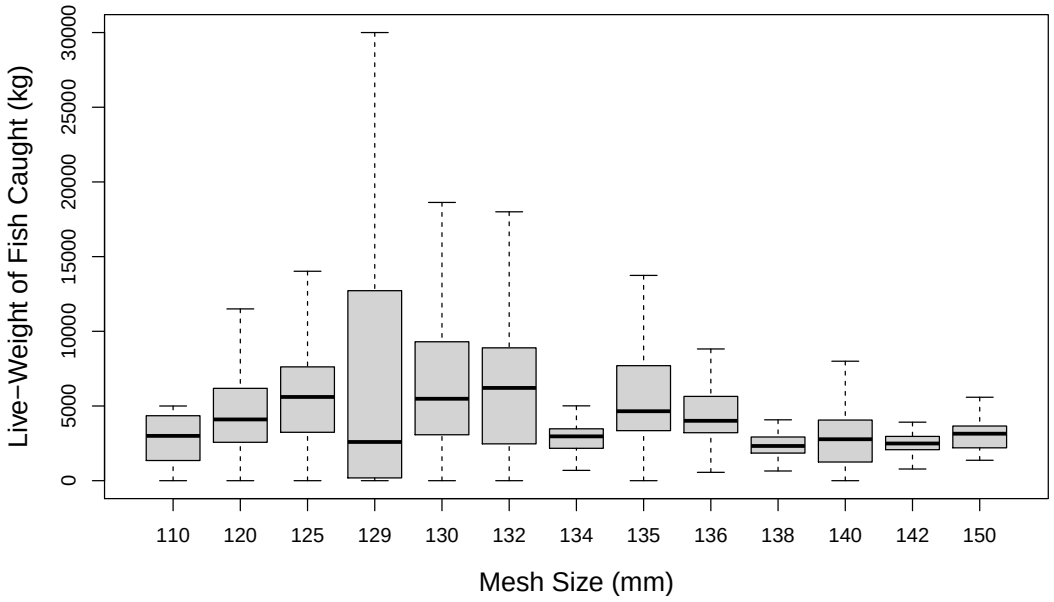

Figure 7: Fishery Example; Box Plots Mesh.

The hyperparameters for the analysis are identical to the previous example. In this case, 60% of the data corresponds to 4,084 rows that are randomly sampled and designated to be rows for the training data. As in each of the examples so far, the RMSE of the point prediction for BART and HBART are close again; BART has an RMSE of 3,812 and HBART 4,140. Moreover, the quantiles predicted by BART seem to be much different from uniform draws as the e-statistic is 13.4. HBART produces a value of 2.35, which is not too large compared to the other examples. This result is also evinced in the predictive qq-plot in Figure 21 of the Appendix.

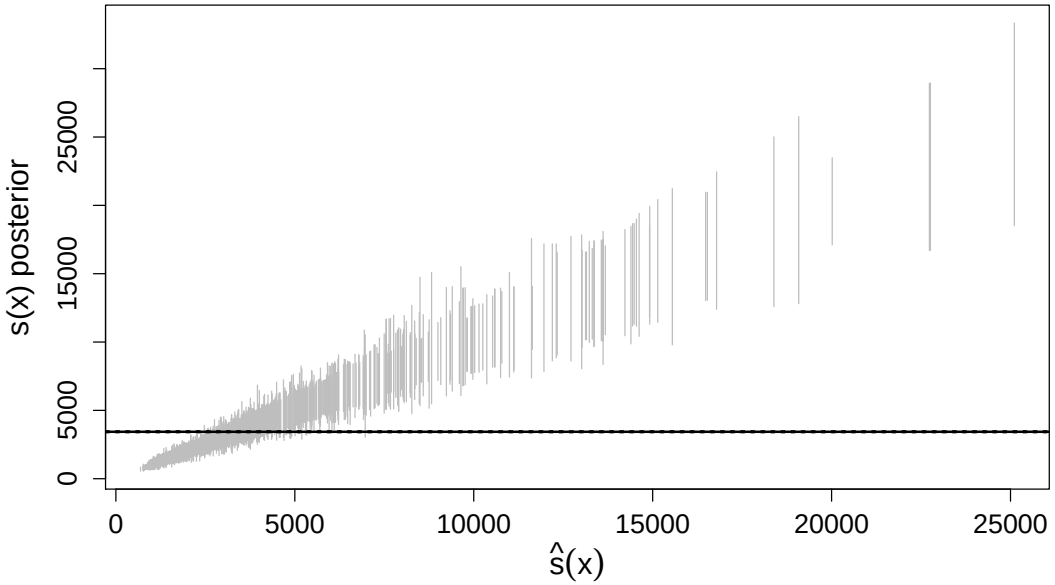

Figure 8: Fishery Example; H-Evidence Plot. The grey vertical lines represent 90% credible sets sorted by the point estimates of the standard deviation. The horizontal, solid line indicates the standard deviation as estimated by BART and the dashed lines showcase the associated 90 % credible set.

The H-evidence plot in Figure 8 corroborates the earlier discussion about heterogeneous data. The homoscedastic BART model yields a relatively small standard deviation compared to the values obtained by HBART. Also, the standard deviation appears to be widely distributed, ranging from values between 687 and 25,107. It is clear that, in this application, ignoring the heteroscedasticity present in the data by applying the homoscedastic BART method, can lead to grave errors in statistical inference and economic conclusions.

3.4 Bankruptcy Prediction

Finally, the SHBART framework is exploited for bankruptcy prediction for U.S. firms after 2000. The dependent variable in this exercise is the event indicator as in (5), derived from the UCLA-LoPucki Bankruptcy Research Database. The dates are coarsened to years for computative simplicity.

The covariates are obtained from the Financial Ratios Suite created by the Wharton Research Data Center. As a starting point – that is, after deleting some identifying variables – the processed dataset contains monthly data for 68 covariates, resulting in over a million observations. To reduce computation time in the analysis, only 1,000 viable firms are selected,

along with the 434 bankrupt firms that were still operating after 2000 and were listed in the financial ratios dataset. Moreover, the final two years of the sample get deleted due to excessive missing data and a dispropotional abundance of bankrupt firms compared to viable firms. The resulting sample consists of exactly two decades of data. Finally, missing data is handled with by deleting all covariates with more than 4% missing data, deleting the bankrupt firms that had missing data for the year they went bankrupt, and dropping the variables research and development/sales, advertising expenses/sales, and labor expenses/sales due to the fact that a lot of observations had a value of 0 for these covariates, whereas this seems improbable. Remaining missing rows are simply deleted.

The boxplots associated with some of the covariates used in the final analysis are displayed in Figure 9. Heterogeneity among viable and bankrupt firms is especially prominent in some of these financial ratios. For example, the distribution of price/earnings ratios for viable firms seems to be shifted upward respective to bankrupt firms. This means that, on average, over the whole observed period, the price over earnings of the stock of a firm that would ultimately go bankrupt was smaller for bankrupt firms than for viable firms. Therefore, the price/earnings ratio may be an intimation for financial distress. However, compared to other ratios in Figure 9, price/earnings seems to represent only a subtle difference.

