ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Business Analytics & Quantitative Marketing

Estimation of Sample Selection and Heterogeneous Treatment Effects using Bayesian Additive Regression Trees

Name student: Quinten Obbink

Student ID number: 474183

Supervisor: Dr. Eoghan O'Neill

Second assessor: Dr. Mikhail Zhelonkin

Date final version: 13-6-2024

Abstract

This research extends the existing sample selection and treatment effects (SSTE) model of Vossmeyer (2016) to estimate Heterogeneous Treatment Effects (HTEs), using Bayesian Additive Regression Trees (BART). In contrast to the existing SSTE model, selection is indicated by a binary variable instead of nonzero values of a censored continuous variable, resulting in a model new to the existing literature. Therefore, a novel estimation procedure is constructed, based on a Random-Walk Metropolis-Hasting algorithm, which addresses the identification issues in combination with binary variables. The resulting SSTE-BART model is also extended to include soft trees and sparse splitting rules. The model is evaluated using an extensive simulation study, along with an application to the National Supported Work (NSW) Demonstration dataset. Both analyses reveal that implementation of soft trees and sparse splitting rules increase performance of the SSTE-BART model, and can accurately estimate the covariance matrix. Moreover, the simulation study confirmed the ability of the model to estimate HTEs accurately.

Contents

1 Introduction

Bayesian causal inference is becoming an increasingly popular aspect of modern day scientific research. Even though much research has been done in this area, it still remains a challenging task. One popular approach for deriving Bayesian causal inference is through the potential outcomes framework. In short, within this framework a pre-specified treatment is used to estimate the potential outcomes of getting the treatment or not. However, only one of those outcomes can be observed at the same time, the other is called the counterfactual.

Previous research has shown that Bayesian estimation of treatment effects is able to provide useful insights for causal inference (Heckman et al. 2014; F. Li et al. 2023). However, estimation of these effects can be biased if there is additional sample selection present (Vossmeyer 2016). If sample selection is ignored, it can lead to biased results as the sample is not representative of the population of interest. Therefore, Vossmeyer (2016) proposes a Bayesian model capable of handling sample selection while still estimating treatment effects. The resulting sample selection and treatment effects (SSTE) model consists of a system of five equations: one sample selection function, one treatment selection function, and three response outcomes for the different potential outcomes (nonselected, selected untreated, selected treated).

Even though the SSTE model is able to estimate treatment effects, the effect of treatment is assumed to be homogeneous across the entire treated sample. In some cases this could be true, but in general this assumption is restrictive, as the effect of treatment can significantly differ between different samples. For that reason, this research will extend the already existing SSTE model to enable estimation of heterogeneous treatment effects. Furthermore, another adjustment to the original SSTE model is made for the assumption that the sample and treatment selection variables are continuous. In practice, it is more common to have binary selection variables, so this will be implemented to develop a more generally applicable model.

To achieve a model structure capable of estimating heterogeneous treatment effects, a well-established approach for Bayesian causal inference is adopted, called Bayesian Additive Regression Trees (BART) (Chipman et al. 2010). BART is an ensemble method that uses a summation of multiple unique regression trees to obtain a flexible model. Regularization is performed on each tree to ensure that individual effects of a single tree are not overwhelmingly influential on the model fit. However, standard BART was developed for univariate models, such that an adjustment is required for a system-of-equations like the SSTE model. This multivariate version of BART was developed by Chakraborty (2016), and is based on the concept of Seemingly Unrelated Regression (SUR) (Zellner 1962). The resulting model is therefore called SUR-BART, and will be the foundation of the adjusted SSTE model.

The specification of binary selection variables adds another level of complexity to the model. In any model with potential outcomes, there is the problem of a not fully identified covariance matrix of the model. This follows from the fact that not all equations are observed simultaneously, as only one of the potential outcomes is observed at the same time. As a result, the corresponding covariances are not identified. Approaches to correctly handle these unidentified elements have been developed (Chib 2007; Chib et al. 2009), and have even been proposed for the original SSTE model (Vossmeyer 2016). However, it has been stated that these approaches fail to work for binary selection variables (Chib et al. 2009), but no concrete solutions yet exist. Therefore, following suggestions from Chib et al. (2009) and P. Li (2011), a novel approach to sample the covariance matrix is constructed using an adjusted Random-Walk Metropolis-Hasting algorithm as mentioned in Chib and Greenberg (1998).

Following from this, the aim of this research can be summarized into the following research question: *"Can heterogeneous treatment effects be captured in the sample selection and treatment effects model, using Bayesian Additive Regression Trees?"*

The performance of the constructed model, which is referred to as the SSTE-BART model, will be evaluated in comparison to other existing models. In addition, some improvements to the SSTE-BART model will be tested, following the approach to a BART-based selection model by O'Neill (2024). Specifically, the SSTE-BART model will be tested for the addition of soft decision trees, as well as sparse splitting rules on the splitting probabilities of the decision rules, as first introduced by Linero and Yang (2018).

The resulting SSTE-BART model has some advantages over other BART models, as there is correlation between treatment selection and the potential outcomes, and even between sample selection and treatment selection. This is not apparent in most other BART treatment effect estimation methods, where there are often no unobservables that can impact both sample selection and treatment selection.

For evaluation of the models an extensive simulation study is conducted, following a similar set-up as executed in Chakraborty (2016). Then the performance of SSTE-BART is also tested on a real-life dataset, namely the National Supported Work (NSW) Demonstration dataset.

Results from the simulation study reveals the addition of sparse splitting rules significantly improves performance of the SSTE-BART model. In addition, the SSTE-BART model has better predictive performance when the underlying data generating process (DGP) is relatively simple compared to a more complex DGP. Furthermore, the implementation of soft trees in the SSTE-BART model, referred to as the SSTE-SoftBART model, outperforms the regular SSTE-BART model. Achieving an average decrease in Mean Squared Error (MSE) across all

outcome predictions, of approximately 22.4% for the complex DGP, and approximately 6.9% for the simple DGP. Moreover, runs of the standard versions of BART and SoftBART obtain predictive results comparable to their SSTE counterparts, indicating there is much room for improvement in the SSTE-BART models. Subsequently, the SSTE-SoftBART model is able to estimate the elements of the covariance matrix fairly accurately and with reasonable convergence properties. Calculations of the treatment effects provide evidence that the SSTE-BART models are indeed able to accurately estimate heterogeneous treatment effects.

The application of the SSTE-BART model to the NSW dataset provides more evidence that the SSTE-SoftBART variation has better predictive performance than the standard SSTE-BART model. However, the models are not able to achieve reasonable MCMC convergence, making it difficult to extract useful insights.

The remainder of this paper is structured as follows: Section 2 contains an overview of the literature surrounding this research. This is followed by a detailed discussion of the required methodology in Section 3. After this, Section 4 introduces the SSTE-BART model, along with the developed estimation procedure. The setting and results of the simulation study are provided in Section 5. Afterwards, Section 6 displays the results of the application to the NSW dataset. Finally, a comprehensive discussion and conclusion of the entire research is provided in Sections 7 and 8.

2 Literature Review

The following section will provide an overview of the existing literature surrounding the topics of sample selection and treatment effects models, Bayesian causal inference, Bayesian Additive Regression Trees, and Markov Chain Monte Carlo (MCMC) sampling algorithms.

At the foundation of a majority of existing research into causal inference, the framework of potential outcomes can be found. The potential outcomes framework was first proposed by Neyman (1923), however only in the context of randomized experiments. This was later extended into the current general framework, widely known as the Rubin Causal Model (RCM) (Rubin 1974), which can be used for both observational and experimental studies. Under this framework, causal effects can be defined as "comparisons of potential outcomes under different treatments on a common set of units" (Rubin 2005).

Using the potential outcomes framework as a basis, models focused on estimating the effect of treatment were developed. An early adaptation is the classic Roy model (Roy 1951), which has been the foundation for other developments in the area of treatment effect estimation (Heckman and Honore 1990; Heckman and Vytlacil 2007). A discussion of the relationship between the

Roy model and other potential outcomes models can be found in Heckman (2008). Another novel application of the potential outcomes framework was introduced by Angrist et al. (1996), where instrumental variables are added to the RCM to identify causal effects.

The majority of these developed methods can be attributed to the Frequentist paradigm, which have established themselves among the most popular approaches in scientific research. This is mostly because the alternative, Bayesian methods, were quite difficult to apply in practice, and require a completely different way of thinking. The disparity between the use of Frequentist and Bayesian methods was noticed by Heckman et al. (2014), who introduce an accessible Bayesian approach for calculating treatment parameters. M. Li and Tobias (2014) also contributed to the growing literature of Bayesian techniques by discussing treatment parameter estimation through the popular Markov Chain Monte Carlo (MCMC) methods.

One of the major advantages of Bayesian approaches for causal inference is the ability to incorporate prior knowledge and/or beliefs into the analysis. This can lead to improved accuracy and efficiency, and results in a better way of handling missing values, as prior information can be leveraged to impute values. In addition, modeling uncertainty in parameters also becomes possible when taking a Bayesian approach (Beck and Katafygiotis 1998).

As a result, research into Bayesian inference for treatment effect estimation has greatly risen in popularity. Vijverberg (1993) was motivated by the fact that in treatment effect models, the counterfactual is never observed. A model was proposed based on the joint distribution of the potential outcomes, where the counterfactuals are actually sampled from a posterior distribution. Conversely, Chib (2007) presented a Bayesian framework without the joint distribution of the potential outcomes. This approach does not require simulating the counterfactuals, which leads to less complex computations and easier prior specifications.

While classic estimation methods of sample selection were already established in the Frequentist literature (Gronau 1974; Heckman 1976), Bayesian frameworks were still underrepresented. However, since then more and more Bayesian methods for sample selection have been developed and empirically tested (Omori 2007; Chib et al. 2009; Van Hasselt 2011).

A consensus was reached in the literature that the sample selection and treatment effect estimation problems share a close affinity (Manning et al. 1987; Leung and Yu 1996). Winship and Mare (1992) was an early adopter of this notion and proposed an approach capable of joining the two methods. Following this, Lee (2012) introduces a nonparametric method for sample selection and treatment effect analysis, hosting several advantages over the usual matching methods. Alternatively, Huber (2014) presents a parametric estimation method, capable of identifying treatment effects under sample selection, calling it a double selection problem.

Within the Bayesian paradigm, a recent paper introducing a model for this double selection problem was Vossmeyer (2016). Even though the model was initially made for a specific application in the financial sector, it is stated that it is not limited to such settings. Vossmeyer (2016) extends the techniques on sample selection proposed by Chib (2007) and Chib et al. (2009), and goes into great depth to design a computationally efficient estimation algorithm, capable of jointly estimating sample selection and treatment effects. This results in the Sample Selection and Treatment Effects (SSTE) model. One shortcoming of the SSTE model, is that it assumes homogeneous treatment effects for every observation in the treated sample. This is potentially restrictive, as the treatment effects might be heterogeneous across the treated sample.

Heterogeneity in treatment effects has been studied before, for example in Hill (2011) where a Bayesian nonparametric modeling procedure is developed, or in Green and Kern (2012) where heterogeneous treatment effects in survey experiments are modeled. More recently, the Causal Forest algorithm was introduced, which is also capable of estimating heterogeneous treatment effects (Wager and Athey 2018). However, in current literature there exists very few research into heterogeneous treatment effects in combination with sample selection. This makes any research into this subject very relevant, and is the reason why it is the focus of this paper.

To try and achieve useful insights for this topic, this research will extend and try to improve the framework proposed by Vossmeyer (2016) to capture heterogeneous treatment effects. This is done by replacing the linear structure of the original model specification with Bayesian Additive Regression Trees (BART) (Chipman et al. 2010). BART is an ensemble method using regression trees and is able to provide robust estimates of complex structures in data. Since its introduction, BART has established itself as one of the most effective and reliable methods in the field of causal inference (Hill et al. 2020). Notably, BART is also applied in the context of heterogeneity of treatment effects (Hill 2011; Green and Kern 2012). Replacing the linear structure of a model with BART has been proven to be successful before, like for instrumental variable models (McCulloch et al. 2021; Spanbauer and Pan 2022). Furthermore, the sample and treatment selection variables are binary instead of censored continuous. Additionally, the response variables for the potential outcomes are assumed to be continuous and unbounded, instead of censored like in Vossmeyer (2016). These statements make a convincing argument to try and improve the original model while also estimating heterogeneous treatment effects.

An issue that arises with the use of BART, is that it is originally made for univariate models, and the sample selection and treatment effects model consists of a system of equations. Therefore, a BART-based approach specifically made for multivariate systems of equations will be used, called Seemingly Unrelated Regression BART (SUR-BART) (Chakraborty 2016). SUR-

BART uses the concept of Seemingly Unrelated Regression by Zellner (1962), which is based on the assumption that because different responses are extracted from the same individual, they have a high chance of being related through an underlying process. Hence, this research incorporates SUR-BART to ensure the correlation in responses is modeled appropriately.

Due to the size of the parameter space resulting from this model, it is common practice to estimate the model using MCMC methods (Vossmeyer 2016; Chakraborty 2016). Conversely, as a result of the added assumption of binary selection variables to the SSTE model, the developed estimation procedure for the covariance matrix by Vossmeyer (2016) can not be used. For fully identified covariance matrices, estimation approaches exist that can deal with such element-wise restrictions (Chan and Jeliazkov 2009). There are even some papers that construct methods which are able to estimate covariances for two selection mechanisms (P. Li 2011; Ding 2014). However, the combination of unidentified elements and binary response variables results in complications. This is further confirmed in Chib et al. (2009), where it is suggested that for this combination, a Metropolis-Hasting algorithm like Chib and Greenberg (1998) should be used. Hence, an adjusted version of the Metropolis-Hasting algorithm from Chib and Greenberg (1998) will be incorporated in the MCMC estimation procedure. Another option could have been the Parameter-Expanded Metropolis-Hasting (PX-MH) algorithm, developed by Zhang et al. (2015). This algorithm is able to estimate the covariance matrix for a mix of binary and continuous variables, while also being able to handle missing values. However, implementation of the PX-MH algorithm for this model brings additional complications, therefore it is not used in this research. The implemented tree sampler for the BART-based algorithms is not required to change much, such that a similar estimation procedure as in Chakraborty (2016) can be used.

Summarizing, this research adds to the already existing literature in multiple aspects. Firstly, the already existing SSTE model is adjusted by replacing the linear structure of the model with BART. This alteration brings more insights in, and stresses the importance of, the joint estimation of sample selection and treatment effects. Alongside of this, the possibility to now estimate heterogeneous treatment effects is a relevant contribution to the area of sample selection. Furthermore, the additional restriction of binary response variables results in the construction of a novel approach for the covariance matrix sampling procedure. Since there is no concrete existing method for this scenario, this research provides information towards a better understanding of such sampling procedures. Lastly, any research into making Bayesian causal inference more accessible is relevant, as it has proven itself to be more effective in certain scenarios. Ultimately, this research will hopefully add to the understanding of Bayesian causal inference in the area of sample selection and treatment effects.

3 Methodology

This section will contain a detailed discussion of the methodological aspects of this research. First, the original Sample Selection and Treatment Effects model by Vossmeyer (2016) will be provided to illustrate the general framework of the model. After this, a brief explanation of the concept of Bayesian Additive Regression Trees (BART) will be provided, to build the foundation for the Seemingly Unrelated Regression BART (SUR-BART) algorithm. This also includes the required prior specifications, as well as a brief mention of the necessary MCMC sampler.

3.1 Original model by Vossmeyer (2016)

The foundation for this research is provided by the framework proposed by Vossmeyer (2016), so its theoretical background will be discussed next. The main idea of the framework is that, in the presence of sample selection and treatment selection, potentially on unobservables, both of these aspects need to be modeled simultaneously. If sample selection is ignored and analysis is done on only the selected sample, inference will be based on a non-representative sample of the population of interest, which leads to specification errors (Vossmeyer 2016). Figure 1 displays a graphical representation of the resulting model. It is important to note, that it is unusual in sample selection models, that the non-selected sample is observed, which makes the SSTE model quite unique. Nevertheless, there exists many settings in which the model can be applied, such as the banking context discussed in Vossmeyer (2016), where the performance of a bank is evaluated based on an application for financial assistance or not. Moreover, models in which the non-selected sample is not observed can still be formulated into an SSTE model, through a combination of the Heckman and Roy selection models.

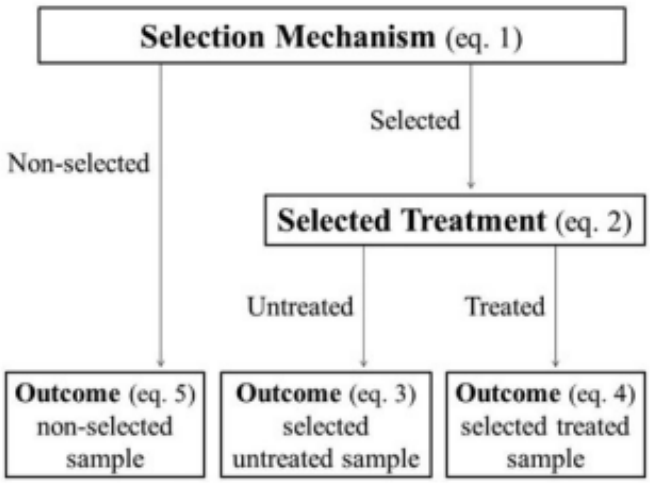

Figure 1: Multivariate treatment effect model in the presence of sample selection (Vossmeyer 2016)

The representation shows the two selection mechanisms: one for selection into the sample (sample selection), and one for the treatment assignment (treatment selection). Three potential outcomes can be extracted as a result: nonselected, selected untreated, and selected treated. This graphical structure can be translated into a system of equations, shown below.

$$\text{Selection mechanism}: \; y_{i1}^* = \boldsymbol{x}_{i1}^{'}\boldsymbol{\beta}_1 + \varepsilon_{i1} \tag{1}$$

$$\text{Treatment selection}: \; y_{i2}^* = \boldsymbol{x}_{i2}^{'}\boldsymbol{\beta}_2 + \varepsilon_{i2} \tag{2}$$

$$\text{Selected untreated sample}: \; y_{i3}^* = (\boldsymbol{x}_{i3}^{'} \; y_{i1})\boldsymbol{\beta}_3 + \varepsilon_{i3} \tag{3}$$

$$\text{Selected treated sample}: \; y_{i4}^* = (\boldsymbol{x}_{i4}^{'} \; y_{i1} \; y_{i2})\boldsymbol{\beta}_4 + \varepsilon_{i4} \tag{4}$$

$$\text{Nonselected sample}: \; y_{i5}^* = \boldsymbol{x}_{i5}^{'}\boldsymbol{\beta}_5 + \varepsilon_{i5} \tag{5}$$

The model contains two definitions for the dependent variable, namely $\boldsymbol{y}_i^* = (y_{i1}^*, y_{i2}^*, y_{i3}^*, y_{i4}^*, y_{i5}^*)^{'}$ and $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5})^{'}$, corresponding to the latent data and the observed data, respectively. The general framework is able to handle continuous, binary, censored and ordered variables. The continuous setting is specified as $y_{ij} = y_{ij}^*$ and the binary one as $y_{ij} = 1\{y_{ij}^* > 0\}$. However, Vossmeyer (2016) specifies a setting which is a combination of continuous and binary, resulting in a censored outcome (Tobin 1958), defined as $y_{ij} = y_{ij}^* \cdot 1\{y_{ij}^* > 0\}$. This setting is not entirely clear, so for the new model an understandable relation will be discussed.

