Maynooth University
National University of Ireland Maynooth

Smoothness and covariance structure modelling in Bayesian machine learning models

A dissertation submitted for the degree of
Doctor of Philosophy

By:

Mateus Maia Marques

Under the supervision of:

Prof. Andrew C. Parnell

Dr. Keefe Murphy

Hamilton Institute
National University of Ireland Maynooth
Ollscoil na hÉireann, Má Nuad

March 2024

Ai, ai que saudade eu tenho da Bahia,
Ai, se eu escutasse o que mamãe dizia

Declaration

I hereby declare that I have produced this manuscript without the prohibited assistance of any third parties and without making use of aids other than those specified.

The thesis work was conducted from April 2020 to March 2024 under the supervision of Professor Andrew C. Parnell and Dr. Keefe Murphy in the Hamilton Institute, National University of Ireland Maynooth.

Mateus Maia Marques.
Maynooth, Ireland,
March 2024

Sponsor

This work was supported by Science Foundation Ireland Career Development Award grant number 17/CDA/4695 and SFI Research Centre Award 12/RC/2289P2.

Collaborations

Andrew C. Parnell: As my supervisor, Prof. Parnell (Maynooth University) supervised and collaborated on the work of all chapters.

Keefe Murphy: As my supervisor, Dr. Murphy (Maynooth University) supervised and collaborated on the work of all chapters.

Jonas Esser: As joint first author of Chapter 4, J. Esser (Vrije Universiteit Amsterdam) jointly developed the main modelling approach, had a hand in writing some R code for plotting purposes, and contributed writing on the background understanding of cost-effectiveness analysis and interpretation of the case study results.

Judith E. Bosmans, **Johanna Maria van Dongen**, and **Thomas Klausch**: As co-authors of Chapter 4, these collaborators contributed their expertise on cost-effectiveness analyses in healthcare to both the literature review and the interpretation of results in the main application. J. M. van Dongen made available the TTCM data analysed in Section 4.6.

Publications

The chapters contained in this thesis have been either published in a peer-reviewed journal, submitted to a peer-reviewed journal, or in preparation for submission. Chapter 3 has been published in the journal *Computational Statistics & Data Analysis*. In Chapter 4, Jonas Esser and Mateus Maia are joint first authors.

Peer-reviewed journal articles:

- **Maia, Mateus**, Keefe Murphy, and Andrew C. Parnell (2024) "GP-BART: A novel Bayesian additive regression trees approach using Gaussian processes". *Computational Statistics & Data Analysis* 190: 107858. https://doi.org/10.1016/j.csda.2023.107858.

Submitted articles (under review):

- **Esser, Jonas, Maia, Mateus**, Andrew C. Parnell, Judith Bosmans, Hanneke van Dongen, Thomas Klausch, and Keefe Murphy (2024+) "Seemingly unrelated Bayesian additive regression trees for cost-effectiveness analyses in healthcare". *The Annals of Applied Statistics* X:X–X. arXiv:2404.02228.

Articles in preparation:

- **Maia, Mateus**, Andrew C. Parnell, and Keefe Murphy (2024+) "Incorporating smoothness in Bayesian additive regression trees via penalised spines".

Contents

Abstract

Bayesian additive regression trees (BART) is a Bayesian tree-based model which can provide high predictive accuracy in both classification and regression problems. Within the Bayesian paradigm, regularisation is achieved by defining priors which ensure that each tree contributes modestly to the overall ensemble, thereby enhancing generalisation. Consequently, BART has proven to be very useful in a wide array of applications.

However, the standard BART model is limited in certain respects. This thesis introduces some novel extensions to the BART framework to address certain key shortcomings. The inherent lack of smoothness, which is intrinsic to the piecewise-constant nature of the decision trees, is the motivation behind two of our proposals. The first involves the incorporation of Gaussian processes while the second uses penalised splines in the terminal nodes. Both of these novel approaches yield demonstrable improvements from the points of view of predictive accuracy and uncertainty calibration in extensive simulations and real-world applications.

Another drawback of the standard BART model is that it is designed for predicting univariate outcomes. We introduce a third extension to embed BART in the seemingly unrelated regression framework to deal with multiple outcomes and model the covariance structure arising from their joint distribution. The method is applied in a causal setting in order to determine the cost-effectiveness of a novel medical intervention.

The incorporation of penalised splines is designed to introduce smoothness to BART's predictions. Concurrently, the extension to model multivariate outcomes within a seemingly unrelated regression framework enhances BART by structuring the covariance among responses. The synthesis of Gaussian processes with BART exemplifies this dual enhancement, simultaneously facilitating smooth predictive surfaces and capturing structured dependency, although the latter is within the feature space.

Acknowledgements

I would like to express my gratitude to my supervisors Prof. Andrew Parnell and Dr. Keefe Murphy. Firstly, I want to thank Prof. Parnell for giving me the opportunity to work with him. Through several meetings, discussions, and projects, I learned how to conduct statistical research and work ethically. From the first meeting, where I saw Andrew doing statistics on the board while wearing a suit, to the last, I am absolutely certain that we achieved our goals successfully, and I will be eternally grateful.

Secondly, I would like to thank Dr. Murphy for all the support, guidance, and patience over these years. With Keefe's assistance, I learned how to achieve and strive for excellence in academic research and output. Additionally, the persistent questions and pedantic discussions during our statistical sessions brought back the joy that initiated my academic journey, which began with just a notebook, a pen, a few books, and an R console. One of the main reasons I pursued the academic path was to have moments like those where we could thoroughly discuss, understand, try, fail, and explore the realm of statistics, and this was always brought back during those times.

I also wish to acknowledge my previous mentors, especially Anderson Ara, who ignited the statistician in me by presenting the 'word of Breiman'. Additionally, this work would not have been possible without funding from Science Foundation Ireland, who continue to provide fertile ground for the growth of quality science.

I also want to extend my thanks to all the Hamilton staff, particularly Rosemary and Kate, not only for their help with bureaucracy but also for the kind smiles and small chats every early morning, which gave me the energy to start each day.

Furthermore, throughout my four years at the Hamilton Institute, I would like to thank my friends Estevão and Alessandra. Although both of them left earlier, they were fundamental during my time at the institute, and their absence was felt every single day. André also played an important role in this, as did, Victor, Samara, Nahia and the *hermanos y hermanas*. Outside the institute, I am incredibly grateful to Maria for being my sunshine on the cloudiest days in Ireland for many, many days.

Finally, I would like to express a few words in Portuguese For those who have always supported and devoted much of their lives so I could reach this point: muito obrigado aos meus pais, Niltinho e Sandra. Vocês foram fundamentais nessa jornada, e sempre nos momentos difíceis busquei em vocês e nos seus exemplos a força para seguir adiante. Meu irmão, você também faz parte disso. Gostaria de dedicar à minha madrinha também. Sai, mesmo estando longe e sem poder me despedir de você como eu gostaria, espero que você esteja orgulhosa onde quer que você esteja.

List of Figures

List of Tables

CHAPTER 1

Introduction

This thesis is presented in the form of three distinct, self-contained chapters. This overall introduction chapter aims to outline the content contained within the chapters that follow and to draw parallels between their overlapping themes and purposes, where appropriate. Broadly speaking, this thesis describes some extensions to the Bayesian additive regression trees framework. We begin by providing relevant background motivation and then present dedicated chapter summaries in Section 1.2.

1.1 Motivation

Binary decision trees are non-parametric statistical models based on recursively partitioning the feature space to create subsets that are homogeneous with respect to a response variable. Their advantages include the ability to handle both categorical and continuous covariates, high interpretability, and the flexibility to approximate functions without specifying a parametric functional form. Additionally, they are adaptable to various tasks, including continuous regression and classification. These properties have made binary decision trees a popular choice among researchers and practitioners for statistical analysis, further supported by the widespread availability of implementations across numerous software packages and programming languages.

Among the earliest examples of tree-based approaches specifically developed for predictive purposes are those introduced by Belson (1959) and Morgan and Sonquist (1963), concentrating on the examination of survey data. Most of the current methodologies, which continue to be widely used and recognized, are based on the adoption of a "greedy search" technique for constructing trees. This technique avoids solving the optimisation problem for the entire tree structure in favor of initiating at the root node and proceeding downward in a sequential manner to identify the optimal tree configuration. Murthy and Salzberg (1995) provided empirical validation that such a greedy search strategy consistently approaches the performance of an ideal tree. Notable implementations of this framework include the CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993) algorithms, with the latter being an extension of the ID3 algorithm (Quinlan, 1986).

Subsequently, Chipman et al. (1998) introduced a Bayesian approach to the CART algorithm by establishing a prior structure for the tree and its parameters, and conducting a stochastic search to navigate the tree space. This method enables deriving a posterior distribution over trees, thereby assigning higher probabilities to more accurate ('good') trees. A similar methodology was also suggested by Denison et al. (1998), with notable distinctions outlined by Chipman et al. (1998). These include basing the tree specification on the actual count of terminal nodes using a truncated Poisson distribution and employing Reversible Jump Markov Chain Monte Carlo methods (RJ-MCMC; Green, 1995) for tree-space exploration.

Simultaneously, in the late 1990s, Breiman introduced several ensemble methods that leveraged tree-based models as their core components, such as bagging (Breiman, 1996) and arcing (Breiman, 1998) ensembles. These techniques, along with the work of Ho (1998), laid the groundwork for what would later emerge as one of the most acclaimed and widely used methods — random forests, officially introduced by Breiman (2001). In a different vein, Friedman (2001) proposed another class of tree ensembles characterised by additive functions, where small regression trees, produced through CART, contribute in an additive manner to modelling the variance of the target variable. The rise of ensemble models marked a significant phase in the development of novel and robust statistical models.

Within a Bayesian context and drawing inspiration from earlier ensemble methods, Chipman et al. (2010) introduced Bayesian additive regression trees (BART), constituting a robust ensemble of Bayesian CART trees. BART adeptly handles both regression and classification tasks by modelling a univariate response with shallow trees, which are regularised through specific priors on the tree structure and conditional prior distributions for leaf parameters given the tree structure. This setup has the same spirit as the approach of Friedman (2001), with each tree making a modest contribution to the overall ensemble. The ensemble's additive nature facilitates sampling the joint distribution of all trees through a Bayesian backfitting algorithm (Hastie and Tibshirani, 2000), which serves as the models' primary sampling mechanism. BART has demonstrated excellent predictive performance and uncertainty calibration, benefits attributed to its Bayesian underpinnings. Moreover, BART retains most of the adaptability and advantages of tree ensembles, making it a versatile tool in the literature on tree-based models (Dorie et al., 2019; Sparapani et al., 2020; Kim, 2022; Wu et al., 2021; Cao et al., 2023).

Over a decade since its initial publication, BART has inspired an active research environment, leading to numerous extensions for various scenarios not originally encompassed by its assumptions. An excellent review of recent advances to BART is provided by Linero (2017). Key developments include adapting BART for high-dimensional data (Linero, 2018), using probabilistic splitting rules (Linero and Yang, 2018), tailoring BART for survival analysis (Sparapani et al., 2016), covering spatial extensions (Müller et al., 2007; Kim, 2022), addressing heteroscedasticity (Pratola et al., 2020), generalising the framework beyond conditional conjugacy (Linero, 2022a), adding model trees as building-blocks of the ensemble (Prado et al., 2021), and providing visual tools to aid interpretability (Inglis et al., 2024), among others. Theoretical advances have also been developed to fill the lack of understanding as to precisely why BART has worked so well (Ročková and Saha, 2019; Ročková and Van der Pas, 2020). In addition, BART has emerged as a central method in various causal inference applications. One of the pioneering showcases of BART in this domain was due to Hahn et al. (2020), which achieved prominence by surpassing competing methods in causal analysis benchmarks (Dorie et al., 2019). Comprehensive reviews of BART's role in causal inference have been provided by Hill et al. (2020) and Linero and Antonelli (2023).