The other equity related ratios paint a more distinctive picture of the differences between viable and bankrupt firms. The average book/market ratio for a viable firm is more concentrated at lower values, whereas the opposite is true for the price/sales ratio. These observations taken together may indicate that investors are not appealed towards stocks that ultimately go bankrupt. This follows from the fact that, given the calculation of the two ratios, the book value to revenue seems to be failry homogeneous among viable and bankrupt firms, as a low price/sales ratio offsets a large book/market ratio. Another distinctive feature of bankrupt firms is high variability, especially at the lower end, of the return on equity. Bankrupt firms seem to be, on average, less profitable over the entirely operating period and are so at inconsistent rates.

Another premonition for the bankruptcy of a firm is its capital structure. The most salient of differences in financial ratios is the cash balance/total liabilities ratios. This observation conforms with findings in other literature in bankruptcy predictions and substantiates the fact that solvency is a primary category to differentiate viable from non-viable firms (Liang, Lu, Tsai & Shih, 2016). Moreover, on average, firms that go bankrupt have approximately 20 pp. more long-term debt proportional to their total liabilities. These aforementioned differences in the capital structure may be the cardinal reasons why firms become insolvent.

To justify using the mean of the financial ratios for the survival analysis, the mean time series of the 8 predictors are depicted in Figure 10. Most predictors are generally concurrent for both types of firms, substantiating the use of the average. Only the price/sales and book/market ratios seem to deviate from a concurrent pattern over several years. Overall it seems reasonable to assume the validity of averaging the yearly ratios.

To test the SHBART methodology on this specific example, half of the data is used as estimation sample and the other half as a prediction sample. Like in the previous empirical exercises, 2,000 draws are used to ensure the convergence of the Gibbs sampler. To accommodate the long computation time of the method, only 500 draws subsequent draws will be skipped. Finally,

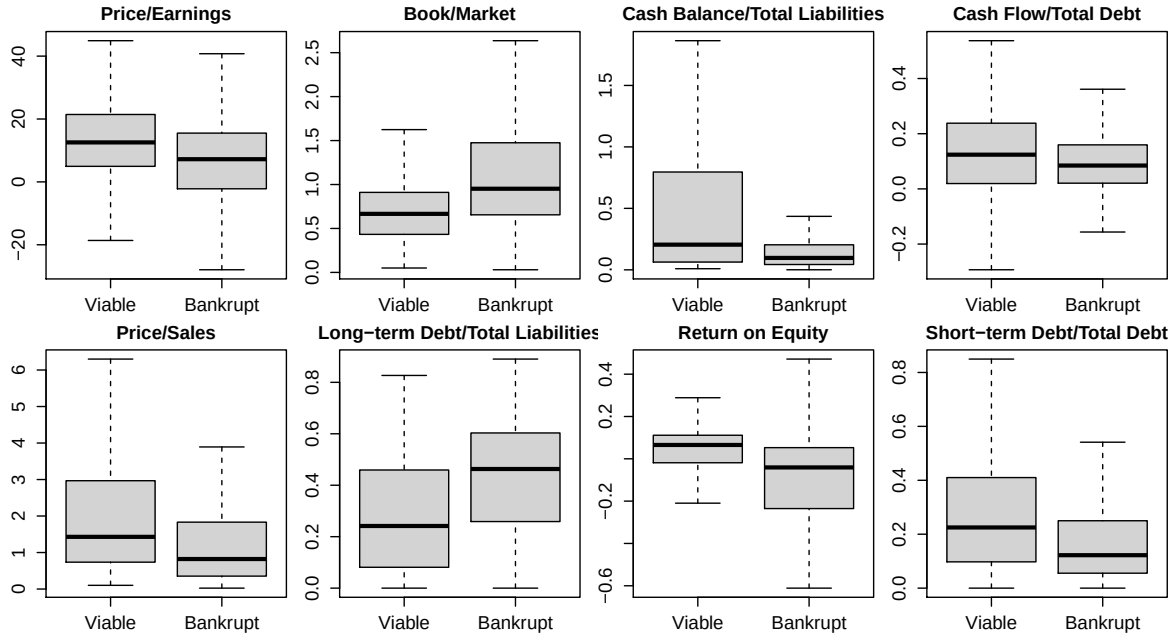

Figure 9: Box Plots for Selected Covariates.

2,000 draws are kept for posterior inference. As in the previous exercises, the number of trees used to model the mean and variance are 200 and 40, respectively. The hyperparameter τ for the parameters in the leaf nodes for the mean model is set to $2.25/\sqrt{m} = 0.159$, where m denotes the number of trees used to model the mean function (Chipman et al., 2010; Sparapani et al., 2016). Finally, the variance prior uses the hyperparameters $\nu = 3$ and $\lambda = \sqrt{Q_{0.1}(f_{\chi^2(\nu)})/\nu} \approx 0.44$, where $Q_{0.1}(\cdot)$ denotes the 10% quantile of its argument and $f_{\chi^2(\nu)}$ denotes the pdf of the χ^2 distribution with ν degrees of freedom. These hyperparamaters define a conservative prior with high probability for the default probit variance of 1. Note that these hyperparameters are transformed as in Section 3.4 in Pratola et al. (2020).

Figure 11 depicts the H-evidence plot for the bankruptcy prediction model. Although there is an evident upwards sloping trend in the posterior intervals, heteroscedasticity appears to be unsubstantiated. The homoscedastic model estimates a relatively low posterior variance compared to that of the heteroscedastic model. The same applies to the credible sets, which are much wider for the heteroscedastic model than the homoscedastic model.

The median survival probabilities corresponding separately to bankrupt and viable firms are depicted in Figure 12. The decreasing slope for bankrupt firms is in stark contrast to the slope of the survival function of the median viable firm. That is, the model seems to appropriately assign lower survival probabilities to firms that did go bankrupt in the prediction sample. Nevertheless, there is still a lot of overlap between the survival probabilities for subsets of bankrupt and viable firms, reveiling major prediction errors in recognizing financial instability.

A more comprehensive assessment of the influence of individual financial ratios can be made from Figure 13. In this figure the mean survival probability of firms that fall within a certain quantile range is plotted for each year. Darker lines are associated with lower values for the financial ratio. None of the ratios seem to exhibit a distinctive pattern for the corresponding