Even though the framework contains five equations, the system of equations is reduced to two or three equations, depending on the subsample, as a result of the selection mechanisms and the subsequent unobserved potential outcomes. For example, if $y_{i1} = 0$ for some i, then the observation is in the nonselected sample. This means y_{i1} and y_{i5} are observed, but the other three are not. Additionally, if $y_{i1} > 0$ and $y_{i2} = 0$, the observation belongs to the selected but untreated sample. Here, y_{i1}, y_{i2} and y_{i3} are observed, and the other two are not.

The framework also includes exogenous covariates, defined as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4}, \boldsymbol{x}_{i5})$. Here, $\boldsymbol{x}_{ij}$ contains the covariates corresponding to equation j, and are thus only required when equation j is observed. The covariates in $\boldsymbol{x}_{i2}$ are assumed to contain at least one more variable than the covariates in other equations. This additional variable serves as the instrumental variable, used in treatment effect models, that is correlated with the treatment but not with the errors (Chib 2007). A similar additional variable requirement can be found in the sample selection equation for many other sample selection models, which is not present in the original SSTE model. This requirement is often referred to as an exclusion restriction, and it helps to improve identification and estimation of the model. Hence, for the new model an additional variable will also be added to the covariates $\boldsymbol{x}_{i1}$.

Lastly, the errors, defined as $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4}, \varepsilon_{i5})$, are assumed to have a multivariate

normal distribution $\mathcal{N}_5(0, \mathbf{\Omega})$. This assumption provides the most flexible foundation, as it can be relaxed by assuming a t-distribution or a mixture of normal distributions.

3.2 Bayesian Additive Regression Trees

The first major change this research makes to the model by Vossmeyer (2016), is the replacement of the linear structure of the model with Bayesian Additive Regression Trees (Chipman et al. 2010). As mentioned earlier, the original model assumes a homogeneous effect of treatment across all observations in the treated sample. However, this is quite a restrictive assumption, as observations can differ greatly in characteristics such that their response to the treatment is not the same. The implementation of Bayesian Additive Regression Trees, or simply BART, does allow for the estimation of heterogeneous treatment effects. Traditional BART consists of two parts that require explanation: a sum-of-trees model, and regularization priors on the parameters of the model. Before these two parts are discussed, first the concept of regression trees will be briefly mentioned, to be able to fully understand the sum-of-trees model.

3.2.1 Regression Trees

Regression trees belong to the family of decision trees and exhibit similar characteristics. The general idea being that the data is partitioned into different subsets based on some splitting rule. For regression trees, these splitting rules are placed directly at the interior nodes, and can be defined as binary splits of the covariate space. This is why regression trees are also referred to as binary regression trees. These rules are of form $x_i < C$ or $x_i \geq C$, where x_i is often a continuous variable. At each interior node, observations are parsed through to the next node according to the corresponding rule. The final nodes of a tree, known as terminal nodes, contain a parameter value associated with each of the resulting subsets of the covariate space. This parameter represents the prediction of the regression tree, and is traditionally set to be the average of the observations in each subset.

3.2.2 Sum-of-trees model

For the discussion of the following sum-of-trees model, similar notation as in Chipman et al. (2010) will be used. First, a single tree model, using one binary regression tree as explained in the previous section, will be established. After this, the sum-of-trees model follows from a straight-forward summation of many of these single tree models.

Define T as a single binary regression tree, and and $M = \{\mu_1, \mu_2, ..., \mu_b\}$ as the corresponding set of parameter values associated with the terminal nodes of T. Here, each tree T is defined

to have b terminal nodes. The splitting rules within the tree are of identical form as mentioned above. Following from this, each x can be assigned to a single terminal node of T, which contains the value μ_i corresponding with that terminal node. In mathematical notation, this process can be encompassed in a function: for a given T and M, define $g(x; T, M)$ as the function which assigns a $\mu_i \in M$ to x.

The single tree model is then defined as:

$$Y = g(x; T, M) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

From Equation 6, the conditional mean of Y given x is defined as: $E[Y|x] = \mu_i$. The errors are assumed to follow a Normal distribution with mean 0 and variance σ^2. As stated before, the sum-of-trees model is a straight-forward summation of many single trees. Formally, a sum-of-trees model with a number of distinct single trees equal to m can be constructed as:

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

In Equation 7, the function $g(x; T_j, M_j)$ assigns $\mu_{ij} \in M_j$ to x, for each binary regression tree T_j and the corresponding terminal node parameters M_j. The conditional expectation of the sum-of-trees model is given by $E[Y|x] = \sum_j \sum_b \mu_{bj} \mathbb{1}\{i \in \text{terminal node } b\}$. Here, $\mathbb{1}\{\cdot\}$ is an indicator function which ensures that the value associated with a terminal node is only included in the expectation, if the observation is indeed assigned to that terminal node. Generally, a significant advantage of this model structure, is that the sum-of-trees model is able to capture specific variable effects as well as interaction effects (Chipman et al. 2010).

3.3 Seemingly Unrelated Regression BART

One important thing to note, is that the traditional BART approach was developed to fit the univariate case, and as a result does not fit the multivariate case directly. So, in a system of equations where the dependent variables are highly correlated with each other, like the SSTE model (Vossmeyer 2016), standard modeling procedures for BART do not apply anymore. Instead, an approach capable of jointly modeling the correlation structure among the related dependent variables is required. Chakraborty (2016) combines the already existing BART structure with one of the most popular approaches to accurately model systems of equations, called Seemingly Unrelated Regression, SUR in short (Zellner 1962). The concept of SUR is based on the assumption that, because different responses are extracted from the same individual, they have a high chance of being related through an underlying process.

The resulting approach is fittingly called SUR-BART, and its structure looks similar to a five-fold repetition of a traditional BART sum-of-trees model in Equation 7. Specifically, each

equation J has its own sum-of-trees model, which are all captured together inside a vector structure as shown in Equation 8.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m_1} g(\boldsymbol{x}_1; T_j, M_j) \\ \sum_{j=1}^{m_2} g(\boldsymbol{x}_2; T_j, M_j) \\ \vdots \\ \sum_{j=1}^{m_J} g(\boldsymbol{x}_J; T_j, M_j) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_J \end{pmatrix} \tag{8}$$

Here, it is assumed that the errors of the SUR-BART model follow a multivariate normal distribution: $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_J) \sim \mathcal{N}_J(\mathbf{0}, \boldsymbol{\Sigma})$.

3.3.1 Regularization Priors

As mentioned before, the second part of a traditional BART approach is setting regularization priors on the parameters of the model. This is also the case for the SUR-BART approach, so this will be discussed next. Regularization priors are a crucial and key aspect of BART approaches, as they are responsible for regularizing the fit of the tree structure, by limiting the individual effects of each tree. The aim of these priors is to prevent large tree elements from exerting overwhelming influence on the structure of the tree, which would significantly reduce the effectiveness of an additive model representation. The prior specifications for this research follow the distributions and structure used for SUR-BART (Chakraborty 2016). which follow the original BART specifications by Chipman et al. (2010). Additionally, as is standard practice for all Bayesian analyses, specification of priors is used to incorporate prior knowledge on the data into the model to increase the model fit.

Chipman et al. (2010) simplifies the specification of these priors by imposing an independence restriction on the prior parameters. This restriction leads to independence: between all tree components (T_j, M_j), between all (T_j, M_j) and σ, and between the terminal node parameters of every tree. As a result, priors only need to be specified on three aspects: T_j, $\mu_{ij}|T_j$ and σ (Chipman et al. 2010).

The first two aspects are identical for traditional BART as for SUR-BART, but the prior on σ is necessarily different. Instead of a single variance, σ, there now exists an entire JxJ covariance matrix, Σ, which requires an inherently different prior distribution. All default prior distributions will be discussed below, while those for the new model are discussed later.

The prior on the trees T_j, written as $p(T_j)$, is constructed in three parts: (i) the probability that a node at depth d $(d = 0, 1, 2, ...)$ is nonterminal, (ii) the distribution on the assignments of splitting variables at each interior node, and (iii) the distribution on the assignments of splitting rules in each interior node, conditional on the splitting variable. Each aspect has

a distinct distribution, for which the derivation and additional motivation can be found in Chipman et al. (1998). Since the derivations themselves have no added value in this research, they are omitted from this paper and the formula will simply be given. The first aspect, (i) is given by the following equation:

$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1),\ \beta \in [0,\infty) \tag{9}$$

The default values for Equation 9 proposed by Chipman et al. (2010), are $\alpha = 0.95$ and $\beta = 2$. These values strongly favor small trees of size 2 or 3, and have proven to be effective in practice. The prior on the second aspect, (ii), is given by an uniform distribution on the available variables. Similarly, the prior on the third aspect, (iii), is given by an uniform distribution on the discrete set of available splitting values.

The priors over the leaf parameters, $p(\mu_{ij}|T_j)$, are constructed using a conjugate normal distribution. The idea is to shift and scale the dependent variable Y, such that the new values are bounded by the interval $[-0.5, 0.5]$. Then, after centering the prior for μ_{ij} around zero, the following distribution can be set:

$$\mu_{ij} \sim \mathcal{N}(0, \sigma_\mu^2), \qquad \text{where } \sigma_\mu = 0.5/(k\sqrt{m}) \tag{10}$$

The parameter k can be interpreted as the number of prior standard deviations in the range of $[0, 0.5]$. For example, in the default specification of $k = 2$, proposed by Chipman et al. (2010), a 95% prior probability is assigned that $E[Y|x]$ lies in the range of $[-0.5, 0.5]$ (Hill et al. 2020).

The priors on the elements on the covariance matrix are inherently different for the univariate and multivariate case. Chipman et al. (2010) apply a data-informed prior on the univariate variance, using the estimated residual standard deviation from a simple linear regression. It is also shown that the use of data-informed priors leads to improved performance, if calibrated correctly. On the other hand, Chakraborty (2016) puts an Inverse Wishart distribution on the entire covariance matrix, using hyperparameters to produce an uninformative prior structure. The Inverse Wishart distribution is commonly used when working with covariances, as it is a conjugate prior for covariance matrix of the multivariate normal distribution. For additional details on these default settings, the reader is referred to the corresponding papers.

Another important aspect of any additive tree algorithm, is the choice of the number of trees m it uses. In a BART-based algorithm, each tree is iteratively updated at each step, so more trees lead to more necessary computations. However, more trees also result in an improved fit of the model and increases its predictive performance. This leads to a trade-off between computational efficiency and predictive performance. Chipman et al. (2010) noted that the predictive performance of the algorithm improves significantly with each increase in m, until

it levels off at some point and then even decreases for very large values of m. A suggestion for a default choice of $m = 200$ is then made, as it provides good results in their experiments. Even though it should be sufficient in most cases, Chipman et al. (2010) emphasize that cross-validating the number of trees is another valid approach to choose from. In line with this, Chakraborty (2016) argue that a flexible specification, where the number of trees is treated like a parameter, actually increases effectiveness and efficiency. Their SUR-BART algorithm therefore includes a component which automatically selects an optimal number of trees for the model. From this it is a logical conclusion that the choice of number of trees is an important consideration in the model specification, depending on the goal and/or desired result.

3.3.2 Posterior information extraction

The final step towards inference is combining these priors with the likelihood function of the model to construct the posterior distribution. The posterior distribution of the sum-of-trees model, which is given by $p((T_1, M_1), ..., (T_m, M_m), \sigma|y)$, contains all the unknown parameters the model, and provides the basis for all inferential techniques. However, as a result of the many unknown parameters, obtaining the relevant statistics is almost always very computationally expensive. That is why, for BART models it is standard practice to achieve inference through Markov Chain Monte Carlo (MCMC) sampling, such as Metropolis-within-Gibbs algorithms (Hill et al. 2020). Specifically, a Bayesian backfitting algorithm was developed by Chipman et al. (2010), which was adjusted to fit the multivariate case by Chakraborty (2016). Only the general framework of this approach will be provided next, as a detailed and updated version will be extensively discussed when introducing the new model.

The sampling procedure introduced by Chipman et al. (2010) is in essence a Gibbs sampler. The basic idea behind a Gibbs sampler is that for a set parameter space, instances of each variable are sampled iteratively from their respective distributions, conditional on the current values of the other variables. Furthermore, it was shown in Hastie and Tibshirani (2000) that a Gibbs sampler for additive models with variances fixed, is a stochastic generalization of a backfitting algorithm (Breiman and Friedman 1985). Hence, the developed sampling procedure is referred to as a Bayesian backfitting algorithm.

This algorithm can be divided into two main steps, given below. Here, $T_{(j)}$ is defined as the set of all trees except T_j, with $M_{(j)}$ similarly defined for the sets of terminal node parameters. As a result, $T_{(j)}$ is a set containing m-1 trees.

1. First, draw (T_j, M_j) conditionally on $(T_{(j)}, M_{(j)}, \sigma)$, successively for j=1,...,m:
   $(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y$,

2. Then draw σ from the full conditional:

   $\sigma|T_1, ..., T_m, M_1, ..., M_m, y.$

The result of the backfitting algorithm is a series of draws of $(T_1, M_1), ..., (T_m, M_m), \sigma$ which, in distribution, should converge to the posterior $p((T_1, M_1), ..., (T_m, M_m), \sigma|y)$. Therefore, after running the algorithm for a sufficient amount of iterations, the series of draws can be regarded as an approximate sample from the true posterior. A variety of Bayesian inferential metrics can then be approximated using this sample, which are further elaborated in Chipman et al. (2010).

The sampling procedure for SUR-BART consists of the same general steps as BART, but the conditional distributions which are sampled from are slightly different. Specifically, the tree parameters are now sampled, conditional on all additional other sums of trees. Similarly, the draw of the covariance matrix is one direct draw from the Inverse Wishart distribution, where the hyperparameters are constructed using all sums of trees.

4 SSTE-BART

This section will introduce the adjusted SSTE model, starting with the model specification itself. Then, the model will be translated using a SUR structure, specifically SUR-BART, to tackle the problem of correlated dependent variables in a system of equations. The resulting adjusted SSTE model will be referred to as SSTE-BART from now on. After this, the necessary regularization priors will be discussed for the different parameters and aspects of the model. Finally, the entire MCMC sampling procedure will be discussed extensively, as the adjusted model requires a new procedure to be constructed to be able to handle the mentioned alterations.

4.1 Model specification

The adjusted model will consist of the same five distinct equations for the selection mechanisms and the potential outcomes as in Vossmeyer (2016). The system of equations that follows from the addition of BART terms is given below, for observations $i = 1, ..., n$. These equations are different from the original SSTE model in two ways. First, and most obvious, each SSTE-BART equation now is of sum-of-trees form as in Equation 7. Second, the variables y_{i1} and y_{i2} are assumed to binary in SSTE-BART, instead of continuous in the original SSTE model. As a result, these variables are not included as endogenous covariates in Equation 14 and 15. This is done, because now these variables will always have the same value for all observations in the subset associated with that equation. Hence, inclusion of the variable to these equations is

redundant, as they are not able to provide any additional information.

$$\text{Selection mechanism}: y_{i1}^* = \sum_{h=1}^{m_1} g_1(\boldsymbol{x}_{i1}^{'}; T_h, M_h) + \phi_1 + \varepsilon_{i1} \tag{11}$$

$$\text{Treatment selection}: y_{i2}^* = \sum_{h=1}^{m_2} g_2(\boldsymbol{x}_{i2}^{'}; T_h, M_h) + \phi_2 + \varepsilon_{i2} \tag{12}$$

$$\text{Selected untreated sample}: y_{i3}^* = \sum_{h=1}^{m_3} g_3(\boldsymbol{x}_{i3}^{'}; T_h, M_h) + \varepsilon_{i3} \tag{13}$$

$$\text{Selected treated sample}: y_{i4}^* = \sum_{h=1}^{m_4} g_4(\boldsymbol{x}_{i4}^{'}; T_h, M_h) + \varepsilon_{i4} \tag{14}$$

$$\text{Nonselected sample}: y_{i5}^* = \sum_{h=1}^{m_5} g_5(\boldsymbol{x}_{i5}^{'}; T_h, M_h) + \varepsilon_{i5}. \tag{15}$$

The relation between the latent data y_{ij}^* and the observed data y_{ij} is also specified differently, as these relations from the original SSTE model are not made entirely clear. Under the assumption that y_{i1} and y_{i2} are binary, the following relations are defined:

$$y_{i1} = \begin{cases} 1, & \text{if } y_{i1}^* > 0 \\ 0, & \text{if } y_{i1}^* \leq 0 \end{cases}, y_{i2} = \begin{cases} 1, & \text{if } y_{i1} = 1, y_{i2}^* > 0 \\ 0, & \text{if } y_{i1} = 1, y_{i2}^* \leq 0 \\ \texttt{NA}, & \text{if } y_{i1} = 0 \end{cases}, y_{i3} = \begin{cases} y_{i3}, & \text{if } y_{i1} = 1, y_{i2} = 0 \\ \texttt{NA}, & \text{otherwise} \end{cases},$$

$$y_{i4} = \begin{cases} y_{i4}, & \text{if } y_{i1} = 1, y_{i2} = 1 \\ \texttt{NA}, & \text{otherwise} \end{cases}, y_{i5} = \begin{cases} y_{i5}, & \text{if } y_{i1} = 0 \\ \texttt{NA}, & \text{otherwise} \end{cases}.$$

In addition, it is suggested to include an offset, denoted by ϕ, in the equations for these binary variables (Chipman et al. 2010). These binary offsets help to adjust the estimations of these binary values, as without an offset, the tree model prior shrinks the latent variable values to zero instead of the offset value. Specifically, the binary offsets are set to the means of their respective observed variables.