Enhancing the computational performance of BART is also an active field of research. He and Hahn (2023) introduces a stochastic hill-climbing technique claimed to be an efficient adaptation of BART, enabling faster model estimation. Moreover, BART is readily accessible in open-source statistical software like R (R Core Team, 2024), bolstered by well-crafted and performance-tuned packages such as `dbarts` (Dorie et al., 2024), `BART` (Sparapani et al., 2021), and `bartMachine` (Kapelner and Bleich, 2016). These packages offer proficient implementations that support a variety of outcome types including continuous, binary, categorical, and time-to-event data, enabling the model to be readily adapted to different application settings.

Despite the successful general performance of BART, the model relies on assumptions that may be violated for different applications, and still possesses some key limitations which have yet to be adequately addressed. In this work, we aim to extend BART in a few different directions and offer further alternatives to practitioners. One such limitation concerns the additive and piecewise-constant nature intrinsic to the tree-based methodology. This was mitigated through two of the proposed approaches in this thesis. First, the novel GP-BART method was developed in which a Gaussian process (GP) prior was integrated into the conditional distribution of the terminal nodes, predicated on the tree structure. Given the inherent capacity of GPs to exhibit smoothness for particular kernel functions, which delineate their covariance structure, the GP-BART model facilitates the incorporation of smoothing within an entirely Bayesian paradigm.

Regarding this same limitation, we introduce another extension, the novel spBART model, which adds smooth effects through the employment of additive functions, as initially advocated by Friedman and Silverman (1989) within generalised linear models. Notably, these functions are represented through classes of smooth functions, with penalised splines (Eilers and Marx, 1996) emerging as a prevalent choice in the additive models literature. Thus, we incorporated these additive elements as foundational components of the Bayesian CART trees that constitute our enhanced spBART model.

Another underlying limitation of the model that we are tackling is with respect to the fact that BART was principally developed for predicting univariate responses. Although recent adaptations by Peruzzi and Dunson (2022), Um et al. (2023), and McJames et al. (2023) have extended the scope of BART to encompass multivariate responses, integrating these across the ensemble of trees, these particular multivariate BART versions are themselves limited by some assumptions which may not be universally valid in all application settings. Motivated by challenges associated with multivariate outcomes in the cost-effectiveness analysis (CEA) framework, we introduce a generalisation we refer to as seemingly unrelated BART (suBART) for the concurrent modelling of multiple responses linked to cost and efficacy considerations. Though the methodology was derived from a specific case study, the suBART model is adaptable, extending to d-dimensional applications and accommodating both continuous and binary outcomes. Expanding upon Chipman et al. (2010)'s analogy between seemingly unrelated regression (SUR) and BART, our suBART framework distinguishes itself by jointly estimating d distinct tree ensembles and capturing the interdependencies among outcomes through a structured error covariance. This represents a departure from existing multivariate BART approaches which incorporate covariances at the terminal node level, under the often restrictive assumption of a shared tree structure across all d responses. The main connection between suBART and GP-BART here is their exploitation of structured covariance modelling, albeit in different ways. The former directly models the covariance structure among the responses, while the GP-BART also takes into account the covariance among instances from the feature space, especially in cases where there is spatial dependency underlying the data which is not covered by the original BART formulation.

1.2 Chapter summaries

This thesis is structured into three chapters, each dedicated to exploring innovative strategies for addressing scenarios where the foundational assumptions of Bayesian additive regression trees (BART) may be compromised, or where the model requires modifications to accommodate a variety of data types. Before delving into the main chapters, which are presented in the format of journal arti-

cles detailing each method, we provide an overview of the standard BART model itself in Chapter 2, beginning with a concise introduction to the underpinning elements of the ensemble methods, notably the classification and regression trees (CART) algorithm. Following this, we present a comprehensive overview of the Bayesian framework as applied to CART, concluding with a detailed exposition of BART itself, emphasising how the previously mentioned assumptions and limitations manifest within the model. The main articles are presented in the following order: Initially, we provide a concise overview of the GP-BART model, illustrating how the integration of GP priors facilitates the incorporation of smoothness and a covariance structure across observations. Following this, we summarise the adaptation of the covariance structure to the outcomes, employing the newly developed suBART model to analyse a dataset under the CEA framework. Lastly, we outline our development of an extension to the BART model which incorporates penalised splines within the terminal node structures to enhance model flexibility and prediction accuracy.

1.2.1 Chapter 3: GP-BART

BART models are based on tree-based structures that employ piecewise constant approximations to model the conditional expectation of a response given a set of covariates, which represent an obstacle to approximating a smooth function. This approach, as per several other regression methodologies, operates under the assumption that observations within a terminal node are drawn from an independent and identically distributed (i.i.d.) population. However, this assumption may not hold in scenarios involving spatial data, where the proximity of observations influences their correlation, as articulated by Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This spatial autocorrelation contradicts the i.i.d. assumption, necessitating models that can explicitly account for such dependencies. GPs offer a compelling solution by imposing a prior over function spaces, assuming that these functions exhibit a joint multivariate normal distribution (Willan et al., 2004). The covariance structure of these priors, defined through a kernel function, is inversely related to the distance between observations, thereby capturing the essence of spatial correlation where closer observations exhibit stronger cor-

relations. Moreover, the smooth surfaces produced by Gaussian processes offer the adaptability required to fine-tune smooth functions, ensuring that models can accurately reflect the underlying data structure with the needed flexibility. In Chapter 3, we demonstrate the capabilities of GP-BART through extensive simulation studies and applications to several real benchmark datasets, and find that it obtains superior metrics for prediction accuracy and uncertainty calibration relative to several tree-based competitors and other purely spatial methodologies.

1.2.2 Chapter 4: suBART

While the traditional BART is limited to univariate responses, it is common to encounter scenarios involving data-generating processes with multiple outcomes, where there is an interest in modelling these outcomes jointly given the existing correlation among them. In the field of cost-effectiveness analysis (CEA), policymakers base their decisions on models that need to account for the average treatment effect in terms of both cost and quality. It becomes evident that these two metrics are correlated, underscoring the necessity for them to be modeled jointly to ensure an accurate representation of the model. In the conventional CEA literature, the prevalent methodology involves the use of seemingly unrelated regression models to accommodate the joint distribution of responses. However, this approach is fundamentally parametric and built upon linear assumptions, posing challenges to correctly specifying the functional form of the model, especially if there are non-linear effects or multiple low-order interactions. In contrast, BART handles such complexities with remarkable efficiency. Despite its advantages, the standard BART framework does not support modelling multiple outcomes directly. To address this limitation within the CEA context, and overcome other limitations of existing multivariate BART approaches, we introduce an adaptation of BART referred to as suBART in Chapter 4. This extension enables the modelling of each individual response by associating it with distinct ensembles of trees, thus offering a sophisticated approach to analysing multiple outcomes simultaneously. We evaluate the suBART model for multivariate continuous outcomes and a further probit suBART extension which accommodates multivariate binary outcomes in a number of simulated settings, and subsequently apply the suBART model to data from an observational case study in the health economics context.

1.2.3 Chapter 5: spBART

In addressing the smoothness assumption through the GP prior in GP-BART, we encountered significant computational demands, leading to infeasible scenarios for larger datasets. An alternative strategy for achieving smoothness within a multiple regression context is through the incorporation of additive functions (Friedman and Silverman, 1989). These function classes are regarded as smoothers due to their ability to reparameterise the original feature space, thus accommodating non-linear behaviors effectively. Prior research, such as that by Prado et al. (2021), has demonstrated that the integration of model trees, which incorporate linear terms within the terminal nodes of the BART framework can mitigate the lack of smoothness stemming from its piece-wise construction. Accordingly, by proposing the integration of penalised splines into the BART framework, we aim to expand the versatility of the model tree BART approach. The resultant proposal in Chapter 5, which we term spBART, enhances BART's adaptability to complex data structures. Moreover, spBART can be perceived as facilitating model specification within the penalised-splines approach. This is because we develop a sampling strategy capable of determining the relevant sets of basis functions, thereby avoiding the need to pre-specify the model structure. As per GP-BART and spBART, we demonstrate the capabilities of spBART through extensive simulation studies, including a setting in which the main effects vary smoothly and/or contain discontinuities. Through an application, we further show that spBART outperforms or competes with other explicitly tree-based methods and other approaches from the literature on splines.

CHAPTER 2

A review of Bayesian additive regression trees

Before delving into the main chapters of this thesis, we begin by establishing the groundwork for the BART model with a concise review of the tree models which compose the ensemble. We start by introducing decision trees, the fundamental building blocks. We focus on their notation and key characteristics under a non-Bayesian perspective. Next, we review the Bayesian CART algorithms proposed by Chipman et al. (1998) and Denison et al. (1998). Here, we explain in further detail their core components and highlight the unique features of these approaches, particularly their prior specifications and the underlying learning algorithms. Finally, we present a concise review of BART describing the model and its main assumptions.

2.1 Decision trees

Decision trees can be seen as non-parametric statistical models based on partitioning algorithms which recursively partition the data into homogeneous subsets with respect to a response variable. Let $\mathbf{x}_i = \{x_i^{(1)}, \ldots, x_i^{(p)}\}$ be a p-dimensional predictor vector, with $\mathbf{X} \in \mathbb{R}^{n \times p}$ encapsulating the entirety of the design matrix, and $\mathbf{y}$ standing as the n-dimensional response vector. Here, each pair $(\mathbf{x}_i, y_i)$ constitutes an individual data observation, where $i = 1, \ldots, n$.

A binary decision tree, which we denote as $\mathcal{T}$, is constructed based on a set of splitting rules that define partitions or nodes within the feature space. The nodes in a decision tree are categorised into root, internal, and terminal nodes. The root node is the starting point of a tree. Internal nodes are intermediate nodes that are not terminal; each one possesses a specific splitting rule, which generally leads to two child nodes. For a continuous predictor, the splitting rule $\{x_i^{(j)} \leq c\}$ is composed by a splitting variable $(x_i^{(j)})$ and a splitting threshold (c), which directs the observations towards the left, if the splitting rule is satisfied, and to the right otherwise. In the case of a categorical predictor, the splitting rules are instead determined by a subset of categories C, whereby the observations are dichotomised into either $\{x_i^{(j)} \in C\}$ or $\{x_i^{(j)} \notin C\}$ and then allocated to the left or right child nodes accordingly. Terminal nodes, also known as leaves, are the endpoints of a tree, as there are no subsequent child nodes below them. Figure 2.1 illustrates the components described above for a single decision tree.

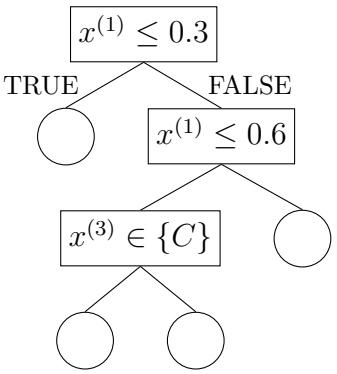


Figure 2.1: Example of a binary decision tree of depth 3 with three internal nodes (rectangles) and four terminal nodes (circles), with splitting rules of different types.

Learning the tree is usually accomplished via a greedy search. The nature of the response $\mathbf{y}$ defines which criteria can be used to select the splitting rules used to learn the tree, as well as which predicted value will be assigned to the observations in the leaf nodes. As all observations within a leaf are generally assigned the same predicted value, the main goal of a recursive partitioning algorithm is to search for splitting rules which optimise a given loss function in each partition of the tree. For any partition $\mathcal{R}$ within a tree $\mathcal{T}$ which subsets the data $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, the most common loss functions are presented below:

- Gini index (categorical outcomes):

$$\mathcal{C}_{\text{Gini}}(\mathcal{R}) \coloneqq \sum_{m=1}^{M} \hat{p}_m(1-\hat{p}_m),$$

  where $\hat{p}_m$ is the proportion of the category m in the partition $\mathcal{R}$.

- Shannon entropy (categorical outcomes):

$$\mathcal{C}_H(\mathcal{R}) \coloneqq -\sum_{m=1}^{M} \hat{p}_m \log \hat{p}_m.$$

- Mean squared error (continuous outcomes):

$$\mathcal{C}_{MSE}(\mathcal{R}) \coloneqq \sum_{(\mathbf{x},\, y)\, \in\, \mathcal{R}} (y-\bar{y})^2,$$

  where $\bar{y} = \frac{1}{|\mathcal{R}|} \sum_{(\mathbf{x},\, y)\, \in\, \mathcal{R}} y$ and $|\cdot|$ indicates cardinality.