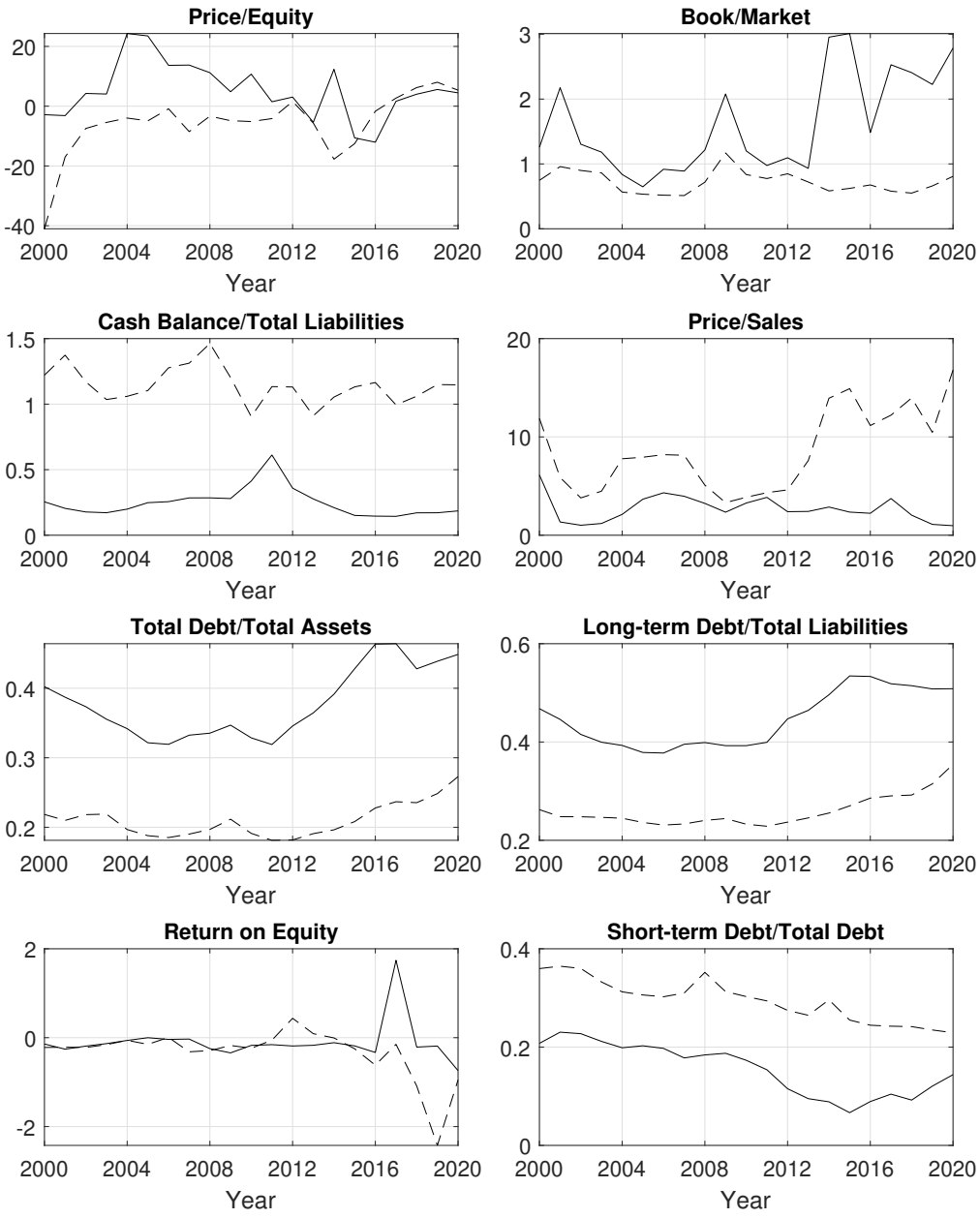

Figure 10: Time Series Plots for Selected Covariates. The solid line represents the yearly time series for bankrupt firms and the dashed line represents the time series for viable firms.

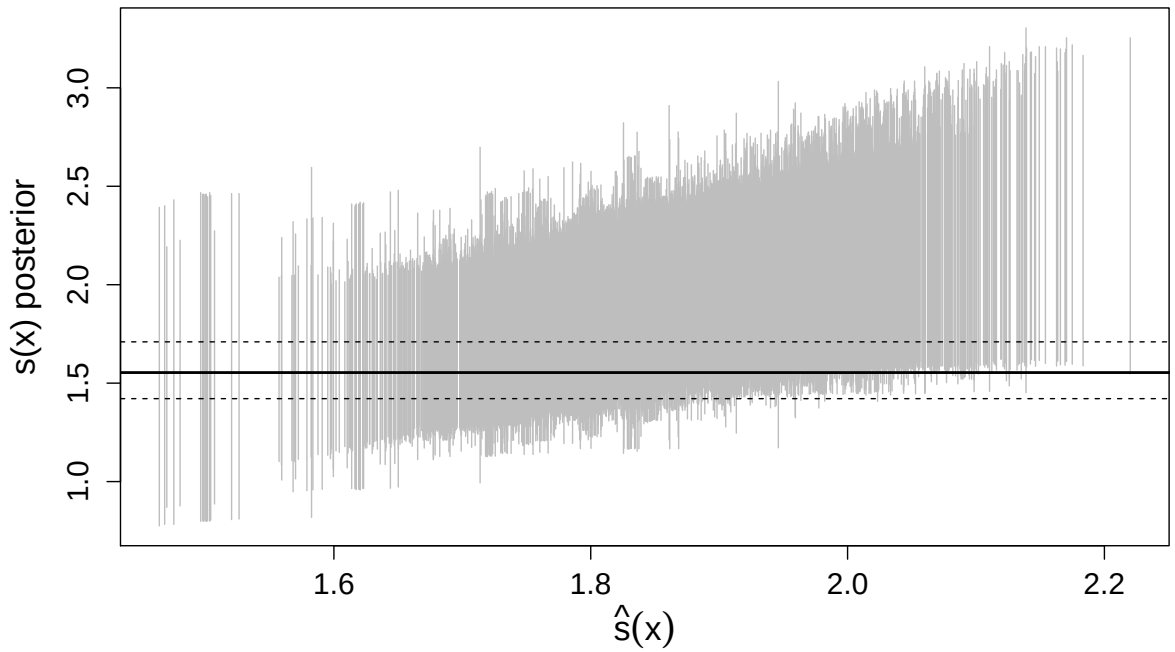

Figure 11: Bankruptcy Example; H-Evidence Plot. The gray vertical lines represent 90% credible sets sorted by the point estimates of the standard deviation. The horizontal, solid line indicates the standard deviation as estimated by BART and the dashed lines showcase the associated 90 % credible set.

survival probabilities. Only a vague arrangement can be discerned for the book/market and price/sales ratio. On average it seems to be the case that a lower book/market ratio is associated with a higher survival probability. On the other hand, firms with high price/sales ratios seem to have a higher chance of surviving at most points in time. Note that these results align with the exploration in Figure 9. The lack of a distinct arrangement for these variables does not imply that the model is inadequate at incorporating the relationships of the ratios to bankruptcy. It does, however, suggest that a uni-dimensional take on the contribution of financial ratios on the viability of a firm is inadequate, as the model outperforms random predictions.

The Brier scores corresponding to each year are displayed in Figure 14. In accordance with the preceding discussion about heteroscedasticity, the performance for the homoscedastic and heteroscedastic models is almost indistinguishable in the bar plot. Overall, the Brier scores seem to fluctuate with the fraction of at-risk firms that go bankrupt. This concurrent movement suggests that the high Brier scores are mainly driven by bankruptcy observations. A possible explanation for this phenomenon is that the model underestimates the default probability, such that it is too conservative. This idea is reinforced by the hazard function in Figure 23 in the Appendix. The fraction of bankrupt firms often exceeds the median default probability. Some remarkable years are those of 2009-2010 and 2020. These years can be associated with periods of financially instable economic environments, or at least their aftermath. First, the financial crisis of 2008 is a logical explanation for the spike in the fraction of bankrupt firms and, consequently, the error in bankruptcy prediction. Secondly, 2020 corresponds with the first whole year during the Covid-19 crisis.