As mentioned before, this model on its own is not able to be estimated, so it will be translated into a SUR-BART structure. The resulting model is presented in Equation 16.

$$\begin{pmatrix} y_{i1}^* \\ y_{i2}^* \\ y_{i3}^* \\ y_{i4}^* \\ y_{i5}^* \end{pmatrix} = \begin{pmatrix} \sum_{h=1}^{m_1} g_1(\boldsymbol{x}_{i1}^{'}; T_h, M_h) + \phi_1 \\ \sum_{h=1}^{m_2} g_2(\boldsymbol{x}_{i2}^{'}; T_h, M_h) + \phi_2 \\ \sum_{h=1}^{m_3} g_3(\boldsymbol{x}_{i3}^{'}; T_h, M_h) \\ \sum_{h=1}^{m_4} g_4(\boldsymbol{x}_{i4}^{'}; T_h, M_h) \\ \sum_{h=1}^{m_5} g_5(\boldsymbol{x}_{i5}^{'}; T_h, M_h) \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} \tag{16}$$

The model specification still contains the exogenous covariates $\boldsymbol{x_i}$. To reiterate, the covariates of the first and second equation both contain an additional variable, the exclusion restriction, to help improve identification and estimation of the model. Additionally, the model now includes

the BART parameters, starting with the single binary tree T_h and the set of all terminal node values of a single tree M_h. Here $M_h = (\mu_{h1}, \mu_{h2}, ..., \mu_{hb_h})$ corresponds with the parameters for the b_h leaves of the hth tree. It is important to note that each equation has its personal unique tree structure, independent of the other equations. So the cardinality of each set of trees can differ between each equation, and is defined as m_j for $j = 1, ..., 5$. The $g_j(\boldsymbol{x}_{ij}^{'}, T_h, M_h)$ function assigns the value corresponding to the terminal node of tree h for equation j, for observation i.

Finally, the errors are assumed to follow a multivariate normal distribution, as is both the default in the original model and in the SUR-BART algorithm. Thus, $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4}, \varepsilon_{i5}) \sim \mathcal{N}_5(0, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is the covariance matrix of the corresponding system of equations. However, due to the nature of a model which incorporates sample selection and treatment effects, not all elements of the covariance matrix are identified. For instance, for each observation only one potential outcome is observed, so the covariances between these variables do not exist. Similarly, the dependent variable associated with the non-selected sample is not related to the treatment selection variable, so the corresponding covariance is also not identified. Additionally, the binary assumption placed on y_{i1} and y_{i2}, places a similar restriction on the corresponding elements of the covariance matrix. Specifically, under this assumption the variances of y_{i1} and y_{i2} are required to be equal to one. This results in the following covariance matrix, where unidentified elements are shown as a dot:

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & \Omega_{12} & \Omega_{13} & \Omega_{14} & \Omega_{15} \\ \Omega_{21} & 1 & \Omega_{23} & \Omega_{24} & \cdot \\ \Omega_{31} & \Omega_{32} & \Omega_{33} & \cdot & \cdot \\ \Omega_{41} & \Omega_{42} & \cdot & \Omega_{44} & \cdot \\ \Omega_{51} & \cdot & \cdot & \cdot & \Omega_{55} \end{pmatrix} \tag{17}$$

There are some problems associated with these unidentified elements, which will be discussed in the next sections, along with their respective solutions.

4.2 Prior specification

The specification of the model can be finalized by specifying priors on the different parameters of the model. These can be split into two parts: parameters for the sampler of the sum-of-trees parameters, and the covariance matrix. For the first part, recall that the sampler of the sum-of-trees parameters uses the same framework as the original BART algorithm. So, initially this research will follow the recommendation by Chipman et al. (2010), to use their default prior specifications as they have empirically proven their effectiveness.

The prior on the covariance matrix requires additional explanation, due to the unidentified elements in combination with the binary assumption of y_{i1} and y_{i2}. For the standard SSTE model, outlined in Equations 1 to 5, an Inverse Wishart prior is placed on fully-identified subsets of the covariance matrix. However, for the new SSTE-BART model with binary sample and treatment allocation variables, a different prior specification is required, as two covariance matrix elements are restricted to 1, and no prior is specified, or inference is made, on unidentifiable covariances. However, under the binary assumption a different prior specification is required. To achieve this, instead of placing a prior on the entire matrix, or on sub-matrices, this research places independent priors on all unique and identified elements of the covariance matrix. This ensures that all assumptions and restrictions can hold, and with enough observations and MCMC iterations, the prior should have little impact on the results.

For the unrestricted diagonal elements of the covariance matrix, which are the variances of y_{i3}, y_{i4} and y_{i5}, the same prior used in Chipman et al. (2010) is used. This results in an Inverse Chi-squared prior distribution, where the hyperparameters are set using a data-informed approach. Specifically, the residual standard deviation of an ordinary least squares regression of the corresponding dependent variable on the covariates, $\hat{\sigma}$, is used. The hyperparameter λ is then calculated such that, for given degrees of freedom v, the qth quantile of the prior on the variance is located at $\hat{\sigma}$. For this research, the parameters (v, q) are set at $(3, 0.99)$, following the aggressive approach by Chipman et al. (2010).

For the unrestricted, and identified, off-diagonal elements of the covariance matrix, diffuse priors are used. They are set to be centered around zero, and include some dependence on the size of the prior variances. Reason for this is that for larger possible values of the variance, the corresponding covariances are also likely to be larger. The diffuse distribution chosen here is a normal distribution with mean zero and variance set to 10, following the application in Chib and Greenberg (1998), to induce the diffusion.

Table 1: Overview of prior specifications of the parameters of the SSTE-BART model

Prior	**Prior specification**
$p(T_h)$	*(i)* $\alpha(1+d)^{-\beta}$, with $\alpha = 0.95, \beta = 2$ *(ii)* uniform prior on available variables *(iii)* uniform prior on available splitting values
$p(\mu_{ih}\|T_h)$	$\mu_{ih} \sim \mathcal{N}(0, \sigma_\mu^2)$, where $\sigma_\mu = 0.5/(k\sqrt{m})$, with $k = 2$
$p(\Omega)$	*(i)* $\Omega_{jj} \sim v\lambda/\mathcal{X}_v^2$ with $(v, q) = (3, 0.99)$, for $j = 3, 4, 5$ *(ii)* $\Omega_{ij} \sim \mathcal{N}(0, 10)$ for identified off-diagonal elements

One important aspect which must not be forgotten, is the fact that some restrictions are still necessary to these 'unrestricted' elements. Specifically, a covariance matrix must be positive definite, which also requires sub-matrices to be positive definite. As a result of this requirement, there are some restrictions on the possible values for the elements of the covariance matrix. This will be discussed in more detail in Section 4.4.3.

Table 1 contains an overview of the necessary regularization priors, along with the default values for the sum-of-trees sampler as mentioned in Section 3.3.1.

4.3 Soft Trees and Sparse Splitting Rules

For this research, two potential improvements to the SSTE-BART model will be tested as an additional feature. Specifically, the implementation of soft trees and sparse splitting rules, introduced by Linero and Yang (2018) as an improvement of the general BART framework. This was further adapted in a selection model context by O'Neill (2024), which provides the foundation for the implementation here.

Soft trees are obtained by replacing the 'hard' decision rules in the tree structure, $x_i \leq C$, with *soft decision rules*. These soft decision rules are obtained by incorporating the cumulative distribution function of a symmetric random variable x, $\psi(x)$. This results in the following form for soft decision rules: $\psi\left(\frac{x_j - C}{\tau}\right)$. A prediction from a soft tree can be described as a weighted linear combination of all parameter values in the terminal nodes. The weights are set as functions of the distances between covariates and splitting points. From this, a prediction from a single tree function can be rewritten as:

$$g(\boldsymbol{x}_i; T_h, M_h) = \sum_{\ell=1}^{L_h} \mu_{h,\ell} \phi_\ell(\boldsymbol{x}_i, T_h, \ell), \tag{18}$$

with

$$\phi_\ell(\boldsymbol{x}_i, T_h, \ell) = \prod_{b \in \mathcal{A}_\ell} \psi\left(\frac{x_{j_b} - C_b}{\tau_b}\right)^{\mathcal{I}\{x_{j_b} > C_b\}} \times \left\{1 - \psi\left(\frac{x_{j_b} - C_b}{\tau_b}\right)\right\}^{\mathcal{I}\{x_{j_b} \leq C_b\}}, \tag{19}$$

where L_h is the number of terminal nodes in the h^{th} tree, $\mu_{h,\ell}$ is the ℓ^{th} terminal node parameter of the h^{th} tree, and $\mathcal{A}_\ell$ is the set of parental nodes of terminal node ℓ. The parameter τ controls the sharpness of the decisions, with lower values of τ leading to sharper decisions. For the function $\psi(x)$ the logistic function $\psi(x) = (1 + \exp(-x))^{-1}$ is used in both Linero and Yang (2018).

Sparse splitting rules are implemented in the SSTE-BART model, as a method to achieve improved variable selection on datasets with many variables. These sparse splitting rules are obtained through the use of a Dirichlet prior on the splitting probabilities of the tree sampler. Specifically, the splitting probabilities for p variables become $(s_1, s_2, ..., s_p) \sim \mathcal{D}(\frac{a}{p}, \frac{a}{p}, ..., \frac{a}{p})$.

Here, the parameter a is responsible for the level of sparsity and is assumed to have Beta prior distribution, $Beta(0.5, 1)$. These splitting probabilities are then iteratively updated in each MCMC step, using the current tree structure to calculate new splitting probabilities. For this, the same steps as used in Linero and Yang (2018) are implemented as Linero shows that the Dirichlet prior is able to adapt to unknown levels of sparsity in the variables, and is able to improve predictions in high dimensional datasets.

4.4 MCMC sampling procedure

Similar as for the SUR-BART algorithm, constructing the posterior distribution is the next step towards inference of the SSTE-BART model. As mentioned before, the SSTE-BART model is a combination of the original SSTE model by Vossmeyer (2016) and the SUR-BART model by Chakraborty (2016). So the estimation procedure is a combination of different aspects of the two approaches, along with additional steps required to tackle specific problems that arise. These problems will be introduced briefly, after which the general steps of the estimation procedure will be provided in a concise and cohesive algorithm. After this, each unique step will get discussed in extensive detail to show exactly how the estimation algorithm works.

The first problem that arises, is found within the sampling steps for the sum-of-trees parameters. The SUR-BART sampler requires the full conditional covariance matrix for each draw of each tree parameter. However, since the covariance matrix is not fully identified, sampling becomes impossible to do in one step. To combat this, a sampling procedure similar to weighted BART (Sparapani et al. 2021) where the inputs are updated differently, is constructed. Specifically, this is done by deriving multiple full conditional distributions that contain the necessary identified elements for each sum-of-trees

The second and arguably more problematic issue arises for the sampling steps of the covariance matrix. This is the result of the combination between unidentified elements of the covariance matrix, and the restriction of some of the identified elements to be equal to one. Both of these problems are stand-alone topics of already existing literature and different solutions exist for them. In line with this, Vossmeyer (2016) constructed a sampler capable of handling these unidentified elements of the covariance matrix, based on earlier work into models with selection mechanisms (Chib 2007; Chib et al. 2009). It is further stated in Chib et al. (2009) that for models where covariances are equal to one, such algorithms do not work properly and instead refer to the use of a Metropolis-Hasting algorithm, like in Chib and Greenberg (1998). However, no concrete proof exists that such an algorithm can efficiently achieve convergence, in the presence of the aforementioned problems. Nevertheless, there is also no other

solution available to this problem, so a Metropolis-Hasting algorithm will serve as the starting point for an attempt at a solution. Specifically, a Random-Walk Metropolis-Hasting sampler as describe in Chib and Greenberg (1998) will be implemented, as it is claimed to be effective for low dimensional covariance matrices.

The general steps of one iteration of the resulting MCMC sampling procedure are summarized below in Algorithm 1.

Algorithm 1 General steps of one iteration of the MCMC estimation algorithm for SSTE-BART

1: Sample y_{i1}^* from the full conditional distributions of each possible sample and treatment selection scenario: $(y_{i1} = 0)$, $(y_{i1} = 1, y_{i2} = 0)$, and $(y_{i1} = 1, y_{i2} = 1)$

2: Sample y_{i2}^* from the full conditional distributions of each possible sample and treatment selection scenario: $(y_{i1} = 1, y_{i2} = 0)$, and $(y_{i1} = 1, y_{i2} = 1)$. (Note here, that $y_{i1} = 0$ is not observed for y_{i2} so it is omitted);

3: Define the 'full' residuals $R_{1i}, R_{2i}, R_{3i}, R_{4i}, R_{5i}$ for the sums of trees for each sample and treatment selection scenario, using the previously derived distributions;

4: Using these residuals, the sum-of-trees of each equation can be sampled, to obtain the full conditional samples of the sums of trees, alongside the subsequent tree parameters;

5: Sample $\mathbf{\Omega}$ using an adjusted Random-Walk Metropolis-Hasting algorithm.

From the general steps shown in Algorithm 1, a couple additional observations can be made. First of all, due to the unidentified covariances, it is necessary to split the conditional distributions into all possible sample and treatment selection scenarios. This ensures that for each conditional distribution, all related covariances are identified and subsequent draws can be obtained. Since these covariances are a crucial part of estimating the sums of trees, this is a necessary first step.

After this, note that y_{i3}^*, y_{i4}^* and y_{i5}^* are not sampled at all. This is done, because sampling these is only necessary if there is some form of censoring on the potential outcomes. For the SSTE-BART model it is assumed that the potential outcomes do not contain any form of censoring, and are observed if in the selected sample. As a result, the corresponding sampling step can be omitted here.

The third step is a helpful intermediate step, which makes the calculation of the residuals in the fourth step a lot easier. Using the derivation of the residuals in the third step, the sums of trees can now be sampled following similar steps as the SUR-BART algorithm.

4.4.1 Deriving full conditional distributions

The first steps of the sampling algorithm entails the derivation of multiple full conditional distributions. The derivations of each distribution will not be shown here, as these follow from standard results for conditional distributions of multivariate normally distributed variables (Holt and Nguyen 2023). Starting with the derived full conditional distributions for y_{i1}^*, under all possible sample and treatment selection scenarios, where conditioning on some irrelevant variables is removed. Note that g_j refers to the function $g_j(\boldsymbol{x}_{ij}^{'}; T_h, M_h)$:

- y_{i1}^* *for nonselected observations*, i.e. $y_{i1} = 0$:

$$y_{i1}^* | y_{i1} = 0, y_{i5}^*, g_1, g_5, \Omega \sim$$

$$\mathcal{TN}_{(-\infty,0)} \left(g_1(\boldsymbol{x}_{1i}) + \phi_1 + \Omega_{15}\Omega_{55}^{-1}(y_{i5}^* - g_5(\boldsymbol{x}_i)), (1 - \Omega_{15}^2\Omega_{11}^{-1}\Omega_{55}^{-1})\Omega_{11} \right)$$

- y_{i1}^* *for selected untreated observations*, i.e. $y_{i1} = 1$, $y_{i2} = 0$:

$$y_{i1}^* | y_{i1} = 1, y_{i2} = 0, y_{i2}^*, y_{i3}^*, g_1, g_2, g_3, \Omega \sim$$

$$\mathcal{TN}_{(0,\infty)} \left(g_1(\boldsymbol{x}_{1i}) + \phi_1 + \begin{bmatrix} \Omega_{12} & \Omega_{13} \end{bmatrix} \begin{bmatrix} \Omega_{22} & \Omega_{23} \\ \Omega_{32} & \Omega_{33} \end{bmatrix}^{-1} \begin{bmatrix} y_{i2}^* - g_2(\boldsymbol{x}_i) - \phi_2 \\ y_{i3}^* - g_3(\boldsymbol{x}_i) \end{bmatrix}, \Omega_{11} - \begin{bmatrix} \Omega_{12} & \Omega_{13} \end{bmatrix} \begin{bmatrix} \Omega_{22} & \Omega_{23} \\ \Omega_{32} & \Omega_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Omega_{21} \\ \Omega_{31} \end{bmatrix} \right)$$

- y_{i1}^* *for selected treated observations*, i.e. $y_{i1} = 1$, $y_{i2} = 1$:

$$y_{i1}^* | y_{i1} = 1, y_{i2} = 1, y_{i2}^*, y_{i4}^*, g_1, g_2, g_4, \Omega \sim$$

$$\mathcal{TN}_{(0,\infty)} \left(g_1(\boldsymbol{x}_{1i}) + \phi_1 + \begin{bmatrix} \Omega_{12} & \Omega_{14} \end{bmatrix} \begin{bmatrix} \Omega_{22} & \Omega_{24} \\ \Omega_{42} & \Omega_{44} \end{bmatrix}^{-1} \begin{bmatrix} y_{i2}^* - g_2(\boldsymbol{x}_i) - \phi_2 \\ y_{i4}^* - g_4(\boldsymbol{x}_i) \end{bmatrix}, \Omega_{11} - \begin{bmatrix} \Omega_{12} & \Omega_{14} \end{bmatrix} \begin{bmatrix} \Omega_{22} & \Omega_{24} \\ \Omega_{42} & \Omega_{44} \end{bmatrix}^{-1} \begin{bmatrix} \Omega_{21} \\ \Omega_{41} \end{bmatrix} \right)$$

Then the same derivations can be obtained for the full conditional distributions of y_{i2}^*. Here, there are only two distributions, as for y_{i2}^* observations for $y_{i1}^* = 0$ do not exist, because to be eligible for treatment, the observation first needs to be selected in the sample. Hence, the full conditional distributions for the remaining two scenarios are defined as follows:

- y_{i2}^* *for selected untreated observations*, i.e. $y_{i1} = 1$, $y_{i2} = 0$:

$$y_{i2}^* | y_{i1} = 1, y_{i2} = 0, y_{i1}*, y_{i3}*, g_1, g_2, g_3, \Omega \sim$$

$$\mathcal{TN}_{(-\infty,0)} \left(g_2(\boldsymbol{x}_{2i}) + \phi_2 + \begin{bmatrix} \Omega_{21} & \Omega_{23} \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{13} \\ \Omega_{31} & \Omega_{33} \end{bmatrix}^{-1} \begin{bmatrix} y_{i1}^* - g_1(\boldsymbol{x}_i) - \phi_1 \\ y_{i3}^* - g_3(\boldsymbol{x}_i) \end{bmatrix}, \Omega_{22} - \begin{bmatrix} \Omega_{21} & \Omega_{23} \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{13} \\ \Omega_{31} & \Omega_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Omega_{12} \\ \Omega_{32} \end{bmatrix} \right)$$