In the process of identifying a new partition, the algorithm selects a predictor j and an associated cut-point k which jointly form a splitting rule that is used to further divide one of the partitions of the tree ($\mathcal{R}$) into two disjoint regions, $\mathcal{R}_{\text{left}}$ and $\mathcal{R}_{\text{right}}$, so that

$$\Delta_{\mathcal{R}}(j,k) = \{\mathcal{C}(\mathcal{R}) - \mathcal{C}(\mathcal{R}_{\text{left}}) - \mathcal{C}(\mathcal{R}_{\text{right}})\}$$

is maximised. Here, $\mathcal{C}(\cdot)$ denotes a generic loss function from the list above. Overall, this optimisation aims to identify the splitting rule that most decreases the loss function. This approach applies to both regression and classification settings. Ultimately, the predictions are made at the terminal node level in light of the chosen loss function. In regression settings using $\mathcal{C}_{\text{MSE}}$, for example, predictions are given by the mean of the y_i observations assigned to the given node while in classification settings using $\mathcal{C}_{\text{Gini}}$ or $\mathcal{C}_H$, predictions are given by the category m whose proportion $\hat{p}_m$ minimises the cost.

In theory, adhering strictly to the optimisation of a loss function to learn the tree structure might favour regions where the leaves would contain only a single observation, leading to significant overfitting. To mitigate this, early stopping

rules are generally employed, such as introducing a stopping complexity parameter ω_{cost}, whereby a split is avoided if $\Delta_{\mathcal{R}} < \omega_{\text{cost}}$. Alternative constraints include setting a predetermined maximum tree depth or specifying a minimal number of observations per leaf.

While these criteria might appear logical, they tend to be less effective compared to what is widely regarded as the most efficient method: grow an 'overfitted' tree and then apply cost-complexity pruning (Breiman et al., 1984). This approach involves pruning a larger tree from the terminal nodes upwards, resulting in a smaller set of subtrees. Following this, through cross-validation, test sample evaluations are applied to select the subtree characterised by the minimal estimated loss. The objective is to eliminate nodes that contribute minimally to the prediction.

2.2 Bayesian CART

As an alternative to the CART algorithm, which is a recursive partitioning method based on a greedy search, Chipman et al. (1998) proposed a Bayesian approach to identify optimal trees, referred to as BCART. The algorithm adopted to learn the tree structures is a principled stochastic search, subject to the prior specification of the tree and its parameters. Although Denison et al. (1998) also independently proposed a similar Bayesian framework for binary decision trees, we will focus on the work of Chipman et al. (1998), as it is the one on which most of the contributions in this thesis are fundamentally based. However, we endeavour in this section to both describe each component of BCART and clarify the main differences with the approach of Denison et al. (1998).

To begin, BCART assumes that

$$y_i = g\left(\mathbf{x}_i; \mathcal{T}, \mathcal{M}\right) + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \tau^{-1}\right) \tag{2.1}$$

for all $i = 1, \ldots, n$. Here, the decision tree $\mathcal{T}$ is described by a set of splitting rules and $\mathcal{M} = (\mu_1, \ldots, \mu_b)$ represents the set of terminal node parameters, where $\ell = 1, \ldots, b$ denotes the number of terminal nodes in tree $\mathcal{T}$. The function $g(\cdot; \mathcal{T}, \mathcal{M})$ assigns the value μ_ℓ to all observations $\mathbf{x}_i$ within the leaf ℓ. For greater flexibility, the residual precision τ can also be defined at the terminal node level τ_ℓ — referred

to as mean-variance shift model (Chipman et al., 1998) — but for conventional notation, we will adopt the common τ across all terminal nodes — referred to as the mean-shift model. In particular, we adopt the same formulation throughout the novel contributions introduced in the following chapters of this thesis.

The objective of the BCART algorithm, which can also be viewed as a (Bayesian) non-parametric statistical model, is to explore and sample from the posterior distribution $\mathcal{T} \,|\, \mathbf{X}, \mathbf{y}$. Although this is not a trivial task, the Metropolis-within-Gibbs algorithm has been proven to be a valuable tool capable of exploring the tree space effectively and updating the terminal node parameters. Given that the BCART model is characterised by $(\mathcal{M}, \mathcal{T})$, it is critical to set a joint prior distribution for these quantities. This can be done via decomposition of the joint prior distribution into $\pi(\mathcal{M} \,|\, \mathcal{T})\pi(\mathcal{T})$, under the assumption that the prior on the tree $\pi(\mathcal{T})$ is independent of the parameter collection $\mathcal{M}$.

Chipman et al. (1998) adopt a branching process prior for $\mathcal{T}$. Letting η denote a generic terminal node and assuming the terminal nodes are independent *a priori*, the prior on the tree structure hinges on the multiplication of independent priors for each node, which are in turn dictated by the rule probability $\mathbb{P}_{\text{rule}}(\eta, \mathcal{T})$ and split probability $\mathbb{P}_{\text{split}}(\eta, \mathcal{T})$. These two components correspond to the probability of assigning a given predictor and split point and the probability of a split to generate left and right child nodes, respectively. Recall Figure 2.1 for examples of splitting rules within a binary decision tree. The default prior setting for $\mathbb{P}_{\text{rule}}(\eta, \mathcal{T})$ assumes a discrete uniform distribution, reflecting the random assortment of predictors and the corresponding grid over the range of the selected predictor, within node η of tree $\mathcal{T}$. It is noteworthy that predictors can be either numerical or categorical.

Another important component of the prior on the tree structure is given by

$$\mathbb{P}_{\text{split}}(\eta, \mathcal{T}) = \frac{\alpha}{(1 + d_\eta)^\beta}, \tag{2.2}$$

where $d_\eta = 0, 1, 2, \ldots$ denotes the depth of the node η, and the pair of hyperparameters $\beta \geq 0$ and $0 \leq \alpha \leq 1$ respectively penalise the shape and size of the tree, making deeper trees less likely to be split further. This prior on the tree structure is a key difference between the BCART algorithm and the approach of Denison

et al. (1998). Denison et al. (1998) instead consider a zero-truncated Poisson distribution on the number of terminal nodes. However, this prior, in contrast to Equation (2.2), does not penalise the topology of the tree *a priori*. In other words, as long as two trees have the same number of leaves, they will have the same prior probability under this alternative approach, regardless of their topologies.

Given the assumption of prior independence across the leaves, the prior for the collection of parameters $\mathcal{M} = (\mu_1, \ldots, \mu_b)$ can be written as

$$\pi(\mathcal{M} \mid \mathcal{T}) = \prod_{\ell=1}^{n} \pi(\mu_\ell).$$

The μ_ℓ parameters are assumed to be i.i.d. and follow a normal distribution $\mu_\ell \mid \sigma, \mathcal{T} \sim (\bar{\mu}, \tau^{-1}/a)$, where a is a hyperparameter selected to ensure that the prior is spread out over the range of $\mathbf{y}$ values. Considering the mean-shift model, the prior assumed for the residual precision is $\tau \sim \text{Gamma}(a_\tau, d_\tau)$, which also achieves conjugacy. Specifying these priors facilitates the characterisation of the posterior distribution as being proportional to

$$\pi(\mathcal{T}, \mathcal{M}, \tau \mid \mathbf{y}, \mathbf{X}) \propto \pi(\mathbf{y} \mid \mathbf{X}, \mathcal{M}, \mathcal{T}, \tau)\, \pi(\mathcal{T})\, \pi(\tau)\, \pi(\mathcal{M} \mid \tau, \mathcal{T}),$$

which does not hold a closed-form. Nevertheless, sampling from this posterior distribution is feasible by sequentially sampling from

$$\pi(\mathcal{T} \mid \mathbf{y}, \mathbf{X}) \propto \pi(\mathcal{T}) \int\int \pi(\mathbf{y} \mid \mathbf{X}, \mathcal{T}, \mathcal{M}, \tau)\, \pi(\mathcal{M} \mid \mathcal{T}, \tau)\, \pi(\tau)\, \mathrm{d}\mathcal{M}\, \mathrm{d}\tau, \tag{2.3}$$

$$\pi(\mathcal{M} \mid \mathcal{T}, \tau) \propto \pi(\mathbf{y} \mid \mathbf{X}, \mathcal{T}, \mathcal{M}, \tau)\, \pi(\mathcal{M} \mid \mathcal{T}, \tau), \tag{2.4}$$

$$\pi(\tau \mid \mathbf{y}, \mathbf{X}, \mathcal{M}) \propto \pi(\mathbf{y} \mid \mathbf{X}, \mathcal{T}, \mathcal{M})\, \pi(\tau). \tag{2.5}$$

The samples from posterior distribution of the trees, as delineated in Equation (2.3), are obtained through a Metropolis-Hastings (MH) step, given the absence of a closed-form solution for the tree structure. The remaining parameters $\mathcal{M}$ and τ, as outlined in Equations (2.4) and (2.5) respectively, can be efficiently sampled using a Gibbs sampling method, thanks to the conjugacy of their priors.

Exploring the posterior distribution of the tree structure requires the use of a MH step. The associated acceptance ratio when proposing a new tree $\mathcal{T}^\star$ is given by

$$\alpha(\mathcal{T}, \mathcal{T}^\star) = \min\left\{1, \frac{\pi(\mathbf{y} \mid \mathbf{X}, \mathcal{T}^\star)\, \pi(\mathcal{T}^\star)\, q(\mathcal{T}^\star \to \mathcal{T})}{\pi(\mathbf{y} \mid \mathbf{X}, \mathcal{T})\, \pi(\mathcal{T})\, q(\mathcal{T} \to \mathcal{T}^\star)}\right\}. \tag{2.6}$$

Equation (2.6) can be succinctly described as containing the ratios (from left to right) of the likelihood, the prior, and the proposal distribution between a new tree $\mathcal{T}^\star$ and the current tree $\mathcal{T}$. Understanding the dynamics of each component of Equation (2.6) is crucial for a comprehensive grasp of how the MH step navigates the posterior distribution of the tree structure. Regarding the likelihood component, the aforementioned conjugacy allows for the terminal node parameters to be marginalised out, thus avoiding the need for reversible-jump Markov chain Monte Carlo (RJ-MCMC; Green, 1995). This approach circumvents the introduction of additional parameters associated with some new tree proposal, marking a significant departure from the methodology of Denison et al. (1998), which relies on RJ-MCMC to explore the tree space, given that μ_ℓ is not marginalised out of the likelihood functions in the MH step.

Under the stochastic search strategy developed by Chipman et al. (1998), a new tree can be proposed based on the following four moves: 'grow', 'prune', 'change' and 'swap'. The last one is an additional move when compared with the methodology of Denison et al. (1998). The type of move is randomly selected (with equal probability) to generate a modification of the current tree, and the nature of the modification is intuitive given the names of the moves. For the grow move, a terminal node is selected at random and divided into two new nodes by assigning a splitting rule, also chosen randomly, in accordance with the $\mathbb{P}_{\text{rule}}(\eta, \mathcal{T})$ defined in the prior on $\mathcal{T}$. The prune move is the reversible counterpart of the grow move; it selects a parent node of two terminal nodes at random and converts it into a single terminal node by collapsing its leaves. The change move selects an internal node at random and assigns it a new splitting rule, also chosen randomly, again following $\mathbb{P}_{\text{rule}}(\eta, \mathcal{T})$. Lastly, the swap move involves randomly selecting a parent-child pair of internal nodes to exchange splitting rules, unless the sibling shares the same rule. Figure 2.2 summarises all possible tree proposal moves.