To compare the performance of the SHBART framework with random predictions, Figure

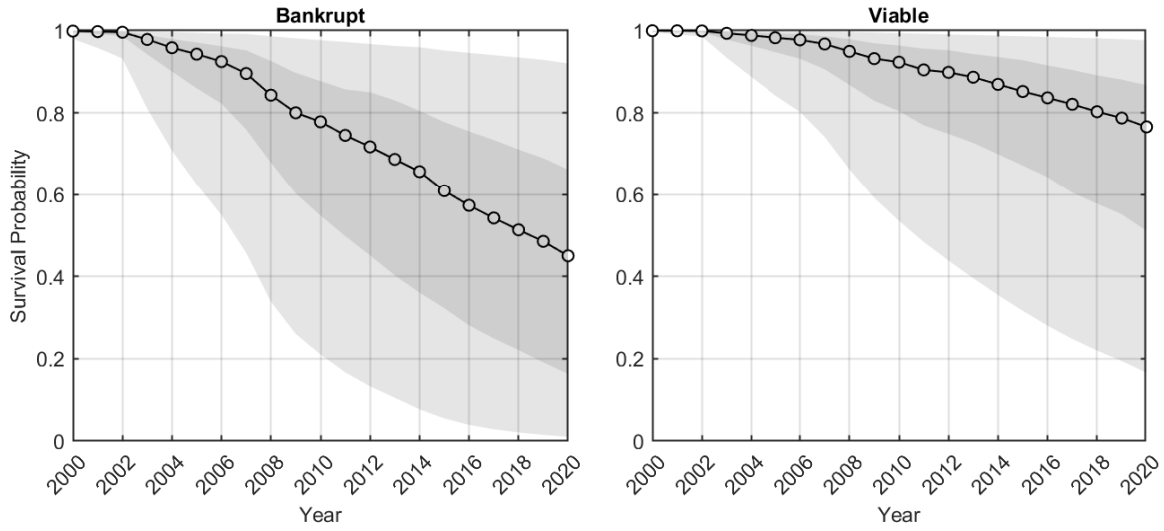

Figure 12: Survival Function for Bankrupt and Viable Firms. The line represents the median survival probability at each year, the dark shaded region depicts survival probabilities between the survival probabilities between the 25% and 75% quantiles, and the light shaded the survival probabilities between the 5% and 95% quantiles.

15 contains ROC-curves for both the HBART and BART model. The concavity of the curves indicates the diagnostic performance of the method. For example, if 20% of the 20% highest default risk firms can actually be classified as bankrupt in some period, then the model performs just as well as a random model. If, on the other hand, 80% of the highest 40% default risk firms go bankrupt somewhere in the sample period, then the model at hand produces an informative prediction. Moreover, the homoscedastic model seems to perform marginally better in terms of predictive performance.

4 Conclusion

This paper has explored four applications of the HBART method proposed by Pratola et al. (2020) and extended it by combining it with a probit framework applied to survival analysis. The results show that relaxing the assumption of homoscedasticity for BART can, in any case, be done without hefty loss of predictive performance, both in its basic form as in the extended survival analysis methodology. The simulated example demonstrated that HBART can flexibly model a nonlinear DGP, whereas BART fails to make accurate predictions. The fishing catch example corroborates these findings in an empirical setting. On the other hand, the alcohol consumption example supports the adaptability of HBART, by finding little to no difference in the predictive performance compared to BART, despite the lack of convincing evidence suggesting heteroscedasticity.

The methodological extension to SHBART can be used to successfully predict bankruptcy using financial ratios. On average, the model is able to distinguishes between viable and bankrupt firms. Because of its flexibility and lack of a priori assumptions, SHBART can be a promising method to model survival data. However, the bankruptcy example showcased no distinct signs of heteroscedasticity, leaving the discussion open as to the relative predictive performance for

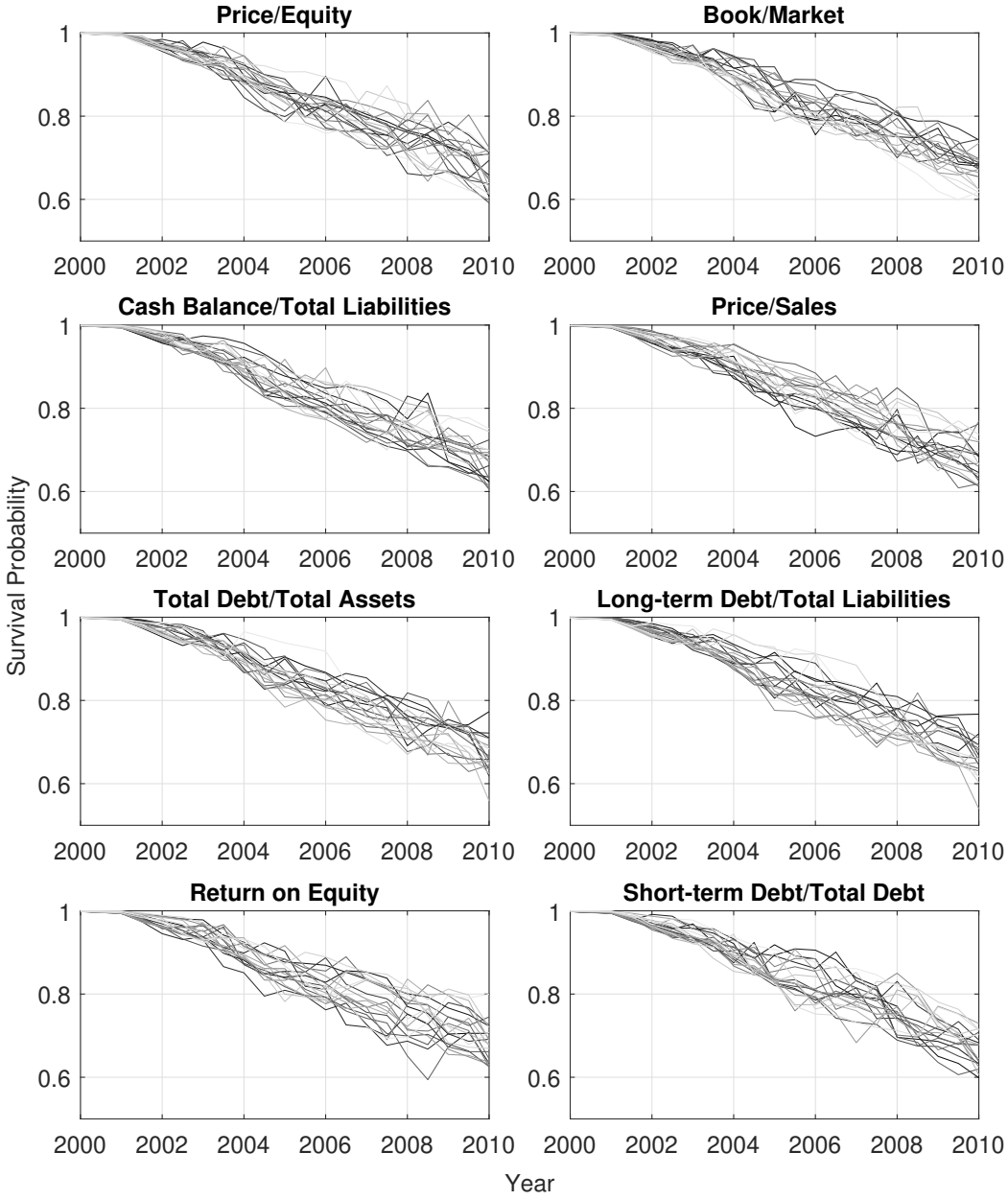

Figure 13: Survival Functions for Firms in Different Bins for the Financial Ratios. Each line represents the mean financial ratio within an interval of 5% of the data, sorted on the corresponding ratio. Lower values for the financial ratio are associated with a darker line.