- y_{i2}^* *for selected treated observations*, i.e. $y_{i1} = 1$, $y_{i2} = 1$:

$$y_{i2}^* | y_{i1} = 1, y_{i2} = 1, y_{i1}*, y_{i4}^*, g_1, g_2, g_4, \Omega \sim$$

$$\mathcal{TN}_{(0,\infty)} \left(g_2(\boldsymbol{x}_{2i}) + \phi_2 + \begin{bmatrix} \Omega_{21} & \Omega_{24} \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{14} \\ \Omega_{41} & \Omega_{44} \end{bmatrix}^{-1} \begin{bmatrix} y_{i1}^* - g_1(\boldsymbol{x}_i) - \phi_1 \\ y_{i4}^* - g_4(\boldsymbol{x}_i) \end{bmatrix}, \Omega_{22} - \begin{bmatrix} \Omega_{21} & \Omega_{24} \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{14} \\ \Omega_{41} & \Omega_{44} \end{bmatrix}^{-1} \begin{bmatrix} \Omega_{12} \\ \Omega_{42} \end{bmatrix} \right)$$

The following distributions for y_{i3}, y_{i4} and y_{i5} are not directly necessary for the sampling procedure, but they another helpful intermediate step towards the calculations of the full residuals in the next step. These full conditional distributions are based on only one specific sample and treatment selection scenario, so only one distribution is required:

- *For selected untreated observations*, i.e. $y_{i1} = 1$, $y_{i2} = 0$:

$$y_{i3}|y_{i1}^*, y_{i2}^,g_1, g_2, \Omega \sim$$

$$\mathcal{N}\left(g_3(\boldsymbol{x}_{3i}) + \begin{bmatrix}\Omega_{31} & \Omega_{32}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}y_{i1}^* - g_1(\boldsymbol{x}_{1i}) - \phi_1 \\ y_{i2}^* - g_2(\boldsymbol{x}_{2i}) - \phi_2\end{bmatrix}, \Omega_{33} - \begin{bmatrix}\Omega_{31} & \Omega_{32}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}\Omega_{13} \\ \Omega_{23}\end{bmatrix}\right)$$

- *For selected treated observations*, i.e. $y_{i1} = 1$, $y_{i2} = 1$:

$$y_{i4}|y_{i1}^*, y_{i2}^,g_1, g_2, \Omega \sim$$

$$\mathcal{N}\left(g_4(\boldsymbol{x}_{4i}) + \begin{bmatrix}\Omega_{41} & \Omega_{42}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}y_{i1}^* - g_1(\boldsymbol{x}_{1i}) - \phi_1 \\ y_{i2}^* - g_2(\boldsymbol{x}_{2i}) - \phi_2\end{bmatrix}, \Omega_{44} - \begin{bmatrix}\Omega_{41} & \Omega_{42}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}\Omega_{14} \\ \Omega_{24}\end{bmatrix}\right)$$

- *For nonselected observations*, i.e. $y_{i1} = 0$:

$$y_{i5}^*|y_{i1} = 0, y_{i1}*, g_1, g_5, \Omega \sim$$

$$\mathcal{N}\left(g_5(\boldsymbol{x}_{5i}) + \Omega_{15}\Omega_{11}^{-1}(y_{i1}^* - g_1(\boldsymbol{x}_{1i}) - \phi_1), (1 - \Omega_{15}^2\Omega_{11}^{-1}\Omega_{55}^{-1})\Omega_{55}\right)$$

4.4.2 Sampling of the sum-of-trees parameters

The full conditional distributions as described above, can be used as the foundation to derive the conditional distributions of the residuals, which are a necessary building block for the derivation of the tree sampler (Chipman et al. 2010; Chakraborty 2016).

Let R_{ij} be the full residual of the g_j equation. Chipman et al. (2010) concluded that the tree parameters T_h and M_h can be drawn in a two-step procedure, using these residuals, which Chakraborty (2016) adjusted into:

1. Draw T_h from $T_h|R_{ij}, (T, M)_{(h)}, \boldsymbol{\Omega}$
2. Draw M_h from $M_h|T_h, R_{ij}, (T, M)_{(h)}, \boldsymbol{\Omega}$

For the SSTE-BART model, the same steps can be used, if the residuals of an equation are calculated conditionally on only the corresponding identified parts of that equation . The proposed draw of T_h is done based on one of four moves: growing a terminal node (GROW), pruning a pair of terminal nodes (PRUNE), changing a nonterminal rule (CHANGE), and swapping a rule between a parent and child node (SWAP) (Chipman et al. 2010). Each of these moves has its own probability, which are by default set at 0.25, 0.25, 0.40 and 0.10,

respectively. However, recent literature has suggested that the SWAP move has little impact on performance, but increases the computational burden of the algorithm significantly (Maia et al. 2024; Kapelner and Bleich 2013). So for the SSTE-BART model, the probability for SWAP will be set to 0, with the other probabilities scaled accordingly. Then, similar as for the derivation of the conditional distributions described previously, the residuals for a specific equation are defined separately for each different sample and treatment selection scenario. Such that only the latent variables that are identified for each equation are included in the formulas for the residuals. All unique specifications for the residuals and for the related variances of each function g_j are given below.

- For g_1, *use all the observations*:

  (a) If $y_{i1} = 0$, then set

  $$R_{1i} = y_{i1}^* - \phi_1 - \Omega_{15}\Omega_{55}^{-1}(y_{i5}^* - f_5(\boldsymbol{x}_i))$$

  and set the variance to $(1 - \Omega_{15}^2\Omega_{11}^{-1}\Omega_{55}^{-1})\Omega_{11}$.

  (b) If $y_{i1} = 1$ and $y_{i2} = 0$, then set

  $$R_{1i} = y_{i1}^* - \phi_1 - \begin{bmatrix}\Omega_{12} & \Omega_{13}\end{bmatrix}\begin{bmatrix}\Omega_{22} & \Omega_{23}\\ \Omega_{32} & \Omega_{33}\end{bmatrix}^{-1}\begin{bmatrix}y_{i2}^* - g_2(\boldsymbol{x}_i) - \phi_2\\ y_{i3}^* - g_3(\boldsymbol{x}_i)\end{bmatrix}$$

  and set the variance to

  $$\Omega_{11} - \begin{bmatrix}\Omega_{12} & \Omega_{13}\end{bmatrix}\begin{bmatrix}\Omega_{22} & \Omega_{23}\\ \Omega_{32} & \Omega_{33}\end{bmatrix}^{-1}\begin{bmatrix}\Omega_{21}\\ \Omega_{31}\end{bmatrix}$$

  .

  (c) If $y_{i1} = 1$ and $y_{i2} = 1$, set

  $$R_{1i} = y_{i1}^* - \phi_1 - \begin{bmatrix}\Omega_{12} & \Omega_{14}\end{bmatrix}\begin{bmatrix}\Omega_{22} & \Omega_{24}\\ \Omega_{42} & \Omega_{44}\end{bmatrix}^{-1}\begin{bmatrix}y_{i2}^* - g_2(\boldsymbol{x}_i) - \phi_2\\ y_{i4}^* - g_4(\boldsymbol{x}_i)\end{bmatrix}$$

  and set the variance to

  $$\Omega_{11} - \begin{bmatrix}\Omega_{12} & \Omega_{14}\end{bmatrix}\begin{bmatrix}\Omega_{22} & \Omega_{24}\\ \Omega_{42} & \Omega_{44}\end{bmatrix}^{-1}\begin{bmatrix}\Omega_{21}\\ \Omega_{41}\end{bmatrix}$$

- For g_2 , *do not make use of nonselected observations*:

  (a) If $y_{i1} = 1$ and $y_{i2} = 0$, then set

  $$R_{2i} = y_{i2}^* - \phi_2 - \begin{bmatrix}\Omega_{21} & \Omega_{23}\end{bmatrix}\begin{bmatrix}\Omega_{11} & \Omega_{13}\\ \Omega_{31} & \Omega_{33}\end{bmatrix}^{-1}\begin{bmatrix}y_{i1}^* - g_1(\boldsymbol{x}_i) - \phi_1\\ y_{i3}^* - g_3(\boldsymbol{x}_i)\end{bmatrix}$$

and set the variance to

$$\Omega_{22} - \begin{bmatrix}\Omega_{21} & \Omega_{23}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{13} \\ \Omega_{31} & \Omega_{33}\end{bmatrix}^{-1} \begin{bmatrix}\Omega_{12} \\ \Omega_{32}\end{bmatrix}$$

.

(b) If $y_{i1} = 1$ and $y_{i2} = 1$, then set

$$R_{2i} = y_{i2}^* - \phi_2 - \begin{bmatrix}\Omega_{21} & \Omega_{24}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{14} \\ \Omega_{41} & \Omega_{44}\end{bmatrix}^{-1} \begin{bmatrix}y_{i1}^* - g_1(\boldsymbol{x}_i) - \phi_1 \\ y_{i4}^* - g_4(\boldsymbol{x}_i)\end{bmatrix}$$

and set the variance to

$$\Omega_{22} - \begin{bmatrix}\Omega_{21} & \Omega_{24}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{14} \\ \Omega_{41} & \Omega_{44}\end{bmatrix}^{-1} \begin{bmatrix}\Omega_{12} \\ \Omega_{42}\end{bmatrix}$$

- For g_3, *only make use of selected untreated observations*, i.e. observations for which $y_{i1} = 1$ and $y_{i2} = 0$ and set

$$R_{3i} = y_{i3} - \begin{bmatrix}\Omega_{31} & \Omega_{32}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}y_{i1}^* - g_1(\boldsymbol{x}_{1i}) - \phi_1 \\ y_{i2}^* - g_2(\boldsymbol{x}_{2i}) - \phi_2\end{bmatrix}$$

and set the variance to

$$\Omega_{33} - \begin{bmatrix}\Omega_{31} & \Omega_{32}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}\Omega_{13} \\ \Omega_{23}\end{bmatrix}$$

- For g_4 *only make use of selected treated observations*, i.e. observations for which $y_{i1} = 1$ and $y_{i2} = 1$ and set

$$R_{4i} = y_{i4} - \begin{bmatrix}\Omega_{41} & \Omega_{42}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}y_{i1}^* - g_1(\boldsymbol{x}_{1i}) - \phi_1 \\ y_{i2}^* - g_2(\boldsymbol{x}_{2i}) - \phi_2\end{bmatrix}$$

and set the variance to

$$\Omega_{44} - \begin{bmatrix}\Omega_{41} & \Omega_{42}\end{bmatrix} \begin{bmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1} \begin{bmatrix}\Omega_{14} \\ \Omega_{24}\end{bmatrix}$$

- For g_5, *only use the nonselected observations*, i.e. observations for which $y_{i1} = 0$. Set

$$R_{5i} = y_{i5} - \Omega_{15}\Omega_{11}^{-1}(y_{i1}^* - \phi_1 - g_1(\boldsymbol{x}_{1i}))$$

and set the variance to $(1 - \Omega_{15}^2\Omega_{11}^{-1}\Omega_{55}^{-1})\Omega_{55}^2$.

With these residual definitions, it becomes possible to implement a BART tree sampler to obtain the sum-of-trees parameters for each equation. The implementation of a tree sampler is similar to the one in the SUR-BART algorithm, which simply comes down to running a singular BART function for each equation, conditioning on all other components previously mentioned. [1]

[1] The implementation of these BART functions is done using the `dbarts` package in `R` (Dorie et al. 2020).

4.4.3 Sampling the covariance matrix

For the sampling step of the covariance matrix Ω, a Random-Walk Metropolis-Hasting algorithm (Chib and Greenberg 1998) will be implemented, to tackle the problems associated with the covariance matrix of this model. Any Metropolis-Hasting algorithm is based on the following concept: a new sampled value is proposed based on the current sample, and is either accepted or rejected depending on a calculated acceptance probability. This can be summarized in the following general algorithm, with notation following Chib and Greenberg (1998), which will provide the foundation for the remainder of this sampler:

Algorithm 2 General steps of one iteration of a Metropolis-Hasting algorithm.

1: Sample proposal values $\Omega_{ij}^{'}$ given the current values Ω_{ij}, for all elements of $\mathbf{\Omega}$, from the densities $q(\mathbf{\Omega}^{'}|\mathbf{\Omega}, Z, \beta)$;

2: Accept the proposals $\mathbf{\Omega}^{'}$ with probability

$$\alpha(\mathbf{\Omega}, \mathbf{\Omega}^{'}) = \min\left\{\frac{\pi(\mathbf{\Omega}^{'})f(Z|\beta, \mathbf{\Omega}^{'})I(\mathbf{\Omega}^{'} \in C)}{\pi(\mathbf{\Omega})f(Z|\beta, \mathbf{\Omega})I(\mathbf{\Omega} \in C)} \frac{q(\mathbf{\Omega}|\mathbf{\Omega}^{'}, Z, \beta)}{q(\mathbf{\Omega}^{'}|\mathbf{\Omega}, Z, \beta)}, 1\right\},$$

and reject the proposals and stay with the current values with probability $1 - \alpha(\mathbf{\Omega}, \mathbf{\Omega}^{'})$

In Algorithm 2, $\mathbf{\Omega}$ and $\mathbf{\Omega}^{'}$ are the vectors containing the elements of the covariance matrix that are sampled. Here, $\pi(\mathbf{\Omega})$ is the *prior probability*, defined as the probability the value of $\mathbf{\Omega}$ has within the prior distribution, with similar interpretation for $\mathbf{\Omega}^{'}$. The function $f(Z|\beta, \mathbf{\Omega})$ is defined as the full data likelihood, evaluated at $\mathbf{\Omega}$, and similarly for $\mathbf{\Omega}^{'}$. The indicator function ensures that the proposed and current values fit in the region associated with the prior distribution, and is rejected directly otherwise. And lastly, the function $q(\mathbf{\Omega}^{'}|\mathbf{\Omega}, Z, \beta)$ is called the *proposal probability*, which is the probability that the proposed value is inside the proposal distribution conditional on the current value. The *reverse proposal probability*, $q(\mathbf{\Omega}|\mathbf{\Omega}^{'}, Z, \beta)$, can then be defined as the probability that the current value, is inside the proposal distribution, after substituting the proposed value inside the distribution instead of the current value.

This becomes the Random-Walk Metropolis-Hasting (RW-MH) algorithm through the choice of proposal density $q(\cdot)$. The general objective behind the choice of $q(\cdot)$ is to traverse the parameter space and be able to generate samples that mix well (Chib and Greenberg 1998). One of the simplest ways to achieve this is to generate proposals through a random-walk chain: $\mathbf{\Omega}^{'} = \mathbf{\Omega} + h$, where the proposed and current value are defined as before, and h is a zero-mean increment vector. Chib and Greenberg (1998) suggest that for suitable proposals, the variance of the chosen distribution for h can be set to a multiple of $1/n$. Here, n is the number of observations corresponding to the sample of the proposed element of the covariance matrix (e.g. for Ω_{31} this becomes the size of the untreated sample). So, using this proposal generation

approach in combination with Algorithm 2, this becomes the RW-MH algorithm.

However, a single step of such an RW-MH algorithm is not sufficient in the case of unidentified elements of the covariance matrix, because the likelihoods can not be calculated. To combat this, an aspect of the approach by Vossmeyer (2016) is borrowed. Specifically, Vossmeyer (2016) defines three subsets of the sample which have fully identified covariance matrices, corresponding to the three different sample and treatment selection scenarios. This results in the following three subsets, where the variances corresponding to the binary variables, Ω_{11} and Ω_{22}, are set equal to 1:

$$\boldsymbol{y}_{iC}^* = (y_{i1}^*, y_{i5}^*)', \boldsymbol{y}_{iD}^* = (y_{i1}^*, y_{i2}^*, y_{i3}^*)', \boldsymbol{y}_{iA}^* = (y_{i1}^*, y_{i2}^*, y_{i4}^*)' \tag{20}$$

$$\boldsymbol{\Omega}_C = \begin{pmatrix} 1 & \Omega_{15} \\ \Omega_{51} & \Omega_{55} \end{pmatrix}, \boldsymbol{\Omega}_D = \begin{pmatrix} 1 & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & 1 & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix}, \boldsymbol{\Omega}_A = \begin{pmatrix} 1 & \Omega_{12} & \Omega_{14} \\ \Omega_{21} & 1 & \Omega_{24} \\ \Omega_{41} & \Omega_{42} & \Omega_{44} \end{pmatrix}. \tag{21}$$

Using these subsets, the following complete-data likelihood can be constructed, where it is used that the errors are multivariate normally distributed. Then, define the likelihood of subset J as $f(\boldsymbol{y}_{iJ}^*|\boldsymbol{\theta})$, with $\boldsymbol{\theta}$ the subset of all parameters. Additionally, define $N_1 = \{i : y_{i1} = 0\}$ as the set of nonselected observations. Similarly, let $N_2 = \{i : y_{i1} = 1, y_{i2} = 0\}$ be the selected untreated observations, and $N_3 = \{i : y_{i1} = 1, y_{i2} = 1\}$ as the selected treated observations:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{y}^*|\boldsymbol{\theta}) &= \left[\prod_{i \in N_1} f(\boldsymbol{y}_{iC}^*|\boldsymbol{\theta})\right] \times \left[\prod_{i \in N_2} f(\boldsymbol{y}_{iD}^*|\boldsymbol{\theta})\right] \times \left[\prod_{i \in N_3} f(\boldsymbol{y}_{iA}^*|\boldsymbol{\theta})\right] \\ &\propto |\boldsymbol{\Omega}_C|^{-N_1/2} \exp\left\{-\frac{1}{2}\sum_{i \in N_1} \boldsymbol{\eta}_{iC}^{*\prime} \boldsymbol{\Omega}_C^{-1} \boldsymbol{\eta}_{iC}^*\right\} \\ &\quad \times |\boldsymbol{\Omega}_D|^{-N_2/2} \exp\left\{-\frac{1}{2}\sum_{i \in N_2} \boldsymbol{\eta}_{iD}^{*\prime} \boldsymbol{\Omega}_D^{-1} \boldsymbol{\eta}_{iD}^*\right\} \\ &\quad \times |\boldsymbol{\Omega}_A|^{-N_3/2} \exp\left\{-\frac{1}{2}\sum_{i \in N_3} \boldsymbol{\eta}_{iA}^{*\prime} \boldsymbol{\Omega}_A^{-1} \boldsymbol{\eta}_{iA}^*\right\} \end{aligned} \tag{22}$$

Equation 22 uses the following definition for $\boldsymbol{\eta}_{iJ}^*$, which is the vector of errors of the sum-of-trees models for the multiple elements in subset J: $\boldsymbol{\eta}_{iJ}^* = \boldsymbol{y}_{iJ}^* - \phi_j - \sum g_J(\boldsymbol{x}_{iJ}^{'}; T_h, M_h)$. This formulation of the complete-data likelihood makes it now possible to calculate the likelihood required in the accept-reject step in the RW-MH algorithm. Another advantage of this formulation is the ability to perform three distinct accept-reject steps of the RW-MH algorithm, based on these three subsets of the covariance matrix. This is advantageous, because a smaller set of proposed elements leads to an increased chance of a higher acceptance probability, which ultimately leads to better mixing properties of the algorithm. However, it should be noted that in general, it is preferred to sample as many parameters as possible in each block of such a Gibbs sampler.