We now illustrate how each component of Equation (2.6) is computed for each move. We begin with the grow move. The likelihood ratio changes solely for the node selected for growth and its children, as follows:

$$\frac{\pi\left(\mathbf{y} \mid \mathbf{X}, \mathcal{T}^\star\right)}{\pi\left(\mathbf{y} \mid \mathbf{X}, \mathcal{T}\right)} = \frac{\pi\left(\mathbf{y}_{\eta_{\text{L}}} \mid \mathbf{X}, \mathcal{T}^\star\right) \pi\left(\mathbf{y}_{\eta_{\text{R}}} \mid \mathbf{X}, \mathcal{T}^\star\right)}{\pi\left(\mathbf{y}_{\eta_{\text{G}}} \mid \mathbf{X}, \mathcal{T}\right)},$$

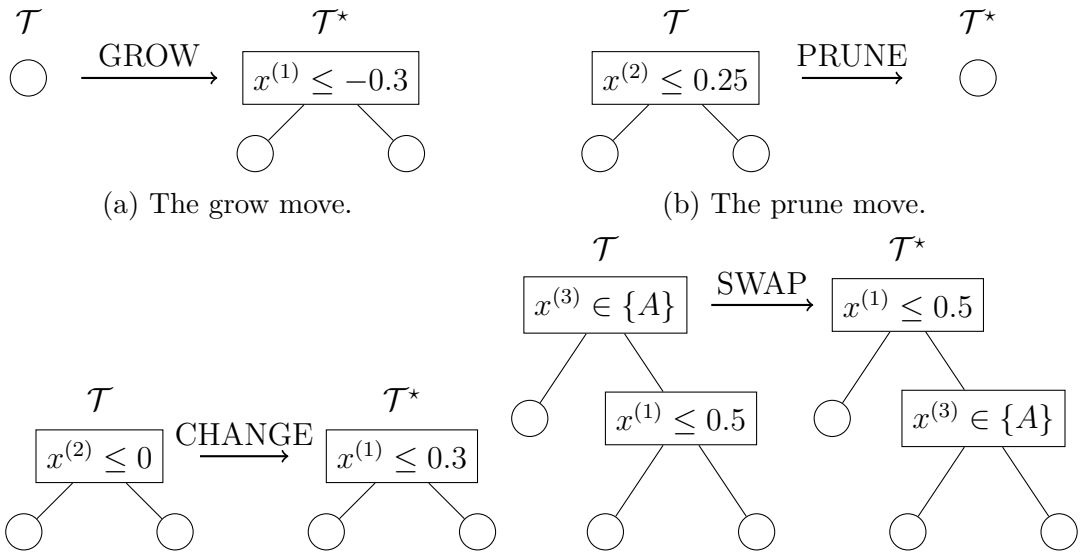
where $\mathbf{y}_{\eta_{\mathrm{G}}}$ denotes the response variable for the observations within the selected leaf, and $\mathbf{y}_{\eta_{\mathrm{L}}}$ and $\mathbf{y}_{\eta_{\mathrm{R}}}$ respectively denote the same for its prospective left and right children. Regarding the tree ratio, this term can be succinctly expressed as follows:

$$\frac{\pi\left(\mathcal{T}^{\star}\right)}{\pi\left(\mathcal{T}\right)}=\frac{\left(1-\mathbb{P}_{\mathrm{split}}\left(\eta_{\mathrm{L}},\mathcal{T}^{\star}\right)\right)\times\left(1-\mathbb{P}_{\mathrm{split}}\left(\eta_{\mathrm{R}},\mathcal{T}^{\star}\right)\right)}{\mathbb{P}_{\mathrm{split}}\left(\eta_{\mathrm{G}},\mathcal{T}\right)}.$$

Lastly, the transition ratio for the grow move is given by

$$\frac{q\left(\mathcal{T}^{\star}\rightarrow\mathcal{T}\right)}{q\left(\mathcal{T}\rightarrow\mathcal{T}^{\star}\right)}=\frac{\mathbb{P}\left(\mathrm{grow}\right)/b_{\mathcal{T}}}{\mathbb{P}\left(\mathrm{prune}\right)/\nu_{\mathcal{T}^{\star}}},$$

where $b_{\mathcal{T}}$ corresponds to the number of leaves of the tree $\mathcal{T}$ and $\nu_{\mathcal{T}}$ is the number of internal nodes that are parents of terminal nodes only in tree $\mathcal{T}^{\star}$. The $\mathbb{P}_{\mathrm{rule}}(\eta,\mathcal{T})$ terms are omitted for the sake of simplification, as they mostly cancel out when comparing the ratios of tree prior and proposal.



(a) The grow move. (b) The prune move.

(c) The change move. (d) The swap move.

Figure 2.2: Illustration of how the tree proposal moves (a) grow, (b) prune, (c) change and (d) swap operate. We stress, however, that the grow and prune moves are not restricted, as respectively depicted here, to growing or producing stumps.

While the details provided above are specific to the grow move, analogous computations apply for the prune move. However, the ratios for the prune move are typically the inverse of those discussed for the grow move. For the change and

swap moves, the relationships are even simpler, as these actions do not alter the topology of the tree, resulting in both the transition ratios and the prior ratios (excluding $\mathbb{P}_{\text{rule}}(\eta, \mathcal{T})$) being consistently equal to one. The primary distinction from the grow and prune moves lies in the likelihood term. Specifically, for the change move, the likelihood must account only for the pair of leaves that underwent modification, since the rest cancel out:

$$\frac{\pi\left(\mathbf{y} \mid \mathbf{X}, \mathcal{T}^{\star}\right)}{\pi\left(\mathbf{y} \mid \mathbf{X}, \mathcal{T}\right)} = \frac{\pi\left(\mathbf{y}_{\eta_{\text{L}}^{\star}} \mid \mathbf{X}, \mathcal{T}^{\star}\right) \pi\left(\mathbf{y}_{\eta_{\text{R}}^{\star}} \mid \mathbf{X}, \mathcal{T}^{\star}\right)}{\pi\left(\mathbf{y}_{\eta_{\text{L}}} \mid \mathbf{X}, \mathcal{T}\right) \pi\left(\mathbf{y}_{\eta_{\text{R}}} \mid \mathbf{X}, \mathcal{T}\right)},$$

where $\mathbf{y}_{\eta_{\text{L}}^{\star}}$ and $\mathbf{y}_{\eta_{\text{R}}^{\star}}$ represent the response variable assigned to the new left and right children associated with the rule proposed during the change move. A similar rationale applies to the swap move, where only the modified nodes are involved in the calculation of the constituent likelihood ratios.

2.3 BART

BART is an additive ensemble of Bayesian trees introduced by Chipman et al. (2010), chiefly for nonparametric regression tasks. BART aims to regularise the contribution of each tree, ensuring that each contributes equally to the overall ensemble. This regularisation results in improved generalisation capabilities, aligning with the principles outlined by Friedman (2001) in the context of gradient boosting. Within the Bayesian framework, BART achieves this regularisation by appropriately configuring the prior distributions on the topology and leaf parameters of its constituent trees. Unlike BCART, which uses only one tree, the combination of multiple trees allows BART to approximate non-linear effects and automatically capture lower-order interactions with minimal assumptions and without requiring pre-specification of the model's functional form.

The BART model is defined as

$$y_i = \sum_{t=1}^{T} g\left(\mathbf{x}_i; \mathcal{T}_t, \mathcal{M}_t\right) + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \tau^{-1}\right),$$

from which it is apparent that the model is an additive extension of the structure of the BCART models defined in Equation (2.1). One of the main differences is the total number of trees T composing the final model prediction, which is typically

large. Chipman et al. (2010) recommend a default of $T = 200$ and note that performance initially tends to improve dramatically as T is increased from one. Each tree is determined by it its own specific collection of parameters $\mathcal{M}_t = (\mu_{t1}, \ldots, \mu_{tb_t})$, where b_t denotes the number of leaves ℓ belonging to the tree $\mathcal{T}_t$.

Having defined the BCART components as building blocks, it is essential to highlight how their priors are adjusted within the BART framework. Firstly, the prior on the tree structure penalises deep trees in an attempt to prevent overfitting, an aspect commonly seen in recursive partitioning algorithms. By default, BART sets the hyperparameters for the prior $\pi(\mathcal{T})$ as $\alpha = 0.95$ and $\beta = 2$. In this setting, the prior probabilities of observing a tree with one, two, or three terminal nodes are approximately 0.05, 0.55, and 0.14, respectively, demonstrating a prior bias towards shallow trees. Another important component of BART relates to the prior distribution for the terminal node parameters, where

$$\mu_{t\ell} \sim \mathrm{N}\left(\mu_\mu, \tau_\mu^{-1}\right), \quad \text{with} \quad \tau_\mu = 4\kappa^2 T,$$

Scaling the response variable to fall within $[-0.5, 0.5]$ aids the elicitation of this prior. Consequently, $\mu_\mu = 0$ is typically assumed. The specification of τ_μ is then based on the induced $\mathrm{N}(0, T\tau_\mu^{-1})$ prior on the conditional expectation $\mathbb{E}[y_i \mid \mathbf{x}_i]$. By setting $\kappa = 2$, there is a 95% confidence level that the mean of the overall prediction from BART will reside within the scaled response interval. Note that the precision of the leaf parameters is proportional to the number of trees, which helps prevent any single tree or a small number of trees from dominating the fit, thereby enhancing the model's generalisation capabilities.

The prior choice for the residual precision is adjusted through its hyperparameters. Assuming the model offers greater flexibility than a linear method, a 'data-informed' prior is utilised. For a fixed shape parameter, the rate parameter of the gamma prior is chosen such that $\mathbb{P}(\tau > \hat{\tau}) = q$, where $\hat{\tau}$ represents a naïve estimate of the residual precision typically derived from an OLS estimator; Chipman et al. (2010) recommend $q = 0.9$.

If one aims to sample from the posterior distribution of the trees and their parameters, it is necessary to specify the prior distributions above assuming prior independence among the collection of terminal node parameters as follows:

$$\begin{aligned}\pi\left(\left(\mathcal{T}_1, \mathcal{M}_1\right), \ldots, \left(\mathcal{T}_T, \mathcal{M}_T\right), \tau\right) &= \left(\prod_{t=1}^{T} \pi\left(\mathcal{T}_t, \mathcal{M}_t\right)\right) \times \pi(\tau) \\ &= \left(\prod_{t=1}^{T} \pi\left(\mathcal{M}_t \mid \mathcal{T}_t\right) \pi\left(\mathcal{T}_t\right)\right) \times \pi(\tau) \\ &= \left(\prod_{t=1}^{T} \prod_{\ell=1}^{b_t} \pi\left(\mu_{t\ell} \mid \mathcal{T}_t\right) \pi\left(\mathcal{T}_t\right)\right) \times \pi(\tau).\end{aligned}$$

In the context of BCART, the focus was primarily on sampling from the posterior distribution $(\mathcal{T}, \mathcal{M}, \tau) \mid \mathbf{y}, \mathbf{X}$. With the transition to an ensemble of trees, the posterior distribution of interest is expressed as

$$\pi\left(\left(\mathcal{T}_1, \mathcal{M}_1\right), \ldots, \left(\mathcal{T}_T, \mathcal{M}_T\right), \tau \mid \mathbf{y}, \mathbf{X}\right).$$

Let $\mathcal{T}_{(-t)}$ represent the collection of all trees $\mathcal{T}_t$ except for tree t, and similarly, let $\mathcal{M}_{(-t)}$ denote the collection of all leaf parameters except for $\mathcal{M}_t$. As per BCART, the joint posterior distribution of the BART model does not have a closed-form. However, the posterior samples can be obtained via the following update scheme:

$$\begin{aligned}1: &\quad \mathcal{T}_1 \mid \mathbf{y}, \mathcal{T}_{(-1)}, \mathcal{M}_{(-1)}, \tau \\ 2: &\quad \mathcal{M}_1 \mid \mathbf{y}, \mathcal{T}_1, \ldots, \mathcal{T}_T, \mathcal{M}_{(-1)}, \tau \\ &\quad \vdots \\ 2T-1: &\quad \mathcal{T}_T \mid \mathbf{y}, \mathcal{T}_{(-T)}, \mathcal{M}_{(-T)}, \tau \\ 2T: &\quad \mathcal{M}_T \mid \mathbf{y}, \mathcal{T}_1, \ldots, \mathcal{T}_T, \mathcal{M}_{(-T)}, \tau \\ 2T+1: &\quad \tau \mid (\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_T, \mathcal{M}_T), a_\tau, d_\tau, \mathbf{y}.\end{aligned}$$