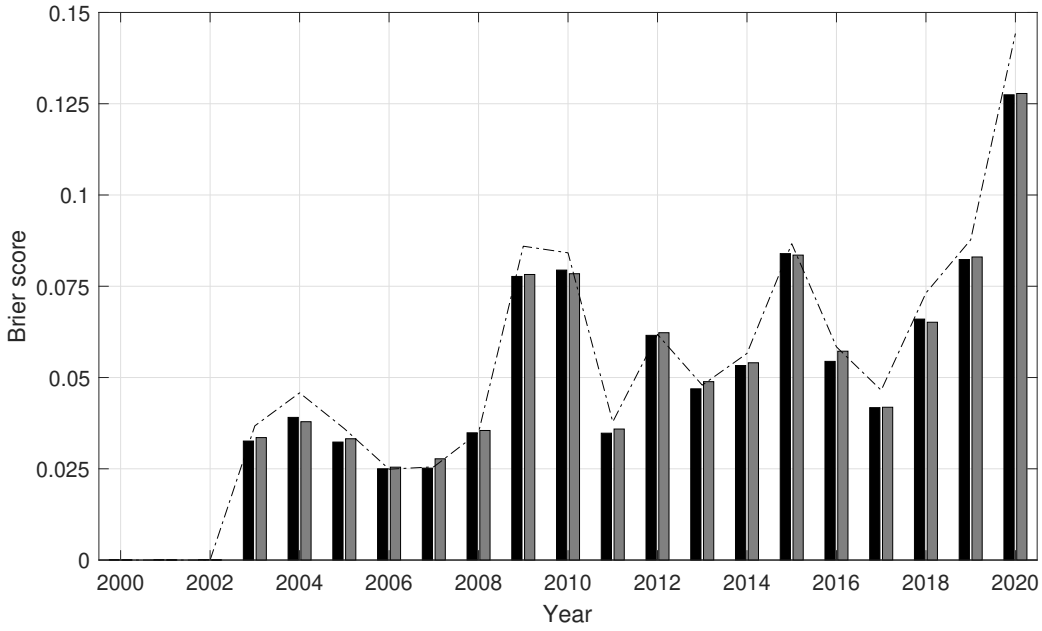

Figure 14: Brier Score in Each Year in the Prediction Sample. The black bars correspond to predictions made by the SHBART model and the gray bars to the predictions produced by the homoscedastic variant. A dashed-dotted line represents the fraction of at-risk firms that went bankrupt in the corresponding year.

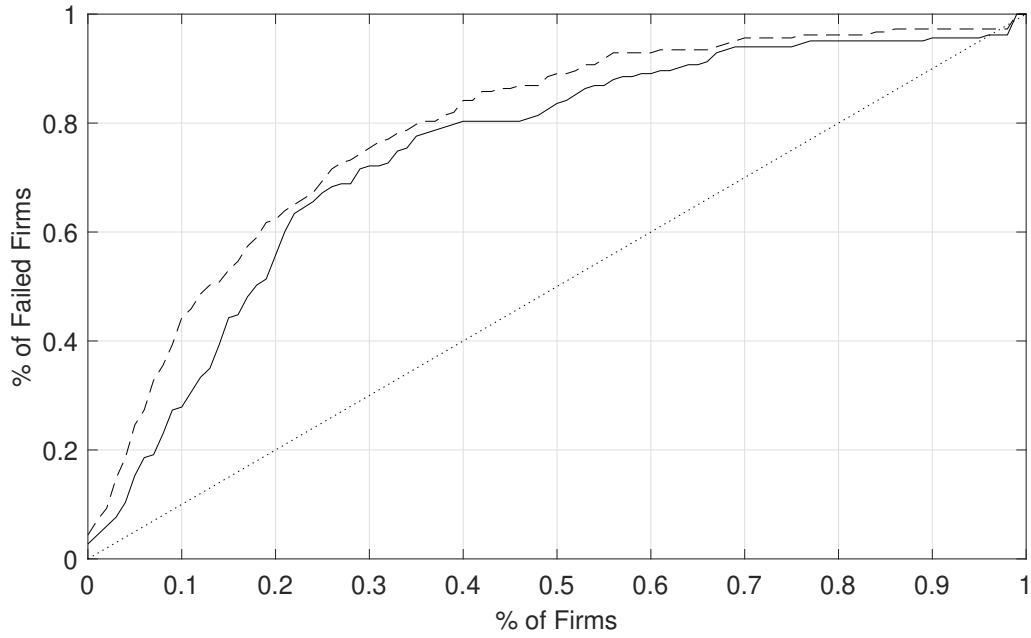

Figure 15: ROC Curves for Bankruptcy Prediction Model. The dashed line corresponds to the homoscedastic variant, the solid line the SHBART method, and the dotted line to predictions according to a random model.

SHBART compared to survival analysis using BART. To further evaluate the SHBART methodology, a comparative study using a range of traditional survival analysis methods could be explored. Ideally, such a study would concern a heteroscedastic outcome, which may not be the case in general bankruptcy prediction.

This paper focused on bankruptcy for U.S. publicly listed firms, thus there are a couple of dimensions that could hinder the generalizability of the performance assessment. First, the data set includes only firms in the U.S. Although the U.S. has a similar economic environment to a lot of western countries, the results of this paper may be unreliable for developing countries or areas with an entire different macroeconomic milieu. Another concern for the generalizability of the results to different economies or countries is that of different regulations regarding government bailouts and economic policy to aid firms in financial distress. Second, the fact that the firms in the dataset are publicly listed may be a hindrance to the interpretation of the results in this paper for firms that are not. Moreover, many financial ratios used in estimating the model are exclusive to publicly listed firms and could not be used in the estimation of firms that are not publicly listed. Third, the BRD contains only data for firms with assets valued more than $100 million in 1980 USD. The results of the performance evaluation of SHBART for smaller firms thus remains inconclusive. In sum, there are at least three dimensions that further empirical research could focus on to generalize the understanding of the performance of SHBART.

There are several more aspects for improvement. For example, to ensure proportional representation of bankrupt firms, selection bias is inevitable. Moreover, a choice-based sample bias was introduced in restricting the data to complete cases. The model developed in this paper could be enhanced with techniques accounting for these biases such as those proposed by Zmijewski (1984). Furthermore, because of constraints on the computation time of the SHBART approach, yearly data was employed, while monthly data was available. Besides being more pragmatic, the predictive performance could be improved upon by using monthly data Chava and Jarrow (2004).