These three distinct steps are summarized in Algorithm 3, which contains the general steps of one iteration of the covariance matrix sampler, where the formulas presented in Algorithm 2 can be inserted accordingly for each corresponding step.

Algorithm 3 General steps of one iteration of the adjusted Random-Walk Metropolis-Hasting algorithm for sampling the covariance matrix

1: Sample proposal values for $\mathbf{\Omega}_D^{'} = \{\Omega_{33}, \Omega_{32}, \Omega_{31}, \Omega_{21}\}$, given their current values $\mathbf{\Omega}_D$, from their proposal densities;
2: Accept the proposal values with probability $\alpha(\mathbf{\Omega}_D, \mathbf{\Omega}_D^{'})$
3: Sample proposal values for $\mathbf{\Omega}_A^{'} = \{\Omega_{44}, \Omega_{42}, \Omega_{41}\}$, given their current values $\mathbf{\Omega}_A$, from their proposal densities;
4: Accept the proposal values with probability $\alpha(\mathbf{\Omega}_A, \mathbf{\Omega}_A^{'})$
5: Sample proposal values for $\mathbf{\Omega}_C^{'} = \{\Omega_{55}, \Omega_{51}\}$, given their current values $\mathbf{\Omega}_C$, from their proposal densities;
6: Accept the proposal values with probability $\alpha(\mathbf{\Omega}_C, \mathbf{\Omega}_C^{'})$

It is important to note, that the value for Ω_{21} is only proposed once, even though it appears twice in the complete-data likelihood. Sampling a value twice brings additional complications along with it, so the choice is made to only propose the value once along with the other values for $\mathbf{\Omega}_D$. This means that, as a result of the sequential nature of the accept-reject steps, the value that is used in the calculations of the acceptance probability for $\mathbf{\Omega}_A$ is the updated value from the previous step. This becomes important in the upcoming discussion of positive definiteness.

The last crucial aspect of this covariance sampler, is concerned with the positive definiteness restriction of any covariance matrix. This means that each set of proposed values for a specific sub-matrix must result in a positive definite matrix. A sufficient condition for positive definiteness of a matrix, is that its determinant must be strictly greater than zero. With this in mind, a solution for adhering to this problem can be found in the choice of proposal distribution for the increment vector h. Specifically, if all but one element of a matrix are fixed, the positive definiteness condition can be achieved through solving the quadratic equation $\det(\mathbf{\Omega}_J) > 0$, for any sub-matrix J . The bounds that result from this quadratic equation can be set as the bounds for a proposal draw from a Truncated Normal distribution, ensuring positive definiteness of the resulting sub-matrix.

It would be theoretically sufficient to place these bounds on one element in each of the sub-matrices in Equation 21. However, this could result in disproportional restrictions on one element compared to the others, leading to poor mixing properties. So, to alleviate some of the restrictiveness on a single element, and to possibly increase the mixing properties of the

algorithm, these bounds are placed on all off-diagonal elements.

For all but one of these elements, these bounds are calculated based on the 2x2 sub-matrix corresponding to its covariances (e.g. Ω_{31} bounds positive definiteness of the sub-matrix of variables 3 and 1). For one element, these bounds are calculated to ensure positive definiteness of the entire sub-matrix as in Equation 21. This is Ω_{21} in Step 1 of Algorithm 3, and Ω_{41} for Step 3 of Algorithm 3. For Ω_{21}, an additional bound is necessary, because it appears twice in the complete-data likelihood of Equation 22. When proposing a value for Ω_{21} in Step 1 of Algorithm 3, a change occurs in $\mathbf{\Omega}_D$ as well as $\mathbf{\Omega}_A$. So for the calculations of the likelihood in the acceptance probability function, both matrices are required to be positive definite, leading to an extra bound. The bounds and corresponding calculations can be found in Appendix A.

The final part to complete the covariance sampler is then the choice of the distribution for the increment vector h. From the definition of the random-walk chain proposal, $\mathbf{\Omega}' = \mathbf{\Omega} + h$, it follows that the distribution of the proposal $\mathbf{\Omega}'$ is the distribution of h, centered around the current value $\mathbf{\Omega}$. This fact will be used for providing the proposal distributions of the elements of the covariance matrix. For the unrestricted diagonal elements, which are the variances of y_{i3}, y_{i4} and y_{i5}, the distribution must ensure that the variances are strictly positive. To this end, a Log-Normal distribution is chosen, centered around the current value. Following the suggestion of Chib and Greenberg (1998), the variances of the distributions of y_{i3}, y_{i4} and y_{i5}, will be set to a multiple of $\frac{1}{N_2}, \frac{1}{N_3}, \frac{1}{N_1}$, respectively. These multiplication factors will be set to increase the possible speed of convergence during testing of the algorithm.

As mentioned before, for all off-diagonal elements the distribution will be a Truncated-Normal distribution, centered around the current value. In addition, the previously mentioned bounds to ensure positive definiteness will be placed on each element. Similar as for the variances, the variances of these proposal distributions will be set to the reciprocal of the sample size of the corresponding subset.

5 Simulation Study

In this section an extensive simulation study is conducted to assess the performance of the SSTE-BART model and the corresponding MCMC sampling algorithm. Data is simulated using two data generating processes (DGPs) to evaluate the performance of the model in different settings. In addition, a comparison will be made against several other already established models, as well as some variations of the SSTE-BART model. First, an overview of the different settings of the DGPs will be discussed, along with other specifications for the simulation study.

5.1 Simulation settings

The first simulation setting is based on one of the settings also used in Chakraborty (2016) The setting is an adaption of the Friedman's five-dimensional test function (Friedman et al. 1983; Friedman 1991), which is one of the most common benchmarks for simulation studies in existing literature. The setting in Chakraborty (2016) only contains four equations, so in addition to adopting those, an additional fifth equation is constructed to fit the SSTE-BART model. Furthermore, similar as in Vossmeyer (2016), the covariates in the treatment selection equation are assumed to have one more covariate than the other equations. This is the variable x_5, which is regarded as the instrumental variable correlated with the treatment but not with the errors. Additionally, the first equation also contains an additional variable, x_6, to simulate the exclusion restriction which can improve identification and estimation of the model (Heckman 1976). The resulting functions are given below:

$$
\begin{aligned}
y_1 &= 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_6 + \epsilon_1 \\
y_2 &= 5\sin(\pi x_1 x_2) + 10(x_3 - 0.5)^4 + 8x_4 + 3x_5 + \epsilon_2 \\
y_3 &= 20\sin(\pi x_1 x_2) + 15(x_3 - 0.5)^2 + 30x_4^2 + \epsilon_3 \\
y_4 &= 15\sin(\pi x_1^2 x_2) + 10(x_3 - 0.5)^2 + 5x_4 + \epsilon_4 \\
y_5 &= 10\sin(\pi \sqrt{x_1} x_2) + 20(x_3 - 0.5)^2 + 12x_4 + \epsilon_5,
\end{aligned} \tag{23}
$$

where the covariates $x_1, x_2, x_3, x_4, x_5, x_6$ are independently and identically distributed on a standard Uniform distribution, iid $U(0,1)$. The errors $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5)$ are assumed to have a multivariate Normal distribution, $\mathcal{N}_5(\mathbf{0}, \boldsymbol{V})$. The covariance matrix $\boldsymbol{V}$ is tested for two different specifications, where one contains higher values than the other. The first one is taken from Chakraborty (2016) and results in $v_{ij} = 0.8^{|i-j|}, i, j = 1, ..., 5$ for the identified elements of $\boldsymbol{V}$, which automatically follows the binary variance restriction. The second one is copied from Vossmeyer (2016), where the diagonal elements are set equal to 0.25, and the identified off-diagonal elements all to 0.1. For this setting, the variances for the first two variables are manually set to 1, to comply with the binary variance restriction.

The second simulation setting is chosen as a counterpart to the highly nonlinear structure of the first setting. Specifically, a very simple DGP where the covariates are different for each function, will be implemented to further assess the ability of the model to estimate the underlying functions. Similar as for the first setting, the first and second equation contain an extra covariate to help identification and estimation of the model. The functions for the simple

DGP that follow are displayed below:

$$
\begin{aligned}
y_1 &= 10x_1 + 5x_3 + 15x_4 + \epsilon_1 \\
y_2 &= 12x_1 + 5x_2 + 8x_5 + \epsilon_2 \\
y_3 &= 8x_1 + 17x_2 + \epsilon_3 \\
y_4 &= 15x_2 + 10x_3 + \epsilon_4 \\
y_5 &= 10x_1 + 5x_3 + \epsilon_5,
\end{aligned} \tag{24}
$$

where the distributions for the covariates and errors are assumed to be the same as for the first simulation setting. Furthermore, the two specifications for the covariance matrix $\boldsymbol{V}$ are similarly tested for the simple DGP.

In addition, for both simulation settings five 'noise' variables are added to the covariate space. These variables, noted as $x_7, x_8, x_9, x_{10}, x_{11}$, are also assumed to follow a standard Uniform distribution. Since these variables to not enter the underlying functions of the DGPs, they have no effect on the response variables. This enables this research to check if the model correctly excludes these variables from the estimated model, or if the model is unable to handle noise appropriately.

For both simulation settings, the distribution of observations into the sub-samples associated with the potential outcomes is assumed to be even. This is done to fairly evaluate the performance of the model, whereas a skewed distribution might produce biased results in favor of a larger subset.

5.2 Results of simulation study

To fit the SSTE-BART models, some initial values need to be specified, namely for the covariance matrix $\boldsymbol{\Omega}$, for the sum-of-tree functions $g_j(\boldsymbol{x}_{ij}, T_h, M_h)$ and for the latent y^*_{i1} and y^*_{i2}. First, for $\boldsymbol{\Omega}$ the initial values are set according to:

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & \cdot \\
0 & 0 & \sigma^2_{3,ols} & \cdot & \cdot \\
0 & 0 & \cdot & \sigma^2_{4,ols} & \cdot \\
0 & \cdot & \cdot & \cdot & \sigma^2_{5,ols}
\end{pmatrix},
$$

where $\sigma_{j,ols}$ is the standard deviation of the same linear regression as in the prior specification of the corresponding elements. These values should provide a good starting point for the MCMC algorithm, to try and estimate any possible covariance matrix. It can be stated that all trees

should start out as stumps, and because the data is centered, all functions $g_j(\boldsymbol{x}_{ij}, T_h, M_h)$ can be initialized to the value 0. The latent outcomes are initialized as a random draw from a normal distribution, truncated to ranges equal to those set in the full conditional draws in Section 4.4.1. The mean is set to the corresponding binary offsets, with standard deviation equal to 1.

For each of the models, it is important to choose a sufficient number of MCMC iterations to enable the algorithm to achieve convergence. In general, more iterations result in a higher chance of convergence, with the downside of increased computational times. Initially, 11,000 MCMC iterations with 1,000 iterations as the burn-in sample provided reasonable convergence, while also providing moderate computational times. However, doubling this to 22,000 MCMC iterations with 2,000 burn-in iterations did improve the predictive results by a margin. The computational time of these models also doubled, making the runtime for some of these models, especially in the model comparison, undesirable. In addition, the predictive results for the model comparison did not improve significantly with the doubling of MCMC iterations. Therefore, the decision is made to run the regular SSTE-BART models with 22,000 MCMC iterations, while performing the model comparisons based on 11,000 MCMC iterations.

In addition, each of the above mentioned DGPs will be simulated for 4,000 observations, where the first half is used as the training set, and the second half for testing. In a simulation study, enough observations are necessary in the training sample to ensure the model is trained properly. Simultaneously, more observations in the testing set lead to a more accurate performance review of the trained model. However, too many observations can lead to deteriorating mixing properties of the MCMC sampler (Hill et al. 2020), so it becomes a trade-off between accuracy, but also computational burden. As a result, 4,000 observations with a 50/50 split provides a sufficient sample size for both sets, while still having a moderate runtime.

For each of the regular BART-based models, SSTE-BART included, the number of trees is kept fixed at $m = 100$. Initial testing revealed that increasing the number of trees to $m = 200$, as suggested by Chipman et al. (2010), did not significantly improve results while doubling the runtime. Contrarily, a lower number of trees did negatively impact results in a significant manner, so for the remaining analyses, the number of trees is kept fixed at $m = 100$. For the SoftBART-based models, the default recommendation is $m = 25$, but initial testing found $m = 50$ to provide significantly better results. Even though the runtime of SoftBART-based model doubles, the results were convincing, so here the number of trees is fixed at $m = 50$.

Lastly, the multiplication factors for the variances of the proposal distributions are also set after some testing. The factors for covariance elements containing a 2, 3, 4, and 5 are respectively, 16, 8, 6, and 6. These values are identical for both the simple DGP and the more

complex Friedman specification, as they provided appropriate convergence rates for all elements.

To evaluate the performance of the models and the different simulation settings, different evaluation metrics are used. For predictive performance, the Mean Squared Error (MSE) is used as the measure for the continuous response variables, y_3, y_4, y_5. However, for the remaining binary response variables, y_1, y_2, it is not advisable to use the standard MSE as a measure. Instead the Brier Score is most commonly used to assess predictive performance of a binary variable. The Brier Score (BS) is calculated using the formula: $BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$, where f_i is the probability of the predicted value for observation i, and o_i is the observed outcome for observation i. This looks similar to the formula for the standard MSE, but is inherently different with the use of probabilities, making it valid for binary response variables.

Furthermore, three commonly used convergence diagnostics are also calculated to evaluate the convergence properties of the model. The first two, Geweke's and Heidelberger & Welch's convergence diagnostic can be calculated directly from a single MCMC chain, only requiring one run of the model. The third one, the Gelman-Rubin convergence diagnostic, traditionally requires multiple MCMC chains with overdispersed starting points. However, Vats and Knudson (2021) recently developed an improved version of the Gelman-Rubin diagnostic, called Stable Gelman-Rubin, which requires only one MCMC chain, while increases stability. So, for computational efficiency, the Stable Gelman-Rubin convergence diagnostic will be used, which is available in the `StableGR` package in `R`.

The results for the performance measures are displayed in Table 2 for each different specification of the simulation setting. Additionally, the posterior means of the elements of the covariance matrix are also provided for each specification in Table 3.

Table 2: Performance measures of the SSTE-BART model for different specifications of the DGP, covariance matrix V of the DGP, and the choice for implementation of the Dirichlet prior.

Model: DGP-V-Dirichlet	**BS y1**	**BS y2**	**MSE y3**	**MSE y4**	**MSE y5**
1: Friedman-Chakraborty-No	0.0608	0.1611	1.5473	1.4427	1.2463
2: Friedman-Chakraborty-Yes	0.0554	0.1602	1.3273	1.2960	1.2554
3: Simple -Chakraborty-No	0.0524	0.1211	1.1468	1.1458	1.0478
4: Simple -Chakraborty-Yes	0.0453	0.1134	1.1352	1.1259	1.0082
5: Friedman-Vossmeyer -No	0.0583	0.1563	0.8476	0.6042	0.4438
6: Friedman-Vossmeyer -Yes	0.0519	0.1557	0.6503	0.4763	0.3957
7: Simple -Vossmeyer -No	0.0456	0.1125	0.3271	0.3024	0.2846
8: Simple -Vossmeyer -Yes	0.0409	0.1058	0.3086	0.2866	0.2819

From Table 2 a couple of observations can be made. The Brier scores for y_1 are all between 0.04 and 0.06, meaning that the SSTE-BART model is very accurate in predicting if an observation will be selected into the sample. A similar observation can be made for the Brier score of y_2, where the values range from 0.10 to 0.16, indicating slightly worse, but still quite good, predictive capabilities. The MSE values of specifications where $\boldsymbol{V}$ is set according to Chakraborty (2016) are significantly higher than for the models based on Vossmeyer (2016). This is to be expected, as the covariance matrix $\boldsymbol{V}$ of the DGP influences the variability in the response variables, where a higher variability in the model can lead to model predictions being more inaccurate, as the relations between the variables can become more obscure. Furthermore, predictions based on the simple DGP are more accurate in all settings, which is in line with the expectation that a simpler model is easier to estimate than a more difficult one. Additionally, the use of a Dirichlet prior on the splitting probabilities of the tree sampler lead to more accurate predictions in all settings for the SSTE-BART model, further proving the results of Linero and Yang (2018).

Table 3: Posterior means of the covariance matrix of the SSTE-BART model. The top four models are based on the Chakraborty covariance specification, and the bottom four on the Vossmeyer specification.

Model	$\boldsymbol{\Omega_{11}}$	$\boldsymbol{\Omega_{21}}$	$\boldsymbol{\Omega_{22}}$	$\boldsymbol{\Omega_{31}}$	$\boldsymbol{\Omega_{32}}$	$\boldsymbol{\Omega_{33}}$	$\boldsymbol{\Omega_{41}}$	$\boldsymbol{\Omega_{42}}$	$\boldsymbol{\Omega_{44}}$	$\boldsymbol{\Omega_{51}}$	$\boldsymbol{\Omega_{55}}$
Actual	*1*	*0.80*	*1*	*0.64*	*0.80*	*1*	*0.51*	*0.64*	*1*	*0.41*	*1*
1	1	0.02	1	0.09	0.19	0.76	0.05	0.39	0.76	-0.03	0.83
2	1	0.02	1	0.29	0.27	0.80	-0.06	0.32	0.90	0.10	0.91
3	1	0.00	1	0.15	0.42	0.83	0.11	0.25	0.77	0.18	0.91
4	1	0.32	1	0.36	0.63	0.93	0.38	0.55	0.88	0.26	1.07
Actual	*1*	*0.10*	*1*	*0.10*	*0.10*	*0.25*	*0.10*	*0.10*	*0.25*	*0.10*	*0.25*
5	1	-0.04	1	-0.03	-0.08	0.30	0.07	-0.13	0.23	-0.04	0.21
6	1	-0.19	1	-0.01	-0.17	0.28	0.08	-0.25	0.26	0.05	0.22
7	1	0.04	1	0.01	0.11	0.21	-0.02	0.02	0.20	-0.08	0.24
8	1	0.16	1	0.08	0.05	0.24	0.06	0.10	0.24	0.09	0.26

The posterior means of the covariance matrix elements in Table 3 provide additional insights in the performance of SSTE-BART. The table is divided in two sections, each corresponding to one of the specifications for the covariance matrix of the DGP, $\boldsymbol{V}$. For both specifications of $\boldsymbol{V}$ it can be seen that the models with the simple DGP, models 3, 4, 7 and 8, are more successful in approximating the true covariances than the models with Friedman DGPs. Furthermore, the variance elements, $\Omega_{33}, \Omega_{44}, \Omega_{55}$, get estimated accurately more frequent than the remaining covariances. The implementation of a Dirichlet prior improves the estimation accuracy for a

majority of the elements in models with the Chakraborty covariance matrix specification, but has no significantly noticeable effect on the other models. The exception being Model 8, where the estimates are significantly more accurate than for its counterpart Model 7, which does not have the Dirichlet prior.