The main challenge of this sampling approach involves obtaining posterior samples for a given tree $\mathcal{T}_t$ and its collection of parameters $\mathcal{M}_t$, given the conditioning on the data, all other remaining trees $\mathcal{T}_{(-t)}$, and the associated collection of parameters $\mathcal{M}_{(-t)}$. The Bayesian back-fitting algorithm (Hastie and Tibshirani, 2000) is employed to address this conditional dependence. With the use of this algorithm, the dependence on $(\mathbf{y}, \mathcal{T}_{(-t)}, \mathcal{M}_{(-t)})$ is simplified through the partial residuals $\mathbf{R}_t = \mathbf{y} - \sum_{k \neq t} g(\mathbf{X}; \mathcal{T}_k, \mathcal{M}_k)$. As a result, the sampling process is revised accordingly to

$$
\begin{aligned}
1: &\quad \mathcal{T}_1 \mid \mathbf{R}_1, \tau \\
2: &\quad \mathcal{M}_1 \mid \mathbf{R}_1, \mathcal{T}_1, \tau \\
&\quad \vdots \\
2T-1: &\quad \mathcal{T}_T \mid \mathbf{R}_T, \tau \\
2T: &\quad \mathcal{M}_T \mid \mathbf{R}_T, \mathcal{T}_T, \tau \\
2T+1: &\quad \tau \mid (\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_T, \mathcal{M}_T), a_\tau, d_\tau, \mathbf{y}.
\end{aligned}
$$

Lastly, another significant consideration arises from the conjugate choice for the prior on $\mathcal{M}_t$, allowing for the expression

$$\pi\left(\mathcal{T}_t \mid \mathbf{R}_t, \tau\right) \propto \pi\left(\tau\right) \times \pi\left(\mathcal{T}_t\right) \int \pi\left(\mathbf{R}_t \mid \mathcal{T}_t, \mathcal{M}_t, \tau\right) \times \pi\left(\mathcal{M}_t \mid \mathcal{T}_t\right) \mathrm{d}\mathcal{M}_t.$$

Marginalising out the terminal node parameters $\mu_{t\ell}$ in this fashion ensures that changes in the tree topology, such as those produced by the grow and prune moves, will not modify the size of the parameter space when proposing a new tree as part of the MH step. Recall that this is a similar feature of BCART models which differentiates them from the approach of Denison et al. (1998) and bypasses the need for RJ-MCMC. Indeed, the sampling process for each step in the aforementioned sequential approach is consistent with that defined in Section 2.2; that is, the trees are sampled via MH steps and both the residual precision and the terminal node parameters are sampled via Gibbs steps. The primary distinction is that $\mathbf{R}_t$ now acts as the 'response variable' in each of these steps. Another difference from BCART — where the proposal moves have the same prior probability — is that the probabilities of the moves grow (0.25), prune (0.25), change (0.4) and swap (0.1) are not equal in BART.

The complete sampling strategy for the BART model is outlined in Algorithm 2.1. Practical implementations of this algorithm are available in a number of open-source R packages, including `BART` (Sparapani et al., 2021), `bartMachine` (Kapelner and Bleich, 2016), and `dbarts` (Dorie et al., 2024).

Algorithm 2.1: BART sampling algorithm

Input: $\mathbf{X}$, $\mathbf{y}$, T, M, and all hyperparameters of the priors.

Initialise: T tree stumps with $\mu_{t1} = 0 \, \forall \, t$, and τ drawn from its prior.

1 **for** $m = 1$ *to* M **do**

2 **for** $t = 1$ *to* T **do**

3 Calculate the partial residuals $\mathbf{R}_t = \mathbf{y} - \sum_{k \neq t} g(\mathbf{X}; \mathcal{T}_k, \mathcal{M}_k)$;

4 Propose a new tree $\mathcal{T}_t^\star$ by a grow, prune, change, or swap move;

5 Accept $\mathcal{T}_t^\star$ with probability

$$\alpha\left(\mathcal{T}_t, \mathcal{T}_t^\star\right) = \min\left\{1, \frac{\pi\left(\mathbf{R}_t \mid \mathcal{T}^\star, \tau\right)\pi\left(\mathcal{T}_t^\star\right)q\left(\mathcal{T}_t^\star \to \mathcal{T}_t\right)}{\pi\left(\mathbf{R}_t \mid \mathcal{T}, \tau\right)\pi\left(\mathcal{T}_t\right)q\left(\mathcal{T}_t \to \mathcal{T}_t^\star\right)}\right\}.$$

6 **for** $\ell = 1$ *to* b_t **do**

7 Update $\mu_{t\ell} \mid \mathbf{R}_{t\ell}, \mathcal{T}, \tau$.

8 **end**

9 **end**

10 Update $\tau \mid \ldots$.

11 **end**

Output: Samples from $\pi\left((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_T, \mathcal{M}_T), \tau \mid \mathbf{y}\right)$.

BART can be readily adapted for classification settings with binary responses $y_i \in \{0, 1\}$ using the probit model, by assuming $\mathbb{P}(y_i = 1 \mid \mathbf{x}) = \Phi(G(\mathbf{x}_i))$, where $G(\mathbf{x}_i) \equiv \sum_{t=1}^{T} g(\mathbf{x}_i; \mathcal{T}_t, \mathcal{M}_t)$. In this setup, the restriction $\tau = 1$ is imposed, and the precision hyperparameter for the leaf parameters is adjusted to $\tau_\mu = \frac{\kappa^2 T}{9}$, maintaining the default value of $\kappa = 2$, such that $G(\mathbf{x}_i)$ will with high probability be in the interval $[-3, 3]$. To accommodate the binary support of the outcome, modifications in the posterior sampling calculations are necessary, including the implementation of the data-augmentation technique by Albert and Chib (1993). This involves introducing latent variables $z_1, \ldots, z_n \sim \mathrm{N}(G(x_i), 1)$, constrained by $z_i > 0$ if $y_i = 1$ and $z_i < 0$ if $y_i = 0$. The sampling algorithm for the probit version of BART is obtained by replacing the y_i values in Algorithm 2.1 with these z_i values and introducing an additional step after line 9 for obtaining posterior draws of $z_i \mid \ldots$ using truncated normal distributions.

Another key feature of BART is its provision of measures to assess variable importance. From the combination of all trees, it can be observed that the most relevant predictors tend to appear more often among the tree splits. Therefore, measures such as the proportion of times a predictor appears among all the splitting rules can be used to indicate which variables and low-order interactions may be more relevant. This approach provides a 'model-free' variable selection within the BART framework in the sense that it does not require the imposition of parametric assumptions.

In summary, BART has been validated as a powerful tool, offering not only exceptional predictive accuracy but also reliable uncertainty calibration. Linero (2017) conducted a comprehensive review of advances to BART, highlighting numerous proposed extensions. Linero (2017) also compared the enhanced performance of BART in an array of application settings against other models commonly referenced in the literature, such as random forests (Breiman, 2001), MARS (Friedman, 1991), boosting (Friedman, 2001), neural networks, and support vector machines (Cortes and Vapnik, 1995).

However, as previously discussed in Chapter 1, the standard BART model described herein suffers from a number of limitations, such as the lack of smoothness given the piecewise-constant nature of the ensemble, and its initial formulation covering only univariate responses. In the chapters which follow, we exploit the scope for extending BART in light of these particular shortcomings.

CHAPTER 3

GP-BART: a novel Bayesian additive regression trees approach using Gaussian processes

The Bayesian additive regression trees (BART) model is an ensemble method extensively and successfully used in regression tasks due to its consistently strong predictive performance and its ability to quantify uncertainty. BART combines "weak" tree models through a set of shrinkage priors, whereby each tree explains a small portion of the variability in the data. However, the lack of smoothness and the absence of an explicit covariance structure over the observations in standard BART can yield poor performance in cases where such assumptions would be necessary. The Gaussian processes Bayesian additive regression trees (GP-BART) model is an extension of BART which addresses this limitation by assuming Gaussian process (GP) priors for the predictions of each terminal node among all trees. The model's effectiveness is demonstrated through applications to simulated and real-world data, surpassing the performance of traditional modelling approaches in various scenarios.

3.1 Introduction

Bayesian additive regression trees (BART; Chipman et al., 2010) is a probabilistic machine learning model that has proved successful in both regression and clas-

sification settings (Zhao et al., 2018; Zhang et al., 2020; Janizadeh et al., 2021). Effectively, BART is a non-parametric Bayesian regression approach which learns through sums of trees (Chipman et al., 1998), where each terminal node contribution is constrained by a regularising prior distribution. Given a vector of predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, the target function $f(\mathbf{x}_i)$ is obtained by aggregating the small contributions of each tree, which is similar in flavour to the small step updates of gradient boosting algorithms (Friedman, 2001).

Considering a univariate response and training observations denoted as $\{\mathbf{x}_i, y_i\}_{i=1}^n$, the standard BART model is given by

$$y_i \mid \mathbf{x}_i \sim \mathrm{N}\left(\sum_{t=1}^{T} h\left(\mathbf{x}_i; \mathcal{T}_t, \mathbf{L}_t\right), \tau^{-1}\right),$$

where the function h assigns a sampled value $\mu_{t\ell}$ to $\mathbf{x}_i$ within terminal node ℓ of the tree $\mathcal{T}_t$ across all T trees and the vector $\mathbf{L}_t = (\mu_{t1}, \ldots, \mu_{tb_t})$ collects the sampled mean parameters from the b_t terminal nodes in tree $\mathcal{T}_t$. Here, $\mathrm{N}(\cdot)$ denotes the normal distribution and τ is a residual precision term. In standard BART, terminal node parameters $\mu_{t\ell}$ are assigned a $\mathrm{N}(\mu_\mu, \tau_\mu^{-1})$ prior, where the hyperparameters are selected to shrink the influence of each tree.

Our novel GP-BART method modifies the standard BART by using the function g (replacing h) which assigns a vector of sampled values $\boldsymbol{\psi}_{t\ell}$ to the $n_{t\ell}$ observations in node ℓ of tree $\mathcal{T}_t$, rather than the single value $\mu_{t\ell}$ used by BART. This is achieved by assuming a Gaussian process (GP) prior over each terminal node with constant mean $\mu_{t\ell}$ and a covariance function whose parameters are defined at the tree level.

In recent years, several extensions and modifications to the original BART model have been proposed to cover different types of data and assumptions (Hill et al., 2020). To deal with the lack of smoothness, Linero and Yang (2018) presented a soft version of the BART model by advocating probabilistic split rules at the tree-building stage. Starling et al. (2020) presented a BART extension, also incorporating GPs, which guarantees smoothness over a single target covariate by applying Gaussian process priors for each terminal node over the targeted variable. Prado et al. (2021) proposed model trees BART that considers piecewise sums of linear functions at the terminal node level instead of piecewise sums of constants,

adding flexibility. Our GP-BART considers GP models at the terminal node level, and can be seen as a piecewise sum of GPs which are inherently smooth.

Notably, our GP-BART approach is coherent with previous work of Linero (2017), who identified that the BART model is itself a GP, conditional on the tree structures, with a non-parametrically learned covariance matrix whereby each element is described by the proportion of times the two corresponding design points are allocated to the same terminal nodes across all trees. Linero (2017) further showed that as $T \to \infty$, BART becomes a GP unconditionally. Therefore, it is natural to assume GP priors over the terminal nodes directly to circumvent the need for large T. More specifically, Linero (2017) also shows that the implied kernel under this relation between BART and GPs is a function of the $\mathcal{L}_1$ distances between design points (similar results were also found in Balog et al. (2016)). Following this, it is natural to allow kernels of other types, especially ones defined by different distance metrics. Here, we employ node-specific anisotropic exponentiated-quadratic kernels relying on squared Euclidean distances. Though these are parameterised, this enables covariance structures, more flexible than the one implied by the standard BART, to be learned non-parametrically when $T > 1$, which would be too difficult to pre-specify under a single GP, or even a sum of GPs without tree splits.

The treed Gaussian process (tGP; Gramacy and Lee, 2008) is another treed approach to GPs which defines all hyperparameters of a single GP at the terminal node level, thereby making it possible to incorporate non-stationarity into the model by varying the residual precision parameter across terminal nodes. However, to deal with the changing dimensions of the parameter space associated with growing and pruning a tree, this model requires the use of a reversible jump algorithm (Green, 1995), which comes with increased computational costs. Our GP-BART can also be seen as an additive ensemble of these treed GPs; though we define our priors and associated hyperparameters differently, the additive nature of the sum of GPs is shown here to yield superior performance. Finally, another example of previous work combining BART and GPs is provided by Wang et al. (2023), who use node-level GPs differently, as an extrapolation strategy for improving BART's predictions for exterior points outside the range of the training data. The authors describe their approach as a 'GPed tree', in contrast to the

'treed GP' of Gramacy and Lee (2008), and by extension GP-BART's ensemble of treed GPs.