Two final points of discussion are the choice of the hyperparamaeters and the use of time-varying covariates. This paper did not explore any cross validation method for the hyperparameters, but opted for the settings suggested by the literature (Chipman et al., 2010; Sparapani et al., 2016; Pratola et al., 2020). Further research could focus on cross-validation using the optimization of, for example, the "area under curve" (AUC) for the ROC curve or Brier scores. Moreover, using time-varying covariates could be a precious addition. The methodology of this paper could be extended to a more dynamic setting by allowing for changing financial ratios. Although there seemed to be no sign of a contemporaneous difference for the financial ratios for bankrupt and viable firms at a yearly frequency, this may not be the case when using monthly data. When using monthly data, the financial ratios for bankrupt firms might experience deviations from the regular pattern close to the bankruptcy date. Therefore, the model can be improved in two ways: one is the use of monthly data, which is said to increase the relative superiority of hazard approaches over traditional methods (Chava & Jarrow, 2004), and another is the inclusion of "anticipation" effects of bankruptcy for the financial ratios.

References

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Bauer, J. & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? a comprehensive test. *Journal of Banking & Finance*, *40*, 432–442.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71–111.

Begley, J., Ming, J. & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of altman's and ohlson's models. *Review of accounting Studies*, *1*, 267–284.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Chava, S. & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of finance*, *8*(4), 537–569.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *Ann. Appl. Stat.*, *4*(1), 266–298.

Fernández, C., Ley, E. & Steel, M. F. (2002). Bayesian modelling of catch in a north-west atlantic fishery. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *51*(3), 257–280.

Kenkel, D. S. & Terza, J. V. (2001). The effect of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *Journal of Applied Econometrics*, *16*(2), 165–184.

Liang, D., Lu, C.-C., Tsai, C.-F. & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European journal of operational research*, *252*(2), 561–572.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Pratola, M. T., Chipman, H. A., George, E. I. & McCulloch, R. E. (2020). Heteroscedastic bart via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, *29*(2), 405–417.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, *74*(1), 101–124.

Sparapani, R. A., Logan, B. R., McCulloch, R. E. & Laud, P. W. (2016). Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in Medicine*, *35*(16), 2741–2753.

Székely, G. J., Rizzo, M. L. et al. (2004). Testing for equal distributions in high dimension. *InterStat*, *5*(16.10), 1249–1272.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 59–82.

A Code

In the attached compressed file or on *https://github.com/nielsvandenheuvel/bsc_thesis*, the code used for generating the results in this paper can be found. Pulling the directory from Github is probably recommended, since here the complete raw datasets are contained for the preprocessing procedure. Alternatively, a small sample of the raw data is attached in the compressed code directory.

Each subdirectory corresponds to a replication or extension part in the paper. Moreover, the surv_rbart directory contains the files for the package used in the extension. This package needs to be build before using. An elaborative guide is provided as a README.md file in the code folder.

The *simulation/simulated_example.R* contains the code for the simulation exercise. It first generates the data used in this example and, subsequently, runs the *HBART* and *BART* analyses with the set parameters. Then, the evaluation metrics as explained in the methodology are computed. Finally, the code produces the H-evidence plot, qq-plots, data plots with real and estimated results, and a traceplot. The corresponding figures in the paper are Figure 1, 2, and 16 - 19.

The *alcohol/alcohol.R* and *fishery/fishery.R* files contain the code used to generate the results in this paper for the alcohol consumption and fish catching examples. They are structured similarly to the *simulation/simulated_example.R* file with regards to estimation. In this case, the code also generates plots for the exploratory analysis like the boxplots in each example. The corresponding figures to alcohol consumption and fishery in the paper are Figure 3 - 5 and 20, and 6 - 8, 21, and 22, respectively.

The *bankruptcy/bankruptcy.R* contains the code for the extension. The file contains some exploratory analysis on top of the traditional code structure from the other empirical examples. Moreover, the *complete.cases* function in R is employed in order to avoid empty cuts. The main results for this paper are generated in *Postprocessing.m*, while the raw data was preprocessed in *Preprocessing.m*. The figures associated with *bankruptcy/bankruptcy.R* are Figure 9 and 11. In *Preprocessing.m* I also generated Figure 10. The remaining figures pertaining to the bankruptcy example – Figure 12 - 15 – are produced by *Postprocessing.m*.

Note that each script is self-contained in the sense that, for the application at hand, it contains all the code used. Only the bankruptcy application consists of three seperate files, as mentioned above. The order of running is *Preprocessing.m*, *bankruptcy.R*, and *Postprocessing.m*. Overall, I recommend running the replication examples first.

Finally, the *surv_rbart* subdirectory contains the package files for *rbart* adapted for probit analysis. Please read the README.md file before using this function in the bankruptcy example (note that I have written some code that already implements the process described here). Moreover, in the README.md file it is also described what changes have been made to the original package.

B Figures

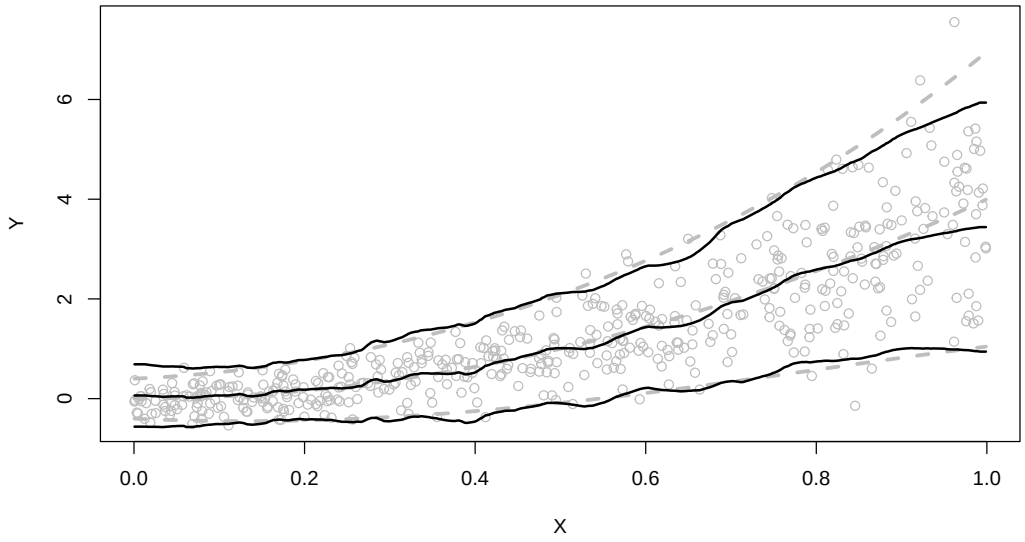

Figure 16: Simulated Example; Prediction Data and Estimates. The dashed lines indicate the true quantities and the solid lines are the estimated ones. For each type of line, the middle line represents the (estimated) mean and the others represent two (estimated) standard deviations away from the mean, in either direction. The estimates are derived from the posterior by taking the mean.