Convergence of the algorithm is checked using the previously mentioned diagnostics, which are performed on the nine sampled elements of the covariance matrix. For all models, Heidelberger's diagnostic returned that stationarity was almost never achieved for any of the nine elements of the covariance matrix, which seems unreasonable. The Geweke diagnostic is more optimistic, claiming a number of stationary elements between 6 and 8, indicating reasonable convergence. Stationarity is almost always achieved for the unrestricted diagonal elements, while it differs quite randomly for the remaining elements. The interpretation of the Gelman-Rubin statistic requires a cut-off point for establishing stationarity, similar to a significance level in statistical testing. In existing literature this value is often chosen arbitrarily to achieve high convergence outputs, and the authors of the original Gelman-Rubin diagnostic recommend in a later paper that a cut-off value of 1.1 is sufficient (Gelman et al. 1995). However, this has been heavily criticized by the developers of the `StableGR` package (Vats and Knudson 2021), stating that this leads to much too lenient convergence claims. Instead they suggest a formula to calculate an appropriate cut-off point, which should be sufficient in most scenarios. The value resulting from this formula will simply be provided here, for the specific formula the reader is referred to Section 5 of Vats and Knudson (2021). In the SSTE-BART model, the appropriate cut-off value is 1.0029, which is much lower and stricter than 1.1. The difference between the cut-off values is drastic, as for a cut-off of 1.0029, only a small portion of the elements is stationary across all models. On the contrary, a cut-off of 1.1, or even 1.05, indicates that all elements are stationary across all models.

To further investigate convergence, as these diagnostics are not able to provide conclusive evidence, the MCMC chains of each of the elements are plotted, to find out if there are some visual indications of convergence. Figure 2 contains the plots for the MCMC chains of Model 2. From Figure 2 it looks like the variances are fluctuating around a value which approximates the true values, indicating convergence. The remaining covariances are different from this, as some similarly fluctuate around a value which is not close to the true value, but what would still indicate some form of convergence has been achieved. However, it would be presumptuous to assume convergence for these elements, as they are mostly non-stationary near the end of the chain, and also in general throughout the entire chain. Plots of other models exhibit similar patterns, so they are not shown in this paper as similar observations can be made for them.

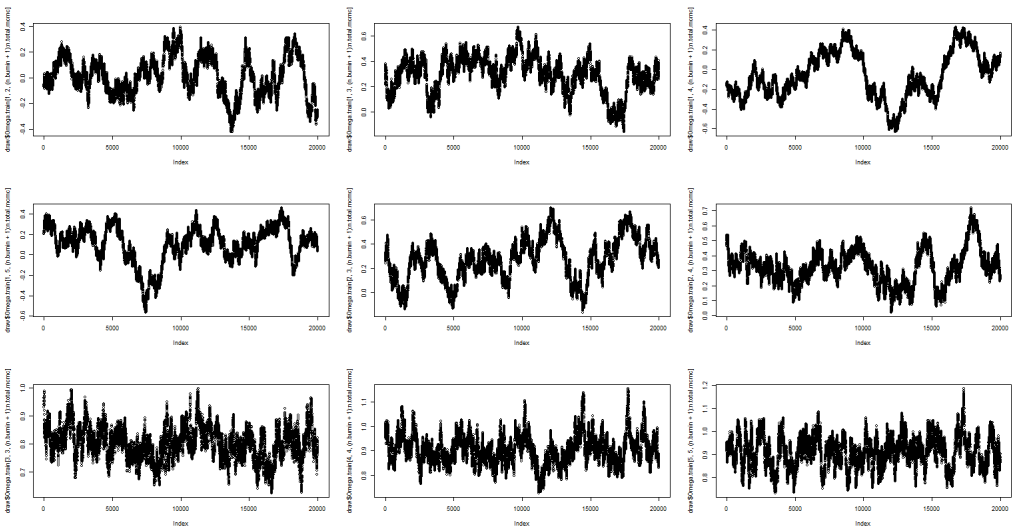

Figure 2: Plots of 20,000 MCMC chains (after burn-in) for the sampled elements of the covariance matrix for Model 2. From left to right, top to bottom: $\Omega_{21}, \Omega_{31}, \Omega_{41}; \Omega_{51}, \Omega_{32}, \Omega_{42}; \Omega_{33}, \Omega_{44}, \Omega_{55}$

The biggest difference between models is mostly not the rate of convergence, but rather the point where the chain convergence to. So it seems like different specifications for the model do not lead to differing convergence rates, but do affect the accuracy of the point of convergence, which is reflected in the posterior means.

5.2.1 Model Comparison

The performance of the SSTE-BART model is compared against several other related methods, which are primarily other BART-based methods. First of all, as mentioned in Section 4.3, soft trees are implemented in the model, resulting in a model which will be called SSTE-SoftBART. This implementation is achieved using the `SoftBart` package in `R`, developed by Linero and Yang (2018). The second model for comparison is also a version of the SSTE-BART algorithm, where the covariance sampler as developed in Vossmeyer (2016) is implemented. This sampler is similar to the one in McCulloch and Rossi (1994), which was criticized by Chib and Greenberg (1998) whose sampler is adopted in the SSTE-BART model. It should illustrate the difference in performance of those samplers, and the resulting model will be referred to as VoSSTE-BART.

Additionally, the standard versions of the regular BART and SoftBART algorithms will be tested as comparisons. Since each of these algorithms can only handle a univariate model, five separate models will be estimated for each of the five equations in the SSTE-BART model. Therefore, each model only contains the observations which are observed for the sample associated with that equation. For the regular BART algorithm, this results in two probit BART models and three regular BART models. These models are easily fitted using the same `dbarts`

package, and most settings are kept at their default values. The only adjusted settings are 100 trees, 11,000 MCMC iterations and a removed 'SWAP' probability.

A similar model construction is done for the SoftBART models, where two probit SoftBART models and three regular SoftBART models are estimated, using the same `SoftBART` package. Again, the parameter settings of each model are kept at their default values, with the exception of 50 trees and 11,000 MCMC iterations.

The final comparison model is the TOBART-2 model, which is not officially released but provided by dr. O'Neill[2]. Essentially, TOBART-2 is a sample selection model based on Type-2 Tobit, where the nonselected observations are censored at zero, and there is only one single outcome equation. In context of the SSTE-BART model, this means no y_{i2} & y_{i5}, and y_{i3} & y_{i4} are combined into a single outcome variable, along with y_{i1} as the selection variable. The parameter settings for the TOBART-2 model are kept at their default, where the number of trees and MCMC iterations are again set at 100 trees and 11,000 iterations.

For the model comparison, a benchmark specification belonging to one of the tested SSTE-BART models must be chosen. While the MSE values in Table 2 differ significantly for the choice of $\boldsymbol{V}$, it is difficult to conclude if one of the choices can be preferred. So for the comparisons, the specification with Chakraborty covariances will be used, as it provides the clearest differences in results for both the predictive performance measures as the posterior means. Additionally, the implementation of a Dirichlet prior on the splitting probabilities has a positive impact on all models, so it will be implemented for the SSTE-SoftBART and VoSSTE-BART models. Here, it is important to note that the `SoftBART` package automatically implements the desired Dirichlet Prior, such that no additional code is required. It is also of interest for these comparison models to be tested on different levels of complexity of an underlying DGP, as results can greatly differ accordingly. Hence, each comparison model will be tested for both the simple and Friedman DGPs. Therefore, the two benchmark models for SSTE-BART are Model 2 (Friedman DGP) and 4 (simple DGP) from Table 2

The results for predictive performance of all comparison models, alongside the benchmark for SSTE-BART, are presented in Table 4 and 5. For the three variations of the SSTE-BART model, the same five performance measures, consisting of the Brier scores and MSE values, can be calculated. As a result of the five separate models for each of the equations in the runs of regular BART and SoftBART, these performance measures can also be calculated for these models. However, for the TOBART-2 model, it is only possible to calculate the Brier score for the sample selection equation, and one MSE value for the single outcome equation. But, by

[2] https://github.com/EoghanONeill/TobitBART

changing the treatment status in the test data, it becomes possible to obtain predictions for both y_{i3} and y_{i4}.

Table 4: Predictive performance results of all comparison models for the Friedman DGP specification.

Model	BS y1	BS y2	MSE y3	MSE y4	MSE y5
SSTE-BART	0.0553	0.1598	1.3917	1.2791	1.2483
SSTE-SoftBART	0.0468	0.1526	0.9662	1.0588	1.0075
VoSSTE-BART	0.0499	0.1590	1.4253	1.2911	1.2448
BART	0.0560	0.1174	1.4537	1.3719	1.2245
SoftBART	0.0436	0.1094	0.9355	1.0540	1.0190
TOBART-2	0.0574		1.7704	1.3789	

Table 5: Predictive performance results of all comparison models for the simple DGP specification.

Model	BS y1	BS y2	MSE y3	MSE y4	MSE y5
SSTE-BART	0.0458	0.1142	1.1387	1.1107	1.0105
SSTE-SoftBART	0.0405	0.1071	1.0697	1.0082	0.9570
VoSSTE-BART	0.0397	0.1118	1.0968	1.1275	1.0059
BART	0.0396	0.0566	1.1342	1.1413	1.0407
SoftBART	0.0361	0.0523	1.0061	0.9996	0.9540
TOBART-2	0.0329		1.0752	1.1366	

From Table 4 it can be seen that the implementation of soft trees (SSTE-SoftBART) results in significantly more accurate predictions compared to the original SSTE-BART model. On average this is a 22.4% decrease in MSE values for the Friedman DGP and an average decrease of 6.9% for the MSE values of the simple DGP. A similar observation can be made from Table 5, where SSTE-SoftBART also outperforms the original SSTE-BART model. The VoSSTE-BART model obtains comparable predictive results as SSTE-BART, showcasing the predictive capabilities of the algorithm even with a theoretically invalid covariance matrix sampler. Furthermore, the difference between predictive accuracy between the simple DGP and complex DGP is greatly reduced in favor of the SSTE-SoftBART model. This indicates that the SSTE-SoftBART model can handle a complex underlying DGP as good, if not better, than a simple DGP. The performance of the regular BART and SoftBART models are mostly comparable to their SSTE counterparts, and even provide a better Brier score for the treatment selection equation y_{i2}. This difference in Brier scores is especially apparent for the simple DGP, where the Brier scores are halved. Revealing that the standard models of BART and SoftBART can

achieve similar performance as more detailed models built around them. This could be an indication that the SSTE-BART and SSTE-SoftBART models are not better than the standard versions at solely predicting the response variables. However, the standard models are not able to capture combined effects of variables and are accompanied by other restrictions.

For the three models based on the SSTE-BART algorithm posterior means of the covariance matrix can be obtained and are displayed in Table 6.

Table 6: Posterior means of the covariance matrix of the comparison models. The top four models are based on the Friedman DGP specification, and the bottom four on the simple DGP specification.

Model	Ω_{11}	Ω_{21}	Ω_{22}	Ω_{31}	Ω_{32}	Ω_{33}	Ω_{41}	Ω_{42}	Ω_{44}	Ω_{51}	Ω_{55}
Actual	*1*	*0.80*	*1*	*0.64*	*0.80*	*1*	*0.51*	*0.64*	*1*	*0.41*	*1*
SSTE-BART	1	-0.11	1	0.28	-0.02	0.78	-0.22	0.41	0.92	0.14	0.93
SSTE-SoftBART	1	0.30	1	0.56	0.60	1.02	0.11	0.45	0.98	0.28	1.02
VoSSTE-BART	1	0.26	1	0.64	0.00	0.79	0.21	0.00	0.91	0.11	0.93
SSTE-BART	1	0.35	1	0.39	0.64	0.94	0.44	0.56	0.89	0.28	1.09
SSTE-SoftBART	1	0.39	1	0.39	0.81	1.03	0.35	0.56	0.95	0.43	1.08
VoSSTE-BART	1	38281	1	1.06	2.10	0.93	1.09	3.41	0.89	3.66	1.07

In line with the predictive performances, for the complex Friedman DGP the SSTE-SoftBART models are able to approximate the true values of the covariance matrix significantly better than the other two models. However, for the simple DGP differences between SSTE-BART and SSTE-SoftBART are smaller, but still a slight edge to latter. The model is still slightly struggling to accurately approximate the off-diagonal elements of the covariance matrix for the complex DGP, but obtains much better results than the regular SSTE-BART model.
From Table 6 it can also be seen, that the VoSSTE-BART model is not able to accurately estimate these off-diagonal elements at all. This is to be expected, as the sampler for the covariance matrix itself is not valid. For the simple DGP, the SSTE-SoftBART estimations are most accurate, but still not near a range of significantly correct estimation.

5.2.2 Treatment Effects

The last part of the results of the simulation study are concerned with treatment effects, which is a big aspect of this research. There exist many different approaches to calculate a variety of treatment effects, so first the choice and definitions of the treatment effects calculated in this research will be given. After this, the corresponding results for these calculations will be provided and discussed briefly.

In a general sample selection and treatment effect model, the effect of a treatment can be found by comparing the treated sample versus the untreated sample. A commonly used measure for this is the Average Treatment Effect (ATE), which is defined as the difference between expected value of the response variable for the treated sample versus the expected value of the response variable for the untreated sample. Formally, this can be written down as: $ATE = E[Y|X = 1] - E[Y|X = 0]$. In the context of this research, this translates into: $ATE = E[y_4] - E[y_3]$. No conditions are necessary to check if an observation belongs to the corresponding sample, as this is already correct as a result of the model structure. This measure can easily be calculated for all models included in this simulation study, and the results are presented in Table 7.

Table 7: Estimated Average Treatment Effects for all comparison models for both DGP specifications.

Model	**ATE Friedman**	**ATE Simple**
Actual	*-8.43*	*3.49*
SSTE-BART	-8.60	3.29
SSTE-SoftBART	-8.74	3.19
VoSSTE-BART	-8.48	3.45
BART	-8.58	3.39
SoftBART	-8.47	3.43
TOBART-2	-8.45	3.39

From Table 7 it can be seen that the ATE is similar across all models, and that they are able to all approximate the average treatment effect of the test set accurately. Surprisingly, the SSTE-BART and SSTE-SoftBART models estimate the ATE worse than the comparison models for both the complex Friedman DGP as the simple DGP.

However, the strength of SSTE-BART lies in the fact that the use of Bayesian Additive Regression Trees allows for Heterogeneous Treatment Effects (HTE). Through BART, each observation can exhibit a unique effect on the given treatment, which in turn can be extracted from the model. Recall that for each observation, only one state for treatment is observed, while the unobserved state is called the counterfactual. However, for simulated data it becomes possible to estimate this value, because the underlying DGP is known to the researcher. As a result, the treatment effect for a single observation i, defined as HTE_i can be calculated, following a similar formula as the ATE: $HTE_i = E[y_{i4}] - E[y_{i3}]$.

As an example, a histogram of the distribution of Heterogeneous Treatment Effects for Model 2 of SSTE-BART is displayed in Figure 3.

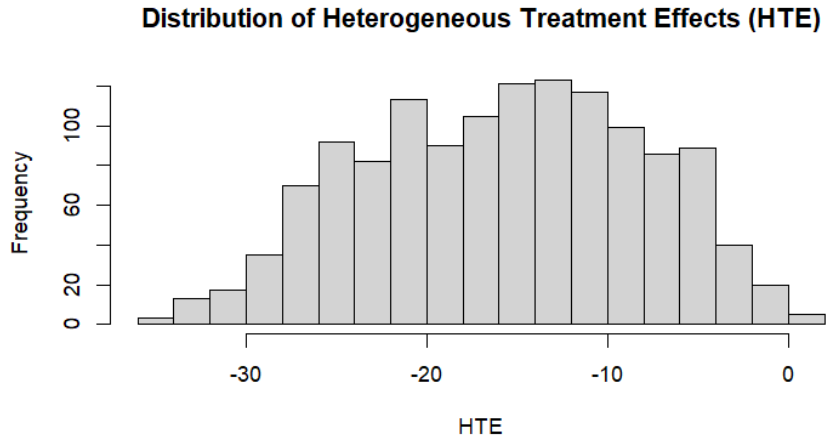

Figure 3: Distribution of Heterogeneous Treatment Effects for Model 2 of SSTE-BART

From Figure 3 it can be confirmed that the SSTE-BART model is indeed capable of estimating heterogeneous treatment effects. Furthermore, following from the setting of the simulation study, the distribution of these treatment effects should follow a Normal-like distribution. The distribution in Figure 3 has a shape looking like a Normal distribution, with the exception of some non-Normal spikes. Other model specifications for the SSTE-BART model show a similar shape, but are excluded from this paper to not compromise its conciseness.

6 Application

This section will discuss the application of the SSTE-BART model on a real-life dataset that fits the sample selection and treatment effect model. The dataset is the well-known National Supported Work (NSW) Demonstration dataset [3], obtained from the experiment performed by LaLonde (1986). The NSW Demonstration was a voluntary labor training program initiated to help individuals struggling with basic job skills, operated by the Manpower Demonstration Research Corporation (MDRC). These individuals could apply to the program, and if chosen were given a temporary job position in a supportive environment. Here they were provided with a decent salary and received frequent counseling. As a result, these individuals got a unique opportunity to develop skills they were lacking in a monitored environment.

6.1 Data preparation

In addition to observations for these treated and untreated individuals, LaLonde (1986) also included a set of control observations obtained from the Panel Study of Income Dynamics

[3]The dataset can be easily loaded from the `rrp` package in `R`.

(PSID). The combination of NSW and PSID observations results in an almost perfectly fit for the structure of an SSTE model. However, one requirement is violated, which is the requirement of non-randomized selection into treatment. The nature of the NSW Demonstration program explicitly stated that it would have been unethical if the treatment selection was not randomized (MDRC 1980). Nevertheless, the dataset should still be able to provide insights in the performance of the model for a real-life dataset, so analysis is continued.

The combined dataset contains observations for $n = 3212$ individuals, where for each individual demographics such as age and received education are available. As mentioned before, the observations can be divided into three categories: applied for the NSW program but not selected (untreated), applied for the program and selected (treated), and not applied for the program at all (nonselected). The sizes for the untreated, treated and nonselected samples are $N_2 = 425$, $N_3 = 297$, and $N_1 = 2490$, respectively.

The variable of interest is denoted as the income of the individual during a specific year. There are three years available, 1974, 1975 and 1978, where the first two years are before the NSW Demonstration program, and 1978 corresponds to the post-treatment year.