We envisage our novel GP-BART framework being particularly suited for spatial data where smoothness in space is expected for certain covariate combinations, and thus useful in situations where GPs are commonly used (e.g., Banerjee et al., 2008; Gelfand and Schliep, 2016; Andugula et al., 2017; Xie et al., 2018). As well as GPs, we introduce a further novelty to allow for rotated splits. Traditional tree-based models can be interpreted as hyper-rectangles since each node is given in parallel-axis directions. This behavior leads to a staircase decision boundary which can inhibit the model's ability to approximate true boundaries. García-Pedrajas et al. (2007) propose non-linear projections of the tree models used in ensemble approaches to overcome this limitation, while Menze et al. (2011) describe an oblique forest model which selects optimal oblique directions using linear discriminant analysis. More recently, Blaser and Fryzlewicz (2016) proposed random rotation ensembles where the direction of rotation is selected randomly, yielding a more general decision boundary. In the GP-BART framework, the incorporation of random projections on various directions allows for splitting rules that are not limited to exclusively parallel axes. This flexibility enables the tree search algorithm to explore a broader sample space of the tree distribution, aiming to mitigate the issue of poor mixing (Wu et al., 2007). The rotation moves can also be interpreted as another way to represent and model complex interactions among variables and should not be seen as strictly restricted to spatial features.

The remainder of this chapter is structured as follows. Section 3.2 describes the GP-BART model, with mathematical formulations and key specifications. Section 3.3 contains the sampling algorithm and describes prediction settings and uncertainty estimation. Sections 3.4 and 3.5 provide comparisons between GP-BART and other methods in simulated and real-data benchmarking scenarios, respectively. Finally, Section 3.6 presents conclusions regarding the proposed algorithm, some limitations, and potential future work. We note that an implementation of our method is available in the R package `gpbart`, which is written in C++ and available at: `https://github.com/MateusMaiaDS/gpbart`, with which all results were obtained.

3.2 Gaussian processes Bayesian additive regression trees

For simplicity, we begin with the notation for a single tree model. Let $\mathcal{T}_1$ be a binary splitting tree with b_1 terminal nodes and let $\mathbf{G}_1 = (\{\mu_{11}, \boldsymbol{\phi}_1, \nu\}, \ldots, \{\mu_{1b}, \boldsymbol{\phi}_1, \nu\})$ denote the sets of parameters associated with each terminal node's GP. Each GP, denoted by $\mathcal{GP}_{1\ell}(\boldsymbol{\mu}_{1\ell}, \boldsymbol{\Omega}_{1\ell}(\boldsymbol{\phi}_1, \nu))$, is characterised by a constant mean vector $\boldsymbol{\mu}_{1\ell} = (\mu_{1\ell}, \ldots, \mu_{1\ell})$ and a covariance function $\boldsymbol{\Omega}_{1\ell}(\boldsymbol{\phi}_1, \nu)$, $\forall\ \ell = 1, \ldots, b_1$, where $\boldsymbol{\phi}_1 = \{\phi_{11}, \ldots, \phi_{1p^\star}\} \in \mathbb{R}^{p^\star}$ and ν are, respectively, the vector of length parameters and precision parameters of the chosen stationary kernel. Notably, this parameterisation allows for variable-specific length parameters $\phi_{1j}\ \forall\ j = 1, \ldots, p^\star$, where $p^\star \leq p$ is the number of continuous predictors, under which the kernel is still stationary but no longer isotropic.

In the standard BART, since the trees follow a binary structure, each new node is determined by split rules of the form $\{\mathbf{x}^{(j)} \leq c_{\mathbf{x}^{(j)}}\}$ vs. $\{\mathbf{x}^{(j)} > c_{\mathbf{x}^{(j)}}\}$ for continuous predictors, where $c_{\mathbf{x}^{(j)}}$ is a scalar uniformly sampled from the range of a specific covariate $\mathbf{x}^{(j)}$ in the matrix $\mathbf{X}$ of training set predictors. Dummy variables are typically used to represent categorical predictors, which yields rules of the form $\{\mathbf{x}^{(j)} \in d_{\mathbf{x}^{(j)}}\}$ vs. $\{\mathbf{x}^{(j)} \notin d_{\mathbf{x}^{(j)}}\}$, where $d_{\mathbf{x}^{(j)}}$ denotes one of the variable's possible outcome levels.

For a single tree $\mathcal{T}_1$ with b_1 terminal nodes, the model is written as $y_i \mid \mathbf{x}_i \sim \mathrm{N}(g(\mathbf{x}_i; \mathcal{T}_1, \mathbf{G}_1), \tau^{-1})$, where the function $g(\cdot)$ assigns the predicted values $\boldsymbol{\psi}_{1\ell}$ from $\mathcal{GP}_{1\ell}$ to the observations belonging to terminal node ℓ. The description of the tree structure for GP-BART, which generalises the above to allow for rotated splitting rules, is deferred to Section 3.2.1.

Expanding such a model into a sum-of-trees structure is achieved via

$$y_i \mid \mathbf{x}_i \sim \mathrm{N}\left(\sum_{t=1}^{T} g\left(\mathbf{x}_i; \mathcal{T}_t, \mathbf{G}_t\right), \tau^{-1}\right),$$

where the parameters $\mathbf{G}_t = (\{\mu_{t1}, \boldsymbol{\phi}_t, \nu\}, \ldots, \{\mu_{tb_t}, \boldsymbol{\phi}_t, \nu\})$ now characterise the terminal node GPs of each tree $\mathcal{T}_t$, now denoted by $\mathcal{GP}_{t\ell}(\boldsymbol{\mu}_{t\ell}, \boldsymbol{\Omega}_{t\ell}(\boldsymbol{\phi}_t, \nu)), \forall\ \ell = 1, \ldots, b_t$, where $\boldsymbol{\mu}_{t\ell} = (\mu_{t\ell}, \ldots, \mu_{t\ell})$ is again a constant vector, $\boldsymbol{\phi}_t = \{\phi_{t1}, \ldots, \phi_{tp}\} \in$

$\mathbb{R}^p$ is now specific to each tree, in addition to each variable, and g now assigns the predicted values $\boldsymbol{\psi}_{t\ell}$ from $\mathcal{GP}_{t\ell}$. The GP-BART model can be interpreted as a piecewise sum of non-linear GPs whereby each of the T trees will make a small contribution to the overall $\mathbb{E}[y_i \mid \mathbf{x}_i]$, whereas BART can be interpreted as a less flexible piecewise sum of constants. Consequently, GP-BART typically requires fewer trees than the standard BART model.

As in standard BART, we require prior distributions for the tree structure and terminal node parameters; i.e., $(\mathcal{T}_1, \mathbf{G}_1), \ldots, (\mathcal{T}_t, \mathbf{G}_t)$. We assume ν is fixed and select the following shrinkage priors assuming independence between trees and terminal nodes:

$$\begin{aligned} \pi\left((\mathcal{T}_1, \mathbf{G}_1), \ldots, (\mathcal{T}_t, \mathbf{G}_t), \tau\right) &= \pi(\tau) \prod_{t=1}^{T} \pi\left(\mathcal{T}_t, \mathbf{G}_t\right) \\ &= \pi(\tau) \prod_{t=1}^{T} \pi\left(\mathbf{G}_t \mid \mathcal{T}_t\right) \pi\left(\mathcal{T}_t\right), \end{aligned} \tag{3.1}$$

where

$$\pi\left(\mathbf{G}_t \mid \mathcal{T}_t\right) = \pi\left(\boldsymbol{\phi}_t\right) \prod_{\ell=1}^{b_t} \pi\left(\boldsymbol{\psi}_{t\ell} \mid \mu_{t\ell}, \mathcal{T}_t, \boldsymbol{\phi}_t, \nu\right) \pi\left(\mu_{t\ell} \mid \mathcal{T}_t\right). \tag{3.2}$$

We follow Chipman et al. (2010) in our selection of priors for $\mathcal{T}_t$ and τ and adopt data-driven priors for the node-level $\boldsymbol{\mu}_{t\ell}$ in such a way that considerable probability is assigned around the range of the observed $\mathbf{y}$ given the induced prior from the sum of GPs. Associated hyperparameters are omitted from Equations (3.1) and (3.2), for brevity, but we now fully define each prior in turn.

3.2.1 The tree structure

The prior $\pi\left(\mathcal{T}_t\right)$ is specified following the standard setting given by (Chipman et al., 1998), with slight modifications to incorporate the rotated splitting rules. Thus, the tree prior distribution is implicitly defined by a generating stochastic process. In the standard BART algorithm, the tree generation is initialised with a root node. Thereafter, the structure is learned via grow, prune, change, and swap moves. New trees are proposed by growing a new terminal node, removing a pair of terminal nodes, changing the split rule for an internal node, or swapping the split rules for a pair of internal nodes, where the type of move is chosen at

random. Each proposed tree is then accepted or rejected via Metropolis-Hastings (MH); see Chipman et al. (1998) for further details. Notably, the swap move is not incorporated by GP-BART due to computational complexity, as shown by Kapelner and Bleich (2016), and the tendency of GP-BART to yield shallower trees, for which proposing such swap moves would not be feasible.

We also introduce two modified moves, termed "grow-rotate" and "change-rotate", as replacements for the original "grow" and "change" moves, in order to enhance the predictive performance over standard BART. We begin by selecting a pair of covariates j and j' among the set of p possible covariates in $\mathbf{X}$ at random with equal probabilities. The rotated splitting rules are restricted to the case where the covariate j in the selected pair is continuous. Subsequently, an angle θ is sampled with equal probability from a predefined grid of 20 equally spaced values within the interval $[0, \pi]$. To rotate both predictors with respect to θ, it is possible to transform the original coordinate system to $(\mathbf{x}_r^{(j)}, \mathbf{x}_r^{(j')})$ by multiplying $\left(\mathbf{x}^{(j)}, \mathbf{x}^{(j')}\right)$ by the rotation matrix

$$\boldsymbol{\mathcal{R}}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{3.3}$$

Then, within the projected feature space, one of these predictors from the pair $(\mathbf{x}_r^{(j)}, \mathbf{x}_r^{(j')})$ is sampled, again with equal probability. The rotated splitting rules are restricted to the case where the covariate selected from the pair is continuous. A split rule is then selected by sampling a cutpoint from a uniform distribution $c_{\mathbf{x}_r^{(.)}} \sim \text{Uniform}(a_{\mathbf{x}_r^{(.)}}, b_{\mathbf{x}_r^{(.)}})$, where $a_{\mathbf{x}_r^{(.)}}$ and $b_{\mathbf{x}_r^{(.)}}$represent the minimum and maximum values of the transformed selected split variable $\mathbf{x}_r^{(.)}$ within the branch. These rules in the projection space correspond to rotated rules in the original space. If θ is chosen from the set $\boldsymbol{\theta}_0 = \{0, \pi/2, \pi\}$, the rotation direction remains originally axis-aligned, effectively returning to the standard BART splitting rules with a univariate cutpoint as per Section 3.2. Thus, the standard BART moves can be viewed as a specific case of their projected counterparts. Indeed, in cases where axis-aligned splits are sufficient, proposed projections at $\theta \in \boldsymbol{\theta}_0$ tend to be accepted instead of any other θ direction.