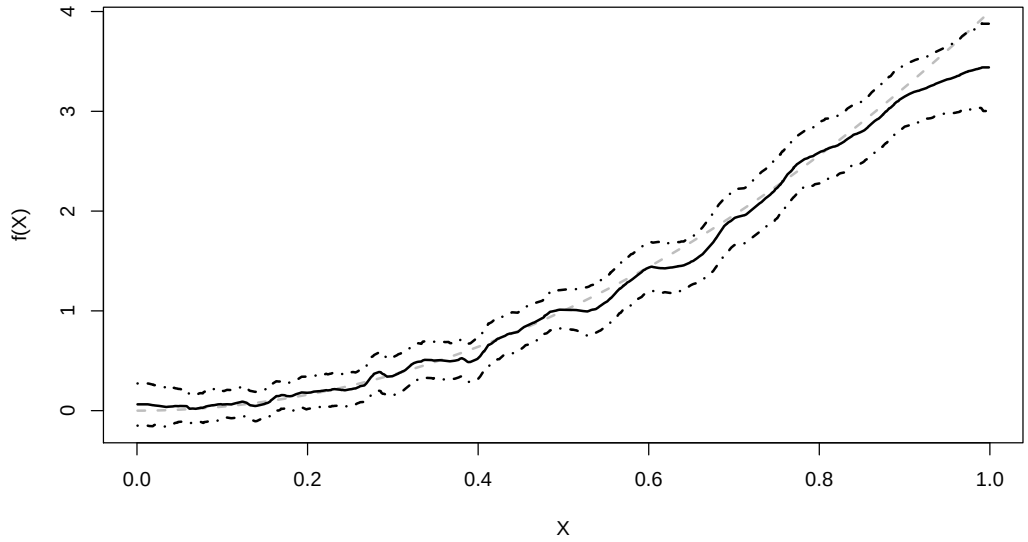

Figure 17: Simulated Example; HBART Estimated Mean. The black line represents the estimated mean, the grey dashed line the true mean, and the black dashed line the point-wise 95% posterior intervals.

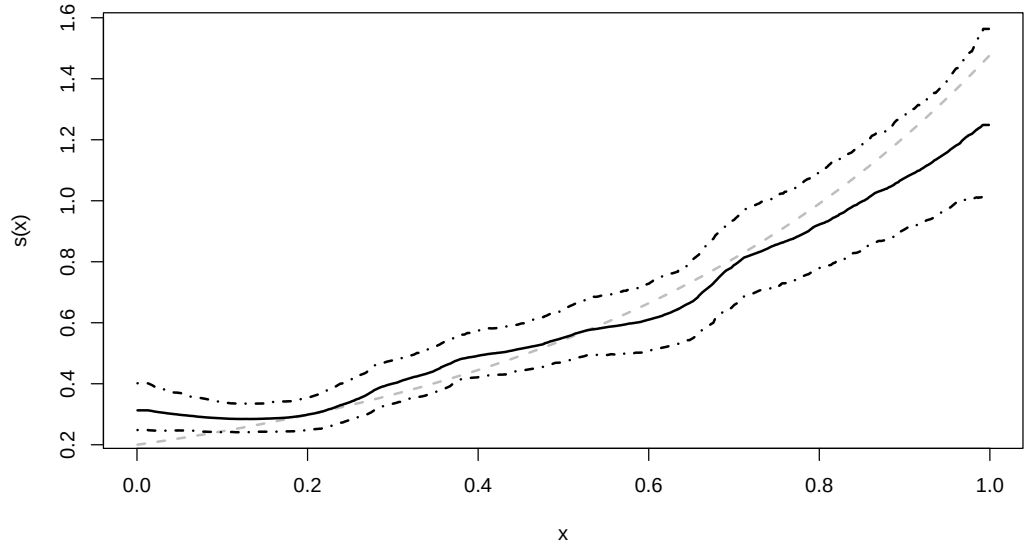

Figure 18: Simulated Example; HBART Estimated Standard Deviation. The black line represents the estimated standard deviation, the grey dashed line the true standard deviation, and the black dashed line the point-wise 95% posterior intervals.

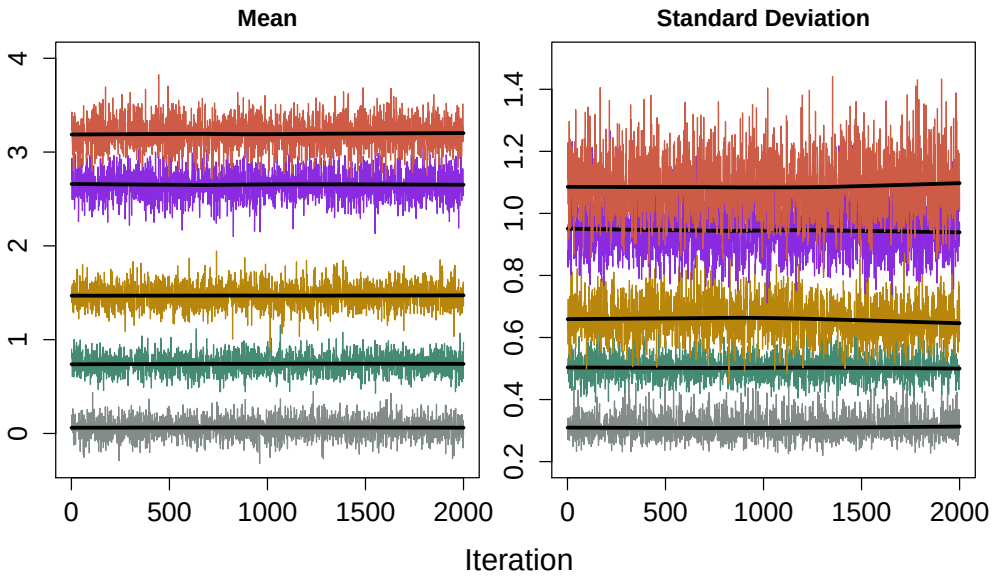

Figure 19: Simulated Example; Traceplots of MCMC Draws. The coloured lines correspond to values for either the mean or standard deviation for the covariate values $X = 0.02, 0.42, 0.63, 0.79, 0.91$, in order from bottom to top. The black lines at the centre of each coloured line represent the corresponding lowess curve (scatterplot smoothing).

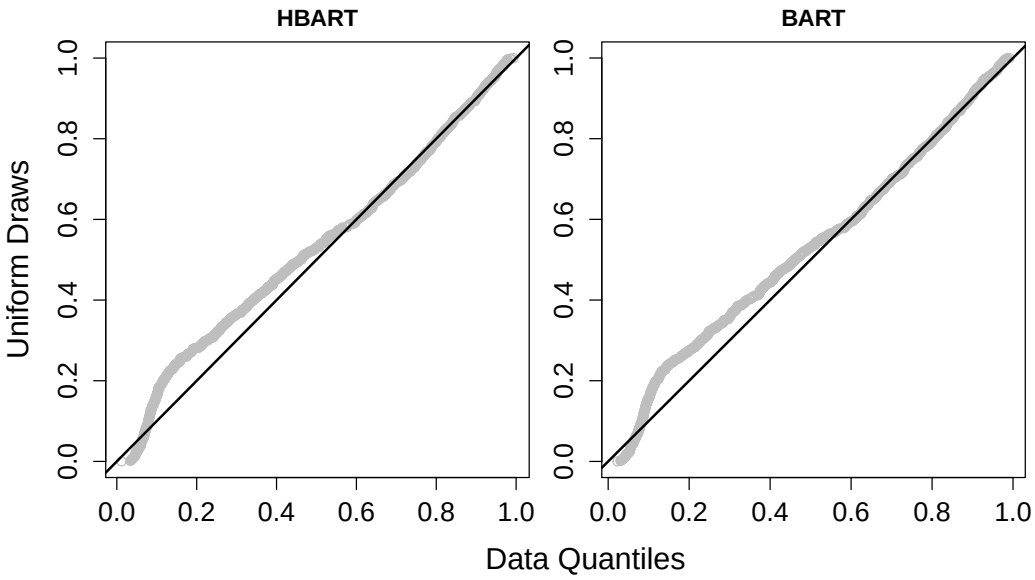

Figure 20: Alcohol Example; Predictive QQ-Plots.