To analyze the effect of treatment, the response variable for the analyses is constructed as the difference between income in 1975 (pre-treatment) and 1978 (post-treatment). Additional cleaning of the dataset is done, as observations for which the income in either 1975 or 1978 is missing are removed to improve accuracy. This is the reason why the year 1974 is not used here, as it contained many missing observations, which would lead to a significant portion of the data being removed. The updated sample sizes for the three potential outcomes are $N_2 = 150$, $N_3 = 182$, and $N_1 = 2122$, respectively. The difference in size between the selected and nonselected samples is quite high, as ideally the observations would be equally distributed. Recall that the nonselected observations all come from the PSID, such that a random subset of these observations should still provide a representative subset of the nonselected sample. Therefore, a number of observations similar to the other two categories, namely $N_1 = 171$, is randomly sampled from the PSID set of observations. Furthermore, because the order of magnitudes differ greatly between the variables, both the response variable and the covariates are standardized.

6.2 Results for the NSW dataset

The dataset is run on the same set of comparison models as for the simulation study, to further analyze the performance of the SSTE-BART model. Similarly, all models are run for a total of 11,000 MCMC iterations, with 1,000 observations treated as the burn-in sample. Since the

dataset is not very large, the training set must be of sufficient size in order for the model to estimate an accurate model, so a train-test split of 80/20 is chosen. The number of trees is also kept exactly the same as for the simulation study, with $m = 100$ for the BART-based models, and $m = 50$ for the SoftBART-based models. Finally, after initial runs of the SSTE-BART model, the multiplication factors for the variances of the proposal distributions are fixed at 10, 3, 4, and 6. These correspond to elements of the covariance matrix containing a 2, 3, 4, and 5 respectively.

Additionally, the fit of the SSTE-BART models is completed by specifying initial values for the covariance matrix $\boldsymbol{\Omega}$, for the sum-of-tree functions $g_j(\boldsymbol{x}_{ij}^{'}, T_h, M_h)$ and for the latent variables y_{i1}^* and y_{i2}^*. For this application, the exact same initial values as for the simulation study are chosen, as reasonable results were obtained there. Moreover, those initial values were chosen to fit a broad number of different data specifications, so it should be an appropriate choice. The predictive performance results of each model are presented in a similar form as for the simulation study, and are displayed in Table 8.

Table 8: Predictive performance results of all comparison models for the standardized NSW dataset.

Model	**BS y1**	**BS y2**	**MSE y3**	**MSE y4**	**MSE y5**
SSTE-BART	0.1221	0.2457	0.4686	0.4870	0.4193
SSTE-SoftBART	0.1084	0.2502	0.3765	0.2604	0.1676
VoSSTE-BART	0.1224	0.2531	0.3578	0.3885	0.5014
BART	0.0814	0.2558	0.4754	0.3439	0.3550
SoftBART	0.0798	0.2516	0.3851	0.2652	0.1373
TOBART-2	0.0541		0.4525	0.2957	

From Table 8, similar observations can be made as for the simulated data in the previous section. It can be seen that SSTE-SoftBART outperforms the regular SSTE-BART model by some margin, with MSE values for y_{i3}, y_{i4}, y_{i5} decreased by 19.7%, 46.5% and 60.0%, respectively. Again, the standard models of BART and SoftBART achieve similar predictive performance results as their SSTE counterparts, with standard BART outperforming SSTE-BART just slightly. Both standard BART as standard SoftBART obtain a lower Brier score for y_{i1}, similar as for y_{i2} in the simulation study. The TOBART-2 model achieves an even better Brier score for the prediction of sample selection, while also predicting the outcomes fairly comparable to the rest. Next, the estimated posterior means of the elements of the covariance matrix $\boldsymbol{V}$ are presented in Table 9.

These estimates of the covariances can aid in understanding relations between different equations

Table 9: Posterior means of the covariance matrix of the comparison models for the standardized NSW datasaet.

Model	Ω_{11}	Ω_{21}	Ω_{22}	Ω_{31}	Ω_{32}	Ω_{33}	Ω_{41}	Ω_{42}	Ω_{44}	Ω_{51}	Ω_{55}
SSTE-BART	1	0.39	1	-0.39	-0.15	1.00	0.26	0.43	1.31	0.20	0.75
SSTE-SoftBART	1	0.05	1	-0.33	-0.25	1.11	-0.05	-0.00	1.03	0.14	1.00
VoSSTE-BART	1	-0.10	1	-0.02	0.01	0.93	0.00	0.00	0.89	0.00	0.81

in the model. For instance, the negative values for the estimate of Ω_{31} indicate a negative relation between the third and first equation. Which can be translated as, the effect of applying for the NSW Demonstration program, leads to a lower income after being rejected for the program. This conclusion is logical, in the sense that an individual who is rejected after application is more likely to compromise on salary as this individual does not have the basic job skills necessary to improve their position. In addition, estimates of Ω_{32} and Ω_{42} can reveal selection of treatment on unobservables, if these values are nonzero. For example, the estimates of Ω_{32} for SSTE-BART and SSTE-SoftBART are both nonzero, indicating that there could indeed be selection of treatment on unobservables. In existing literature surrounding the NSW Demonstration dataset, there has been quite some attention towards assessing the sensitivity to unconfoundedness, which is the term generally used for selection of treatment on observables (Imbens 2003; Masten et al. 2024). This indicates that there most likely is selection of treatment on unobservables in the NSW dataset, which is also shown in the estimates for Ω_{32} and Ω_{42}. However, as the elements differ significantly between each model and predictive results showed a disability to correctly estimate the underlying model, concrete conclusions based on these values must be taken with care. The Average Treatment Effect (ATE) is also calculated for each of these models, and the results are presented in Table 10.

Table 10: Estimated Average Treatment Effects for all comparison models on the NSW dataset.

Model	ATE
SSTE-BART	-4049.19
SSTE-SoftBART	1389.30
VoSSTE-BART	-352.84
BART	856.31
SoftBART	542.09
TOBART-2	345.88

Table 10 displays an interesting pattern, where there is a wide range of estimated ATEs across

all models. This could be the result of the relatively small size of the final dataset, or because the models are simply not able to estimate the underlying model of the data correctly. Existing research into the dataset has often revealed an ATE ranging between $\$1,500 - \$1,800$ (LaLonde 1986; Dehejia and Wahba 1999). The SSTE-SoftBART model is the only one successful in approximating this range, but the wide range of ATE estimations still suggests a poor ability to estimate the underlying model from the NSW dataset.

Heterogeneous Treatment Effects (HTEs) are difficult to calculate for a real-life dataset, as the counterfactuals are never observed. There exist different machine learning algorithms capable of calculating HTEs (Künzel et al. 2019; Syrgkanis et al. 2019), but these models are often not easily accessible and require in-depth knowledge of the machine learning techniques used. Therefore, as this research has already demonstrated its capabilities to estimate heterogeneous treatment effects, the calculations for the NSW Demonstration dataset are omitted.

7 Discussion

This section provides a thorough discussion of the obtained results from the simulation study in Section 5 and the real-life data application in Section 6. Additionally, some potentially limiting aspects of the SSTE-BART model will be discussed.

Starting with the simulation study, recall that the SSTE-BART model is tested on two different specification of the underlying DGP of the data. A simple DGP is constructed, as well as a more complex variant based on the Friedman function. In addition, the implementation of sparse splitting rules through the use of a Dirichlet prior on the splitting probabilities is tested. Furthermore, two specifications of the covariance matrix for the DGP, $\boldsymbol{V}$, are tested, based on the settings used in Vossmeyer (2016) and Chakraborty (2016).

First of all, there is a clear difference in the MSE values between the different specifications of $\boldsymbol{V}$. For the Vossmeyer specification, which consists of smaller covariances, the MSE values are much lower, indicating that the predictions are more accurate. This is not surprising, as the response variables contain little variability, making it easier for predictions to approximate the true values. Furthermore, the complexity of the DGP also strongly impacts predictive performance. For all combinations of specifications, models trained on the simple DGP obtain better accuracy in predicting the response variable than the complex DGP counterparts. This could mean that the standard SSTE-BART model is not able to estimate the underlying functions of the model accurately if it is complex. The incorporation of a Dirichlet prior on the splitting probabilities of the regression trees proves to be an improvement for all model specifications, which is in line with the conclusions in Linero and Yang (2018). Additionally, only a fixed

number of trees was considered for each model, while it has been shown in existing literature that each component in a model can require a different number of trees (Chakraborty 2016). Hence, it could be of interest to research different settings for the number of trees, to try and improve performance of the model.

Combining these findings, the model which provides the best performance results for both specifications of the covariance matrix, is the model with a simple DGP and Dirichlet prior on the splitting probabilities. The posterior means for the elements of the covariance matrix provide additional insights in the performance of the model. It is shown that the best predicting models are able to accurately estimate the elements corresponding to the variances of the variables y_3, y_4 and y_5. For the other model specification, these variances are sporadically estimated near their true values, but are more often inaccurate. The same can be said for all estimates of the covariances, which the SSTE-BART model has trouble estimating accurately. The best predicting models again perform the best compared to the other models, but still fail to consistently approximate the true values. This is most likely a result of the implementation of the Random-Walk Metropolis-Hasting algorithm constructed for the sampler of covariances. The performance of such a sampler in the scenario with unidentified elements along with binary response variables, was also questioned by Chib et al. (2009). In addition, this research places multiple bounds on the proposal values for these elements, to ensure positive definiteness of the covariance matrix. It is possible that these restrictions impact the ability of the model to accurately estimate the covariances.

The comparison of versions of the SSTE-BART model with other competing models has not confirmed the claim of ignoring sample selection leading to less accurate results Vossmeyer (2016). The runs for the standard models of BART and SoftBART are able to equally accurate predict the response variables, and in most cases are even more accurate in estimating the selection variables. This indicates that estimating the SSTE structure simultaneously in the SSTE-BART model, is not able to outperform separate models for each aspect of the SSTE model. This is surprising, as allowing for combined effects of variables in a model, generally leads to more accurate estimation of the underlying model.

Furthermore, the implementation of soft trees in the SSTE-BART model, referred to as the SSTE-SoftBART model, significantly outperforms the other SSTE-BART models. Concretely, the MSE values for the simple DGP simulated data are on average approximately 6.9% lower for SSTE-SoftBART than for regular SSTE-BART. While for the complex DGP, this percentages moves to a 22.4% decrease, further confirming the capabilities of soft trees as stated in Linero and Yang (2018). Moreover, the predictive performance of SSTE-SoftBART for the complex

DGP is equal to that for the simple DGP. This indicates that the SSTE-SoftBART model is able to handle data with a complex underlying DGP as good as a simple DGP. A notable observation that can be made, is that the predictive performance of the SSTE-BART model with the wrong covariance sampler, VoSSTE-BART, performs equally to the standard SSTE-BART model. This shows that a theoretically invalid estimation procedure, can still obtain results similar to a theoretically valid one.

It should be noted however, that the obtained results for predictive performance are not nearly as good as those obtained in the papers this research was based upon (Vossmeyer 2016; Chakraborty 2016). This could be the result of this research requiring the construction of a novel approach to the estimation procedure for the parameters of the model. Multiple choices are made which differ from ones made for the models in the original papers, like the choice of prior distributions, or which variables to sample in the Gibbs sampler.

The results for the posterior means of the covariance matrix can only be calculated for the SSTE-BART models, and not for the other comparison models. Still, a similar pattern can be found, where SSTE-SoftBART is estimating the covariances more accurately then its counterparts. However, for the simple DGP the estimations are significantly more accurate than for the complex DGP. The approximation of the true values of the covariances are also closer for SSTE-SoftBART, with the posterior means of the simple DGP being extremely close. This indicates that the implementation of soft trees to the SSTE-BART model not only increases predictive performance, but is consequently also able to more accurately estimate the covariance matrix. This raises the question if the RW-MH sampler might not be the only cause of bad estimates, and that the fit of the sum-of-trees models actually plays a significant role. Moreover, the estimates of the VoSSTE-BART models are somewhat good for the complex DGP, while they are some orders of magnitude off for the simple DGP. This is another indication that a theoretically invalid sampler for the covariance matrix produces untrustworthy results. Additionally, the values for Ω_{32} and Ω_{42} are set to a nonzero value while simulating the data. As a result, the data is modeled to have selection of treatment on unobservables, which lead to biased results from BART-enhanced algorithms as these models can only account for treatment selection on observables. This can often be improved by including estimates of propensity scores as splitting variables, or through the use of methods like IV-BART (Spanbauer and Pan 2022).

The calculation of treatment effects showcased the ability of the SSTE-BART models to correctly estimate the Average Treatment Effect (ATE). However, the other comparison models achieved equally accurate estimations, and in some cases even more accurate. Specifically, the standard runs of the BART and SoftBART models approximated the true ATE of the test set

more accurately than their SSTE counterparts. This further demonstrates the capabilities of these algorithms in itself. The difference with the more complex SSTE-BART models could be explained by the fact that the effects of treatment are easily separable, not requiring a very in-depth model to extract them.

Furthermore, Heterogeneous Treatment Effects (HTEs) are confirmed to be captured in the SSTE-BART models, and display an expected normal distribution. This is a convincing argument that the implementation of BART to the original SSTE model has indeed resulted in the ability to estimate HTEs.

The application of the SSTE-BART model on the NSW dataset provided a set of mixed results. First of all, the predictive performance results display that SSTE-SoftBART also predicts more accurately compared to the regular SSTE-BART model. With MSE values lower on average by approximately 37%. Additionally, all comparison models are able to equally accurate predict the response variable, with MSE values comparable to the SSTE-BART models. This adds to the argument that the ability of the SSTE-BART to estimate combined variable effects, does not result in improved predictive performance. Similar as in the simulation study, the VoSSTE-BART model predicts equal, if not better, than the standard SSTE-BART model.

It is important to note that these predictive results are obtained from standardized variables, such that it is difficult to judge the relative performance of these models. This becomes also apparent for the posterior means of the elements of the covariance matrix. This standardization is done for better interpretation of the results, as the non-standardized results were of an extremely high order of magnitude, making inference very difficult. This also shows that the used dataset is perhaps not entirely correct to be used in this setting. Recall that one of the requirements of the dataset was already violated, where the treatment was randomized instead of being non-random. Moreover, the final dataset only consists of around 500 observations, which is quite small to extract meaningful and significant results. Especially after dividing the dataset into their respective potential outcomes, resulting in an average of 167 observations in each sample. One other option to evaluate the performance of the model on such a small dataset more accurately, would be to use k-fold cross-validation on the train-test split of the data.

These less than ideal properties of the dataset could potentially be the cause for the extreme fluctuations in the calculated ATE across the models, as well as the variability in the posterior means of the covariances. Hence, additional research is required to further assess the performance of the SSTE-BART model on real-life data.

8 Conclusion

This research extends the existing SSTE model (Vossmeyer 2016) to capture heterogeneous treatment effects, by replacing the linear structure of the model with BART terms. In addition, the model is adjusted further to fit more general applications, by assuming the selection variables to be binary instead of continuous. Following from this, this research tries to answer the following research question: *"Can heterogeneous treatment effects be captured in the sample selection and treatment effects model, using Bayesian Additive Regression Trees?"*

Moreover, the resulting SSTE-BART model is further tested for improvements by implementing soft trees and imposing sparse splitting rules. To be able to estimate these models, a novel MCMC estimation procedure, based on the RW-MH algorithm, is developed, capable of handling identification issues in combination with binary selection variables.

An extensive simulation study revealed that the variant of the SSTE-BART model containing the implementation of both soft trees, SSTE-SoftBART, resulted in significantly better performance than the standard SSTE-BART model. Consequently, the SSTE-SoftBART model was able to achieve comparable results to the other competing models, while also providing more accurate estimations of the covariance matrix. However, standard versions of BART and SoftBART provided equally good results, indicating the SSTE-BART models can still be improved by some margin. Furthermore, it was shown that Heterogeneous Treatment Effects are indeed correctly captured in the SSTE-BART models, achieving the main goal of this research.

The application of the SSTE-BART models on a real-life dataset further confirmed the superior performance of the SSTE-SoftBART model. However, no additionally useful insights were able to be extracted from the dataset, which still raises questions on the performance of SSTE-BART on real-life data.

To conclude, this research has succeeded in correctly estimating heterogeneous treatment effects in a sample selection and treatment effects model, through the use of Bayesian Additive Regression Trees. Additionally, a successful attempt at constructing a novel estimation procedure for a covariance matrix with unidentified elements and binary variables has been provided. The SSTE-SoftBART model has proven itself to be a good performing model, but it can still be improved as the results are not yet improved compared to Vossmeyer (2016).

8.1 Future Research

Even though an estimation procedure for a restricted covariance matrix was successfully constructed, it has much room for improvement. Approaches similar to Chan and Jeliazkov (2009) and Zhang et al. (2015) have the potential to provide more accurate samplers, if correctly ad-

justed to fit the restrictions. Furthermore, as the SSTE-BART model does not outperform the standard versions of BART and SoftBART, future research into generally improving the SSTE-BART algorithm is desired. In addition, this research only tested one specification for the prior and proposal distributions. More research could be done to see if the performance is influenced by a different choice for these distributions.

This research specifically used Bayesian Additive Regression Trees to try and capture heterogeneous treatment effects, but there exist other methods capable of this. Examples consist of other tree-based methods like Causal Random Forest (Hahn et al. 2020), or other regression-based methods like Debiased/Double Machine learning (Chernozhukov et al. 2018).

Moreover, combining the SSTE-BART model with other Machine Learning algorithms to be able to also estimate heterogeneous treatment effects for real-life data, is another interesting step towards improving the model (Künzel et al. 2019; Syrgkanis et al. 2019).

For the application on a real-life dataset, there were very few datasets available online, so any future research into finding fitting datasets to tests the performance of SSTE models is very relevant.

References

[1] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. "Identification of causal effects using instrumental variables". In: *Journal of the American statistical Association* 91.434 (1996), pp. 444–455.

[2] James L Beck and Lambros S Katafygiotis. "Updating models and their uncertainties. I: Bayesian statistical framework". In: *Journal of Engineering Mechanics* 124.4 (1998), pp. 455–461.

[3] Leo Breiman and Jerome H Friedman. "Estimating optimal transformations for multiple regression and correlation". In: *Journal of the American statistical Association* 80.391 (1985), pp. 580–598.

[4] S Chakraborty. "Bayesian additive regression tree for seemingly unrelated regression with automatic tree selection". In: *Handbook of Statistics*. Vol. 35. Elsevier, 2016, pp. 229–251.

[5] Joshua Chi-Chun Chan and Ivan Jeliazkov. "MCMC estimation of restricted covariance matrices". In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 457–480.

[6] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[7] Siddhartha Chib. "Analysis of treatment response data without the joint distribution of potential outcomes". In: *Journal of Econometrics* 140.2 (2007), pp. 401–412.