As stated, the above applies only when the selected covariate from the pair is continuous. While we omit categorical variables from the GPs, we do allow them

to be used to form splitting rules. If the sampled covariate within the pair is categorical, the angle θ is irrelevant; we assume $\theta \in \boldsymbol{\theta}_0$ and need not sample it. When the selected covariate is binary, the splitting rule is simply a partition of its levels. However, we make a further modification when the sampled pair contains a nominal variable. We use the reparameterisation suggested by Wright and König to identify optimal cutpoints for such predictors in treed methods (see Wright and Ziegler (2017) and Wright and König (2019) for more details). This ultimately leads to split rules of the form $\{\mathbf{x}^{(\cdot)} \in \{\mathcal{S}\}\}$ vs. $\{\mathbf{x}^{(\cdot)} \notin \{\mathcal{S}\}\}$, where $\{\mathcal{S}\}$ denotes a subset of the levels of the given covariate and also allows for continuous covariates to be rotated with respect to nominal ones. Otherwise, GP-BART utilises the default moves from BART outlined in Section 3.2 when $p = 1$.

To summarise, the prior for T_t can be divided into five aspects; namely, (i) the distribution on the pair of candidate splitting variables at each interior node, (ii) the distribution on the selected splitting variable, conditioned on the chosen pair, (iii) the distribution on the rotation angle θ, given the selected variable, and (iv) the distribution on the splitting cutpoint, conditional on the chosen pair, variable, and angle. For these four aspects, the relevant priors coincide with the equiprobable discrete proposal distributions described above. Furthermore, (v) the prior probability of an individual node at depth $d = 0, 1, 2, \ldots$ being non-terminal is controlled by the hyperparameters α and β through

$$\Pr\left(\text{non-terminal node}\right) \propto \alpha\left(1+d\right)^{-\beta}, \;\; \alpha \in (0,1), \;\; \beta \in [0,\infty). \tag{3.4}$$

The tree prior $\pi\left(\mathcal{T}_t\right)$ is then given by a product of the probabilities of each node, since Equation (3.4) assumes independence between nodes. Following some evaluation of alternative parameterisations, we fix the default values $\alpha = 0.95$ and $\beta = 2$, as per the standard BART (Chipman et al., 2010). The proposal distribution for a new tree is described by a discrete sample of the possible grow-rotate, change-rotate, and prune moves, with respective probabilities of 0.3, 0.4, and 0.3. These probabilities align with those associated with the standard moves in `bartMachine` (Kapelner and Bleich, 2016), whereby the 0.1 probability of a swap move in the original BART (Chipman et al., 2010) is equally reapportioned to the grow-rotate and change-rotate moves, without modifying the prior probability of the prune move.

Figure 3.1 summarises the main idea of our proposed statistical model, highlighting the modified terminal node priors and the rotated splitting rules, via four examples of regression trees within the ensemble. Here, there are two continuous predictors, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, and $\mathbf{x}^{(3)}$ is categorical, with the sets $\{\mathcal{S}_1\}$ and $\{\mathcal{S}_2\}$ being subset of its levels. Notably, some split rules from trees $\mathcal{T}_2$ and $\mathcal{T}_3$ are obtained by projecting a randomly sampled non-parallel axis direction θ onto the pair $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, resulting in rotated splitting rules.

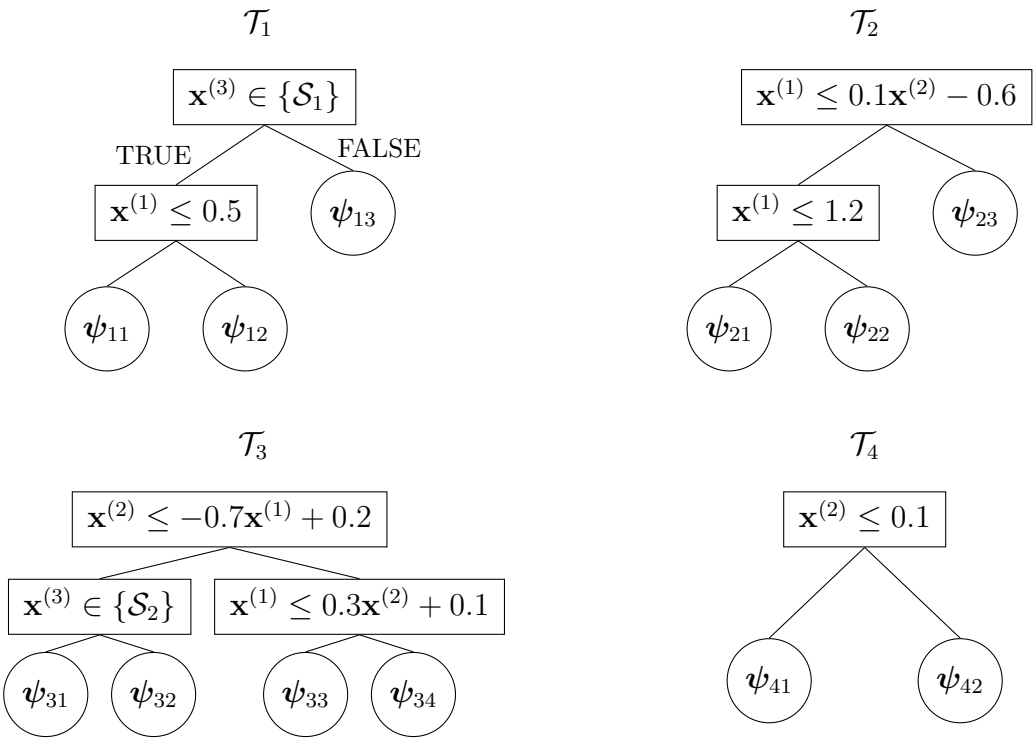


Figure 3.1: Graphical representation of four example trees from a GP-BART model. The splitting rules in each tree can take the form of a univariate cut-point for continuous covariates (subject to $\theta \in \boldsymbol{\theta}_0$), a subset of factor levels for categorical covariates, or rotated split rules obtained by random projections of a pair of covariates, provided the selected covariate from the pair is continuous. Gaussian process priors are assumed for the predicted values for each terminal node in each tree, such that $\boldsymbol{\psi}_{t\ell} \sim \mathcal{GP}_{t\ell}$ *a priori*.

3.2.2 The prior on the Gaussian processes

The main contribution of the GP-BART model is to define

$$\boldsymbol{\psi}_{t\ell} \mid \mathcal{T}_t, \mu_{t\ell}, \boldsymbol{\phi}_t, \nu \sim \text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, \boldsymbol{\Omega}_{t\ell}\right) \tag{3.5}$$

as a GP prior over the set of $n_{t\ell}$ observations belonging to terminal node ℓ of tree $\mathcal{T}_t$, where $\mathbf{1}_{n_{t\ell}}$ is a vector of ones of length $n_{t\ell}$, such that the mean vector is constant. Here, $\boldsymbol{\Omega}_{t\ell} \in \mathbb{R}^{n_{t\ell}} \times \mathbb{R}^{n_{t\ell}}$ is specified as a node-specific, stationary, anisotropic matrix of exponentiated-quadratic covariance terms, with its (i,k)-th element given by

$$\nu^{-1} \exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\frac{\left(x_i^{(j)} - x_k^{(j)}\right)^2}{\phi_{tj}^2}\right\}. \tag{3.6}$$

We normalise all predictors to the $[0,1]$ range to improve the numerical stability of the kernel. Notably, the trees themselves are unaffected by this, as the rules governing their structure are invariant to monotone transformations. We set $\mu_{t\ell} \mid \mathcal{T}_t, \sim \text{N}\left(\mu_\mu, \tau_\mu^{-1}\right)$ to exploit conjugacy and enable all $\mu_{t\ell}$ parameters to be marginalised out. Hence, Equation (3.5) can be redefined as

$$\boldsymbol{\psi}_{t\ell} \mid \mathcal{T}_t, \boldsymbol{\phi}_t, \nu, \mu_\mu, \tau_\mu \sim \text{MVN}\left(\mu_\mu \mathbf{1}_{n_{t\ell}}, \tau_\mu^{-1}\mathbf{1}_{n_{t\ell}}\mathbf{1}_{n_{t\ell}}^\top + \boldsymbol{\Omega}_{t\ell}\right),$$

in order to encourage better mixing. We adopt this likelihood formulation throughout and provide further details in Appendix 3.A.

Chipman et al. (2010) showed that the induced prior distribution on $\mathbb{E}[y_i \mid \mathbf{x}_i]$ over all T trees in a BART model allows for some expert knowledge to be incorporated about the contribution of each tree which can help to guide the choices of hyperparameter values. However, the presence of the GP priors on $\boldsymbol{\psi}_{t\ell}$ in GP-BART yields a different induced prior which we write as

$$\mathbb{E}\left[y_i \mid \mathbf{x}_i\right] \sim \text{N}\left(T\mu_\mu, T\left(\nu^{-1} + \tau_\mu^{-1}\right)\right).$$

Following the Chipman et al. approach, the key idea is to select the hyperparameters such that $\mathbb{E}[y_i \mid \mathbf{x}_i]$ is between $y_{\min}$ and $y_{\max}$ with high probability. The confidence interval for $\mathbb{E}[y_i \mid \mathbf{x}_i]$, $\forall\, i = 1, \ldots, n$, has boundaries

$$\begin{cases} T\mu_\mu - k\sqrt{T}\left(\nu^{-1} + \tau_\mu^{-1}\right)^{1/2} = y_{\min} \\ T\mu_\mu + k\sqrt{T}\left(\nu^{-1} + \tau_\mu^{-1}\right)^{1/2} = y_{\max} \end{cases}$$

for a chosen k. We adopt $k = 2$, which represents an approximate 95% confidence interval. Following Chipman et al., we re-scale $\mathbf{y}$ such that $y_{\min} = -0.5$ and $y_{\max} = 0.5$, set $\mu_\mu = 0$ and hence set the precision parameters to

$$\nu = \tau_\mu = 8k^2T,$$

in order to balance the contribution of both parameters.

Though ν and τ_μ are both referred to as precision parameters, their roles and interpretations differ, with ν and τ_μ – both of which are fixed rather than estimated – being the parameters that control the precision of the GPs and the $\mu_{t\ell}$ parameters, respectively. As we increase the number of trees T, the scale ν^{-1} of each GP decreases, regularising the model by setting the contribution of each GP to be small. Likewise, the precision of the $\mu_{t\ell}$ parameter is proportional to the number of trees, shrinking the mean of each terminal node as more tree components are added into the model. Setting both parameters in this way reduces the chance of only one single tree dominating the model.

3.2.2.1 The prior on the length parameter

As shown in Equation (3.6), $\boldsymbol{\phi}_t$ controls the rate of decay with respect to the $\mathcal{L}_2$ distances between pairs of design points, such that larger values of ϕ_{tj} will quickly decrease the contribution of variables which are uncorrelated with the true generation function $f(\mathbf{x}_i)$. Thus, to enable the use of automatic relevance determination (ARD) over the variables used in the GPs while balancing computational considerations, we derive a discrete prior for the length parameter ϕ_{tj} for a given tree t and covariate j from a mixture of gamma distributions:

$$\begin{aligned} \kappa \times \mathrm{Ga}\left(a_{\phi_1} = 3, d_{\phi_1} = 2.5\right) + \\ (1-\kappa) \times \mathrm{Ga}\left(a_{\phi_2} = 5000, d_{\phi_2} = 100\right), \end{aligned} \tag{3.7}$$

where $\mathrm{Ga}(a, d)$ denotes a gamma distribution with expectation a/d. The two components govern smaller and larger values of ϕ_{tj}, respectively, and we set the mixture weight κ to 0.3 throughout.

Ultimately, we define a discrete prior for $\pi(\phi_{tj})$, with support given by $\mathbf{S}_\phi = \{0.1, 0.5, 1, 2, 3, 4, 50\}$ in order to reflect the high-probability regions of the mix-

ture in Equation (3.7) and the fact that the precise magnitude of ϕ_{tj} is only important for smaller values. Furthermore, the prior probabilities are specified to be proportional to $d_\phi(k)$, where $d_\phi(k)$ is the density of the mixture of gamma distributions in Equation (3.7) evaluated at $k \in \mathbf{S}_\phi$. This leads to probabilities of $\Pr(\phi_{tj} = k) = (0.022, 0.206, 0.236, 0.035, 0.014, 0.002, 0.485)$, which reflect the fact that, *a priori*, we expect each variable to have an equal chance of contributing meaningfully to the GPs. The ϕ_{tj} sampling processes is also done using MH, with the proposal distribution for new parameters given by an equiprobable discrete distribution which reflects the support of our induced discrete prior $\pi(\phi_{tj})$ — i.e., each value in $\mathbf{S}_\phi$ is sampled with equal probability — and helps to avoid spurious length parameter values.