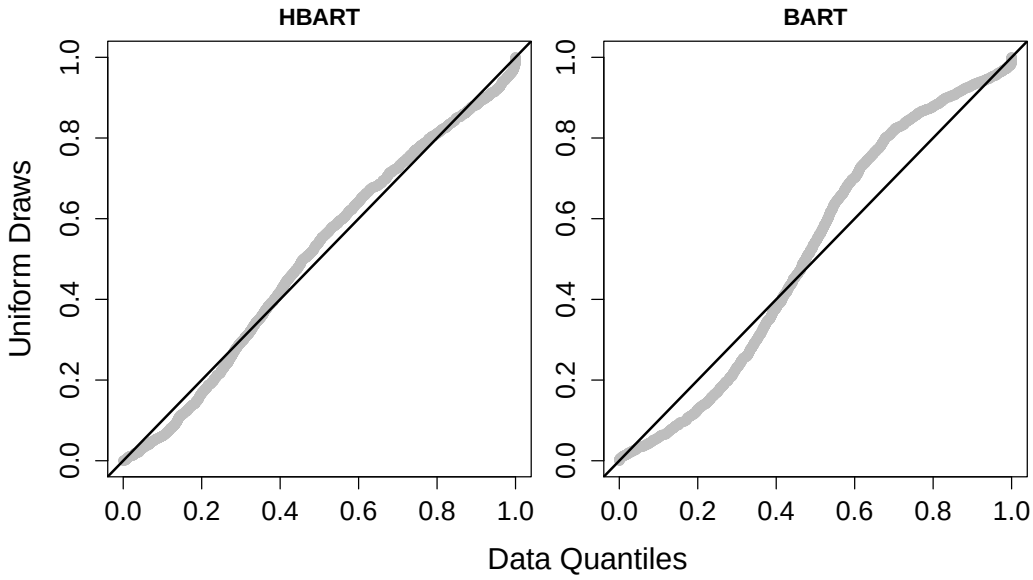

Figure 21: Fishery Example; Predictive QQ-Plots.

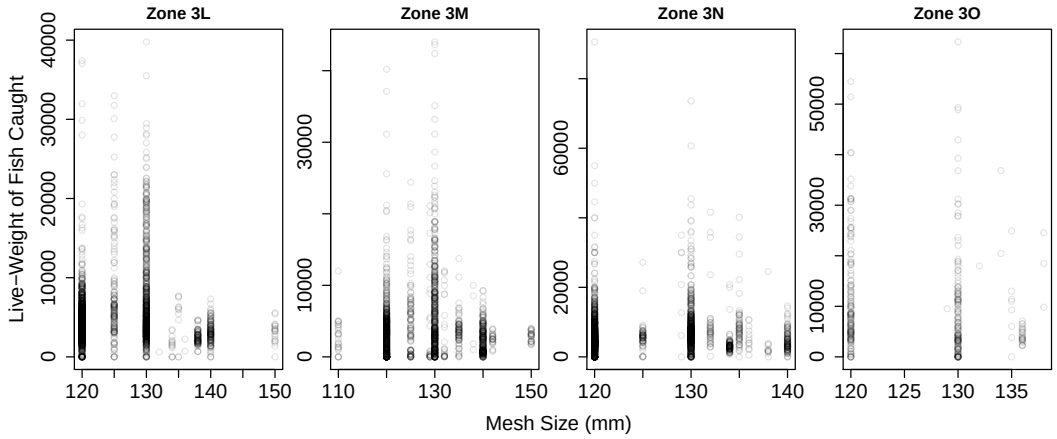

Figure 22: Fishery Example; Fishing Yield per Mesh Size for Different Zones. This plot dissects the fishing yield as a function of mesh size even further. It is not directly evident from this figure that there is heterogeneity in catching fish as a result of an adjustment to the mesh size in different zones. A motivation for plotting these results is the suggestion that some zones may have a higher concentration of smaller fish, which would confound the heterogeneity in yield as a result of different mesh sizes. It turns out that this figure cannot provide any evidence for this, as the yield for the same mesh size in each subfigure seems to be the same. On the other hand, it is interesting to note that mesh sizes of 110mm and 150mm are only used in zones 3L and 3M, respectively.

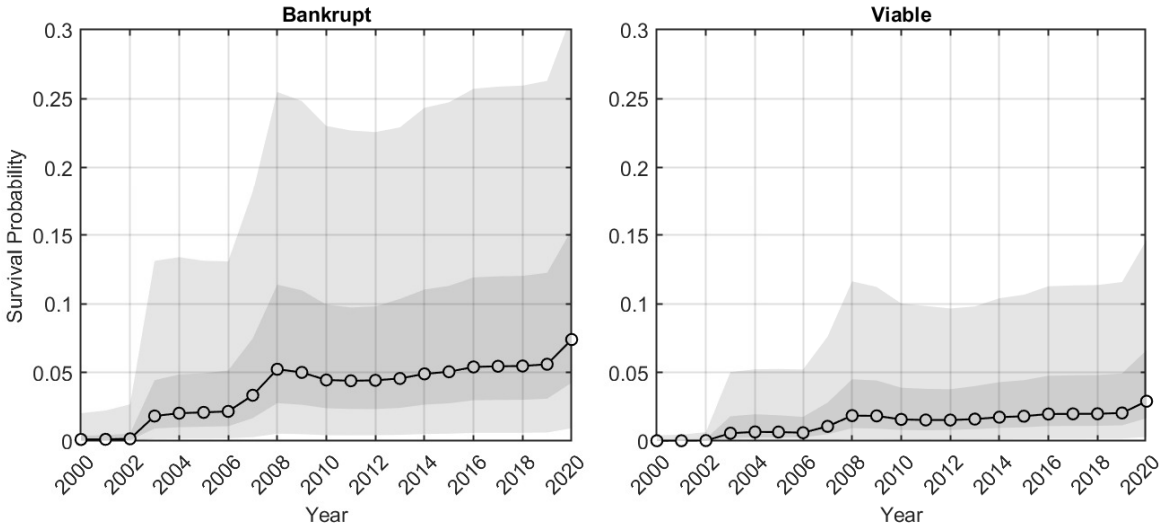

Figure 23: Hazard Function for Bankrupt and Viable At-Risk Firms. The dashed-dotted line in the first panel represents the fraction of at-risk firms that went bankrupt in the corresponding year. Please note the difference in the scale for the in default probability for the two panels.