[8] Siddhartha Chib and Edward Greenberg. "Analysis of multivariate probit models". In: *Biometrika* 85.2 (1998), pp. 347–361.

[9] Siddhartha Chib, Edward Greenberg, and Ivan Jeliazkov. "Estimation of semiparametric models in the presence of endogeneity and sample selection". In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 321–348.

[10] Hugh A Chipman, Edward I George, and Robert E McCulloch. "BART: Bayesian additive regression trees". In: (2010).

[11] Hugh A Chipman, Edward I George, and Robert E McCulloch. "Bayesian CART model search". In: *Journal of the American Statistical Association* 93.443 (1998), pp. 935–948.

[12] Rajeev H Dehejia and Sadek Wahba. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs". In: *Journal of the American statistical Association* 94.448 (1999), pp. 1053–1062.

[13] Peng Ding. "Bayesian robust inference of sample selection using selection-t models". In: *Journal of Multivariate Analysis* 124 (2014), pp. 451–464.

[14] Vincent Dorie, Hugh Chipman, Robert McCulloch, and A Dadgar. "dbarts: discrete Bayesian additive regression trees sampler". In: *R package version 0.9-19* 3 (2020), pp. 30–43.

[15] Jerome H Friedman. "Multivariate adaptive regression splines". In: *The annals of statistics* 19.1 (1991), pp. 1–67.

[16] Jerome H Friedman, Eric Grosse, and Werner Stuetzle. "Multidimensional additive spline approximation". In: *SIAM Journal on Scientific and Statistical Computing* 4.2 (1983), pp. 291–301.

[17] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[18] Donald P Green and Holger L Kern. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees". In: *Public opinion quarterly* 76.3 (2012), pp. 491–511.

[19] Reuben Gronau. "Wage comparisons–A selectivity bias". In: *Journal of political Economy* 82.6 (1974), pp. 1119–1143.

[20] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)". In: *Bayesian Analysis* 15.3 (2020), pp. 965–1056.

[21] Trevor Hastie and Robert Tibshirani. "Bayesian backfitting (with comments and a rejoinder by the authors". In: *Statistical Science* 15.3 (2000), pp. 196–223.

[22] James J Heckman. "Econometric causality". In: *International statistical review* 76.1 (2008), pp. 1–27.

[23] James J Heckman. "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models". In: *Annals of economic and social measurement, volume 5, number 4*. NBER, 1976, pp. 475–492.

[24] James J Heckman and Bo E Honore. "The empirical content of the Roy model". In: *Econometrica: Journal of the Econometric Society* (1990), pp. 1121–1149.

[25] James J Heckman, Hedibert F Lopes, and Rémi Piatek. "Treatment effects: A Bayesian perspective". In: *Econometric reviews* 33.1-4 (2014), pp. 36–67.

[26] James J Heckman and Edward J Vytlacil. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". In: *Handbook of econometrics* 6 (2007), pp. 4779–4874.

[27] Jennifer Hill. "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240.

[28] Jennifer Hill, Antonio Linero, and Jared Murray. "Bayesian additive regression trees: A review and look forward". In: *Annual Review of Statistics and Its Application* 7 (2020), pp. 251–278.

[29] William Holt and Duy Nguyen. "Essential Aspects to Bayesian Data Imputation". In: *Available at SSRN 4494311* (2023).

[30] Martin Huber. "Treatment evaluation in the presence of sample selection". In: *Econometric Reviews* 33.8 (2014), pp. 869–905.

[31] Guido W Imbens. "Sensitivity to exogeneity assumptions in program evaluation". In: *American Economic Review* 93.2 (2003), pp. 126–132.

[32] Adam Kapelner and Justin Bleich. "Bartmachine: Machine learning with bayesian additive regression trees". In: *arXiv preprint arXiv:1312.2171* (2013).

[33] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10 (2019), pp. 4156–4165.

[34] Robert J LaLonde. "Evaluating the econometric evaluations of training programs with experimental data". In: *The American economic review* (1986), pp. 604–620.

[35] Myoung-jae Lee. "Treatment effects in sample selection models and their nonparametric estimation". In: *Journal of Econometrics* 167.2 (2012), pp. 317–329.

[36] Siu Fai Leung and Shihti Yu. "On the choice between sample selection and two-part models". In: *Journal of econometrics* 72.1-2 (1996), pp. 197–229.

[37] Fan Li, Peng Ding, and Fabrizia Mealli. "Bayesian causal inference: a critical review". In: *Philosophical Transactions of the Royal Society A* 381.2247 (2023), p. 20220153.

[38] Mingliang Li and Justin L Tobias. "Bayesian analysis of treatment effect models". In: *Bayesian inference in the social sciences* (2014), pp. 63–90.

[39] Phillip Li. "Estimation of sample selection models with two selection mechanisms". In: *Computational statistics & data analysis* 55.2 (2011), pp. 1099–1108.

[40] Antonio R Linero and Yun Yang. "Bayesian regression tree ensembles that adapt to smoothness and sparsity". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.5 (2018), pp. 1087–1110.

[41] Mateus Maia, Keefe Murphy, and Andrew C Parnell. "GP-BART: A novel Bayesian additive regression trees approach using Gaussian processes". In: *Computational Statistics & Data Analysis* 190 (2024), p. 107858.

[42] Willard G Manning, Naihua Duan, and William H Rogers. "Monte Carlo evidence on the choice between sample selection and two-part models". In: *Journal of econometrics* 35.1 (1987), pp. 59–82.

[43] Matthew A Masten, Alexandre Poirier, and Linqi Zhang. "Assessing sensitivity to unconfoundedness: Estimation and inference". In: *Journal of Business & Economic Statistics* 42.1 (2024), pp. 1–13.

[44] Robert McCulloch and Peter E Rossi. "An exact likelihood analysis of the multinomial probit model". In: *Journal of Econometrics* 64.1-2 (1994), pp. 207–240.

[45] Robert McCulloch, Rodney Sparapani, Brent Logan, and Purushottam Laud. "Causal inference with the instrumental variable approach and Bayesian nonparametric machine learning". In: *arXiv preprint arXiv:2102.01199* (2021).

[46] MDRC. *Summary and findings of the national supported work demonstration*. Ballinger Publishing Company, 1980.

[47] Jerzy Neyman. "On the application of probability theory to agricultural experiments. Essay on principles". In: *Ann. Agricultural Sciences* (1923), pp. 1–51.

[48] Eoghan O'Neill. "Type I Tobit Bayesian Additive Regression Trees for Censored Outcome Regression". In: *Statistics and Computing* 34.4 (2024), pp. 1–19.

[49] Yasuhiro Omori. "Efficient Gibbs sampler for Bayesian analysis of a sample selection model". In: *Statistics & probability letters* 77.12 (2007), pp. 1300–1311.

[50] Andrew Donald Roy. "Some thoughts on the distribution of earnings". In: *Oxford economic papers* 3.2 (1951), pp. 135–146.

[51] Donald B Rubin. "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[52] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[53] Charles Spanbauer and Wei Pan. "Flexible Instrumental Variable Models With Bayesian Additive Regression Trees". In: *arXiv preprint arXiv:2210.01872* (2022).

[54] Rodney Sparapani, Charles Spanbauer, and Robert McCulloch. "Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package". In: *Journal of Statistical Software* 97 (2021), pp. 1–66.

[55] Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. "Machine learning estimation of heterogeneous treatment effects with instruments". In: *Advances in Neural Information Processing Systems* 32 (2019).

[56] James Tobin. "Estimation of relationships for limited dependent variables". In: *Econometrica: journal of the Econometric Society* (1958), pp. 24–36.

[57] Martijn Van Hasselt. "Bayesian inference in a sample selection model". In: *Journal of Econometrics* 165.2 (2011), pp. 221–232.

[58] Dootika Vats and Christina Knudson. "Revisiting the gelman–rubin diagnostic". In: *Statistical Science* 36.4 (2021), pp. 518–529.

[59] Wim PM Vijverberg. "Measuring the unidentified parameter of the extended Roy model of selectivity". In: *Journal of Econometrics* 57.1-3 (1993), pp. 69–89.

[60] Angela Vossmeyer. "Sample selection and treatment effect estimation of lender of last resort policies". In: *Journal of Business & Economic Statistics* 34.2 (2016), pp. 197–212.

[61] Stefan Wager and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242.

[62] Christopher Winship and Robert D Mare. "Models for sample selection bias". In: *Annual review of sociology* 18.1 (1992), pp. 327–350.

[63] Arnold Zellner. "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias". In: *Journal of the American statistical Association* 57.298 (1962), pp. 348–368.

[64] Xiao Zhang, W John Boscardin, Thomas R Belin, Xiaohai Wan, Yulei He, and Kui Zhang. "A Bayesian method for analyzing combinations of continuous, ordinal, and nominal categorical data with missing values". In: *Journal of multivariate analysis* 135 (2015).

A Appendix

A.1 Bounds for 2x2 sub-matrices

A.1.1 Bounds for Ω_{32}

To ensure positive definiteness in the sub-matrix $\begin{pmatrix} \Omega_{22} & \Omega_{23} \\ \Omega_{32} & \Omega_{33} \end{pmatrix}$, its determinant must be strictly greater than 0. For a fixed proposal of Ω_{33} and $\Omega_{22} = 1$, this results in the following bounds on Ω_{32}, displayed in Equation 25:

$$\begin{aligned} \texttt{det} \begin{pmatrix} 1 & \Omega_{23} \\ \Omega_{32} & \Omega_{33} \end{pmatrix} &> 0 \\ \Omega_{33} - (\Omega_{32}\Omega_{23}) &> 0 \\ \Omega_{33} &> \Omega_{32}^2 \\ -\sqrt{\Omega_{33}} < \Omega_{32} &< \sqrt{\Omega_{33}} \end{aligned} \tag{25}$$

A.1.2 Bounds for Ω_{31}

To ensure positive definiteness in the sub-matrix $\begin{pmatrix} 1 & \Omega_{13} \\ \Omega_{31} & \Omega_{33} \end{pmatrix}$, its determinant must be strictly greater than 0. For a fixed proposal of Ω_{33} and $\Omega_{11} = 1$, this results in the following bounds on Ω_{31}, displayed in Equation 26:

$$\begin{aligned} \texttt{det} \begin{pmatrix} 1 & \Omega_{13} \\ \Omega_{31} & \Omega_{33} \end{pmatrix} &> 0 \\ \Omega_{33} - (\Omega_{31}\Omega_{13}) &> 0 \\ \Omega_{33} &> \Omega_{31}^2 \\ -\sqrt{\Omega_{33}} < \Omega_{31} &< \sqrt{\Omega_{33}} \end{aligned} \tag{26}$$

A.1.3 Bounds for Ω_{42}

To ensure positive definiteness in the sub-matrix $\begin{pmatrix} \Omega_{22} & \Omega_{24} \\ \Omega_{42} & \Omega_{44} \end{pmatrix}$, its determinant must be strictly greater than 0. For a fixed proposal of Ω_{44} and $\Omega_{22} = 1$, this results in the following bounds on Ω_{42}, displayed in Equation 27:

$$\begin{aligned} \texttt{det} \begin{pmatrix} 1 & \Omega_{24} \\ \Omega_{42} & \Omega_{44} \end{pmatrix} &> 0 \\ \Omega_{44} - (\Omega_{42}\Omega_{24}) &> 0 \\ \Omega_{44} &> \Omega_{42}^2 \\ -\sqrt{\Omega_{44}} < \Omega_{42} &< \sqrt{\Omega_{44}} \end{aligned} \tag{27}$$

A.1.4 Bounds for Ω_{51}

To ensure positive definiteness in the sub-matrix $\begin{pmatrix} 1 & \Omega_{15} \\ \Omega_{51} & \Omega_{55} \end{pmatrix}$, its determinant must be strictly greater than 0. For a fixed proposal of Ω_{55} and $\Omega_{11} = 1$, this results in the following bounds on Ω_{51}, displayed in Equation 28:

$$\begin{aligned} \texttt{det} \begin{pmatrix} 1 & \Omega_{15} \\ \Omega_{51} & \Omega_{55} \end{pmatrix} &> 0 \\ \Omega_{55} - (\Omega_{51}\Omega_{15}) &> 0 \\ \Omega_{55} &> \Omega_{51}^2 \\ -\sqrt{\Omega_{55}} < \Omega_{51} &< \sqrt{\Omega_{55}} \end{aligned} \tag{28}$$

A.2 Bounds for 3x3 sub-matrices

A.2.1 Bounds on Ω_{21}

To ensure positive definiteness in the sub-matrix $\begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix}$, corresponding with the sub-matrix Ω_D, its determinant must be strictly greater than 0. For a fixed proposal of Ω_{33}, and fixed proposal of Ω_{31} and Ω_{32} obtained from the previous bounds, along with $\Omega_{11} = \Omega_{22} = 1$, this results in the following bounds on Ω_{21}, displayed in Equation 29:

$$\begin{aligned}
\texttt{det}\begin{pmatrix} 1 & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & 1 & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix} &> 0 \\
1 \times \texttt{det}\begin{pmatrix} 1 & \Omega_{23} \\ \Omega_{32} & \Omega_{33} \end{pmatrix} - \Omega_{12} \times \texttt{det}\begin{pmatrix} \Omega_{21} & \Omega_{23} \\ \Omega_{31} & \Omega_{33} \end{pmatrix} + \Omega_{13} \times \texttt{det}\begin{pmatrix} \Omega_{21} & 1 \\ \Omega_{31} & \Omega_{32} \end{pmatrix} &> 0 \\
(\Omega_{33} - \Omega_{32}\Omega_{23}) - \Omega_{12}(\Omega_{21}\Omega_{33} - \Omega_{23}\Omega_{31}) + \Omega_{13}(\Omega_{21}\Omega_{32} - \Omega_{31}) &> 0 \\
\Omega_{33} - \Omega_{32}^2 - \Omega_{21}^2\Omega_{33} + \Omega_{21}\Omega_{23}\Omega_{31} + \Omega_{21}\Omega_{13}\Omega_{32} - \Omega_{31}^2 &> 0 \\
-\Omega_{21}^2\Omega_{33} + 2\Omega_{21}\Omega_{23}\Omega_{31} + \Omega_{33} - \Omega_{32}^2 - \Omega_{31}^2 &> 0
\end{aligned}$$

This is a quadratic equation of the form: $-ax^2 + 2bx + c > 0$, which has the following solution:

$$\frac{b}{a} - \sqrt{\frac{(ac + b^2)}{a^2}} < x < \frac{b}{a} + \sqrt{\frac{(ac + b^2)}{a^2}}, \tag{29}$$

where $x = \Omega_{21}, a = \Omega_{33}, b = \Omega_{32}\Omega_{31}$ and $c = \Omega_{33} - \Omega_{32}^2 - \Omega_{31}^2$. To simplify notation for the last part, these lower and upper bounds are referred to as: $lb_{21.D} < \Omega_{21} < ub_{21.D}$.

In addition, as mentioned in Section 4, an extra bound is required to also ensure positive definiteness of the other 3x3 sub-matrix, corresponding with Ω_A: $\begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{14} \\ \Omega_{21} & \Omega_{22} & \Omega_{24} \\ \Omega_{41} & \Omega_{42} & \Omega_{44} \end{pmatrix}$. Here it is important to note, that the values of $\Omega_{41}, \Omega_{42}, \Omega_{44}$ are not the proposal values, but the current values of those elements at the time of this step. The calculation is exactly the same as above, only replacing the 3's with 4's, leading to the same bounds as in Equation 29, but now with: $x = \Omega_{21}, a = \Omega_{44}, b = \Omega_{42}\Omega_{41}$ and $c = \Omega_{44} - \Omega_{42}^2 - \Omega_{41}^2$. These bounds are referred to as: $lb_{21.A} < \Omega_{21} < ub_{21.A}$

Combining these two lower and two upper bounds, such that a proposed value will always result in positive definiteness of both 3x3 sub-matrices, results in the following lower and upper bounds for the proposal of Ω_{21}:

$$\texttt{max}\{lb_{21.D}, lb_{21.A}\} < \Omega_{21} < \texttt{min}\{ub_{21.D}, ub_{21.A}\} \tag{30}$$

A.2.2 Bounds on Ω_{41}

To ensure positive definiteness in the sub-matrix corresponding with Ω_A in the second step of the adjusted RW-MH algorithm: $\begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{14} \\ \Omega_{21} & \Omega_{22} & \Omega_{24} \\ \Omega_{41} & \Omega_{42} & \Omega_{44} \end{pmatrix}$, its determinant must be strictly greater than 0. For a fixed proposal of Ω_{44}, a fixed proposal of Ω_{42} using the previously calculated bounds, and the value of Ω_{21} which was either accepted or rejected in the previous RW-MH step, along with $\Omega_{11} = \Omega_{22} = 1$, this results in the following bounds displayed in Equation 31 :

$$\begin{aligned}
\texttt{det}\begin{pmatrix} 1 & \Omega_{12} & \Omega_{14} \\ \Omega_{21} & 1 & \Omega_{24} \\ \Omega_{41} & \Omega_{42} & \Omega_{44} \end{pmatrix} &> 0 \\
1 \times \texttt{det}\begin{pmatrix} 1 & \Omega_{24} \\ \Omega_{42} & \Omega_{44} \end{pmatrix} - \Omega_{12} \times \texttt{det}\begin{pmatrix} \Omega_{21} & \Omega_{24} \\ \Omega_{41} & \Omega_{44} \end{pmatrix} + \Omega_{14} \times \texttt{det}\begin{pmatrix} \Omega_{21} & 1 \\ \Omega_{41} & \Omega_{42} \end{pmatrix} &> 0 \\
(\Omega_{44} - \Omega_{42}\Omega_{24}) - \Omega_{12}(\Omega_{21}\Omega_{44} - \Omega_{24}\Omega_{41}) + \Omega_{14}(\Omega_{21}\Omega_{42} - \Omega_{41}) &> 0 \\
\Omega_{44} - \Omega_{42}^2 - \Omega_{21}^2\Omega_{44} + \Omega_{21}\Omega_{24}\Omega_{41} + \Omega_{21}\Omega_{14}\Omega_{42} - \Omega_{41}^2 &> 0 \\
-\Omega_{41}^2 + 2\Omega_{41}\Omega_{21}\Omega_{24} + \Omega_{44} - \Omega_{42}^2 - \Omega_{21}^2\Omega_{44} &> 0
\end{aligned}$$

This is a quadratic equation of the form: $-x^2 + 2bx + c > 0$, which has the following solution:

$$b - \sqrt{b^2 + c} < x < b + \sqrt{b^2 + c}, \tag{31}$$

where $x = \Omega_{41}, b = \Omega_{21}\Omega_{24}$ and $c = \Omega_{44} - \Omega_{42}^2 - \Omega_{21}^2\Omega_{44}$.