The aforementioned normalisation of each predictor in $\mathbf{X}$ also aids the elicitation of this prior, by minimising the range of ϕ_{tj} and ensuring all covariates are on the same scale. Furthermore, the discrete proposal reduces the computational burden, as we can partially pre-compute all possible covariance functions. We calculate the fraction in the exponent of Equation (3.6) for each length parameter value in $\mathbf{S}_\phi$, using all n observations of the continuous covariates, and thereafter obtain $\boldsymbol{\Omega}_{t\ell}(\boldsymbol{\phi}_t, \nu)$ by appropriately utilising the quantities relevant to the sampled $\phi_{t1}, \ldots, \phi_{tp^\star}$ values and subset of observations $\mathbf{X}_{(t\ell)}$ belonging to the corresponding terminal node.

3.2.3 The prior on the residual precision

A conjugate gamma distribution $\tau \sim \text{Ga}(a_\tau, d_\tau)$ is assumed for the residual precision parameter. To select the hyperparameters, we follow Chipman et al. (2010) in setting the shape a_τ and rate d_τ such that $\Pr(\tau \geq \hat{\tau}_{OLS}) = \eta_\tau$, where η_τ is a high-probability value (we typically use $\eta_\tau = 0.9$) and $\hat{\tau}_{OLS}$ is the precision calculated from an ordinary linear regression of $\mathbf{y}$ against the same set of predictors $\mathbf{X}$. The intuition behind this estimation strategy comes from the idea that, given the non-linearity of the GP and the piecewise-constant component from BART, we can be optimistic that the precision of the model is greater than that of a linear model.

3.3 Computational algorithms for inference and prediction

Given the observed $\mathbf{y}$, the posterior distribution for the trees and their parameters is given by

$$\pi\left(\left(\mathcal{T}_1, \mathbf{G}_1\right), \ldots, \left(\mathcal{T}_T, \mathbf{G}_T\right), \tau \mid \mathbf{y}\right). \tag{3.8}$$

We define the notation of a generic set $\mathcal{M}_{-t}$ as the the set of all $\mathcal{M}_1, \ldots, \mathcal{M}_T$ elements except $\mathcal{M}_t$, such that $\mathcal{T}_{-t}$ corresponds to the set of $T-1$ trees except $\mathcal{T}_t$ with respective terminal node parameters $\mathbf{G}_{-t}$. The key feature necessary to sample from Equation (3.8) is the "Bayesian backfitting" algorithm of Hastie and Tibshirani (2000), which enables iterative sampling of the t-th tree and its parameters. Hastie and Tibshirani showed that the distribution $\pi\left(\mathcal{T}_t, \mathbf{G}_t \mid \mathcal{T}_{-t}, \mathbf{G}_{-t}, \tau, \mathbf{y}\right))$ can be rewritten in terms of the partial residuals

$$\mathbf{R}_t = \left(\mathbf{r}_{t1}, \ldots, \mathbf{r}_{tb_t}\right) \equiv \mathbf{y} - \sum_{r \neq t}^{T} g\left(\mathbf{X}; \mathcal{T}_r, \mathbf{G}_r\right). \tag{3.9}$$

The general structure of the sampler is thus given by:

$$\begin{aligned}
1: &\quad \mathcal{T}_1 \mid \mathbf{R}_1, \boldsymbol{\phi}_1, \nu, \tau_\mu, \tau \\
2: &\quad \boldsymbol{\psi}_{11}, \ldots, \boldsymbol{\psi}_{1b_1} \mid \mathcal{T}_1, \mathbf{R}_1, \boldsymbol{\phi}_1, \nu, \tau_\mu, \tau \\
3: &\quad \boldsymbol{\phi}_1 \mid \mathcal{T}_1, \mathbf{R}_1, \nu, \tau_\mu, \tau \\
&\quad \vdots \\
3T-2: &\quad \mathcal{T}_T \mid \mathbf{R}_T, \boldsymbol{\phi}_T, \nu, \tau_\mu, \tau \\
3T-1: &\quad \boldsymbol{\psi}_{T1}, \ldots, \boldsymbol{\psi}_{Tb_T} \mid \mathcal{T}_T, \mathbf{R}_T, \boldsymbol{\phi}_T, \nu, \tau_\mu, \tau \\
3T: &\quad \boldsymbol{\phi}_T \mid \mathcal{T}_T, \mathbf{R}_T, \nu, \tau_\mu, \tau \\
3T+1: &\quad \tau \mid \left(\mathcal{T}_1, \mathbf{G}_1\right) \ldots, \left(\mathcal{T}_T, \mathbf{G}_T\right), a_\tau, d_\tau, \mathbf{y}.
\end{aligned}$$

The algorithm is initialized with T stumps (i.e., trees with a single root node), with all mean parameters $\mu_{t1} = 0$ and all length parameters ϕ_{tj} sampled from the discrete proposal distribution described in Section 3.2.2.1. Additionally, the residual precision parameter τ is sampled from its prior distribution. For stumps, only the grow-rotate move is proposed. Thereafter, once trees have reached sufficient depth $d = 1$, new trees $\mathcal{T}_t^\star$ are sequentially proposed by randomly selecting

one of the three available moves: grow-rotate, change-rotate, and prune, and then accepted or rejected according via MH.

Though these moves modify the tree depth, $\mathbf{G}_t$ only changes dimension with respect to the means $\mu_{t1}, \ldots, \mu_{tb_t}$, since ν is fixed and $\boldsymbol{\phi}_t$ is specified at the tree level. Consequently, this does not affect the sampling of $\mathcal{T}_t$, since all $\mu_{t\ell}$ parameters are marginalised out, thereby yielding a tractable tree posterior proportional to $\pi(\mathcal{T}_t)\,\pi(\boldsymbol{\phi}_t)\,\pi(\mathbf{R}_t \mid \mathcal{T}_t, \boldsymbol{\phi}_t, \nu, \tau_\mu, \tau)$, which does not depend on any varying-dimensional parameters at the terminal node level. The predicted values in each terminal node are updated by a Gibbs sampling scheme, with the associated full conditional distribution given by

$$\boldsymbol{\psi}_{t\ell} \mid \ldots \sim \text{MVN}\left(\boldsymbol{\mu}_{\mathcal{GP}_{t\ell}}, \boldsymbol{\Sigma}_{\mathcal{GP}_{t\ell}}\right), \tag{3.10}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\mathcal{GP}_{t\ell}} &= \boldsymbol{\Lambda}_{t\ell}^{\top}\left(\tau^{-1}\boldsymbol{\mathcal{I}}_{n_{t\ell}} + \boldsymbol{\Lambda}_{t\ell}\right)^{-1}\mathbf{r}_{t\ell}, \\
\boldsymbol{\Sigma}_{\mathcal{GP}_{t\ell}} &= \boldsymbol{\Lambda}_{t\ell} - \boldsymbol{\Lambda}_{t\ell}^{\top}\left(\tau^{-1}\boldsymbol{\mathcal{I}}_{n_{t\ell}} + \boldsymbol{\Lambda}_{t\ell}\right)^{-1}\boldsymbol{\Lambda}_{t\ell},
\end{aligned}$$

with $\boldsymbol{\Lambda}_{t\ell} = \tau_\mu^{-1}\mathbf{1}_{n_{t\ell}}\mathbf{1}_{n_{t\ell}}^{\top} + \boldsymbol{\Omega}_{t\ell}$ and $\boldsymbol{\mathcal{I}}_{n_{t\ell}}$ being an identity matrix of the indicated dimension.

Lastly, we sample the length parameters $\phi_{tj}\ \forall\,(j = 1, \ldots, p^\star, t = 1, \ldots, T)$ from their discrete proposal distribution using MH steps. Once all T trees are updated, the precision parameter is sampled using a Gibbs step, with the full conditional given by

$$\tau \mid \ldots \sim \text{Ga}\left(\frac{n}{2} + a_\tau, \frac{1}{2}\left(\mathbf{y} - \hat{\mathbf{y}}\right)^{\top}\left(\mathbf{y} - \hat{\mathbf{y}}\right)\right) + d_\tau\Big), \tag{3.11}$$

where $\hat{\mathbf{y}} \equiv \sum_{t=1}^{T} g(\mathbf{X}; \mathcal{T}_t, \mathbf{G}_t)$ represents the sum of the predictions $\boldsymbol{\psi}_{t\ell}$ across all terminal nodes from all sampled trees.

3.3.1 Algorithm specifications and initialisation

We set the number of trees T to have a default value of 20, since we require fewer trees than BART due to the inherent non-linearity of the GPs and achieved reasonable predictive performance in various scenarios demonstrated in Sections

3.4 and 3.5 using this value. Alternatively, this quantity could be selected via cross-validation, though the computational cost of doing so may be prohibitive.

Employing the rotated splitting rules or using the standard moves from BART is also a setting of the model that can be toggled by the user, as well as which variables are included in the GPs themselves. All variables are allowed to form splitting rules, since it improves the model's prediction in general, especially for spatial data. If the rotated splits are deemed unnecessary, the sampler will not accept them and favour splits with $\theta \in \boldsymbol{\theta}_0$. By default, if there is no strict prior knowledge about the covariates, GP-BART includes all continuous variables in the GPs. Though a more parsimonious model could be achieved if the variables used in the GPs are merely a subset of those used to construct the trees, we do not consider this further here.

We present the full structure of the GP-BART sampler in Algorithm 3.1, where the matrix of covariates $\mathbf{X}$ and response vector $\mathbf{y}$ from the training set enter as inputs. Trees, partial residuals, and hyperparameters are then initialised. For each MCMC sample, a proposed tree $\mathcal{T}_t^\star$ is accepted, if it is valid and contains no empty terminal nodes, with probability $\gamma^\star(\mathcal{T}_t, \mathcal{T}_t^\star)$. The novel aspects of the tree prior we introduce under GP-BART (i.e., priors over the pair of candidate splitting variables and the rotation angle θ) cancel out in the MH acceptance ratio. Consequently, the ratio of priors $\pi\left(\mathcal{T}_t^\star\right)/\pi\left(\mathcal{T}_t\right)$ in $\gamma^\star\left(\mathcal{T}_t, \mathcal{T}^\star\right)$ and the transition probabilities $q(\cdot)$ for all moves remain unchanged from the formulations given by Linero and Yang (2018). The remaining parameters are sampled using Equations (3.10)–(3.11).

A standard number of iterations $N_{\text{MCMC}} = 3500$, of which the first $N_{\text{burn}} = 1500$ are discarded, was found to yield a sufficient number of samples to reliably characterise the posterior in all applications herein. This was verified through examination of the convergence of posterior samples of τ. Though the algorithm is computationally onerous given the matrix inversions associated with the use of GPs, we stress that such operations are of the order $\mathcal{O}(n_{t\ell}^3)$ within a given terminal node, rather than $\mathcal{O}(n^3)$ as they would be under a single GP. Further details of the computational performance of our algorithm in the context of a simulation study are deferred to Section 3.4.2.1.

Algorithm 3.1: GP-BART sampling algorithm.

Input: $\mathbf{X}$, $\mathbf{y}$, T, N_{MCMC}, N_{burn}, and all hyperparameters of the priors.

Initialise: T tree stumps with $\mu_{t1} = 0 \,\forall\, t$, $\phi_{tj} \,\forall\, (t, j)$ drawn with equal probability from $\mathbf{S}_\phi$, and τ drawn from its $\text{Ga}(a_\tau, d_\tau)$ prior.

for *iterations* m *from* 1 *to* N_{MCMC} **do**
  for *trees* t *from* 1 *to* T **do**
    Calculate the partial residuals $\mathbf{R}_t$ via Equation (3.9);
    Propose a new tree $\mathcal{T}_t^\star$ by a grow-rotate, change-rotate, or prune move;
    Accept and update $\mathcal{T}_t = \mathcal{T}_t^\star$ with probability

$$\gamma^\star\left(\mathcal{T}_t, \mathcal{T}_t^\star\right) = \min\left\{1, \frac{\pi\Big(\mathbf{R}_t \mid \mathcal{T}_t^\star, \boldsymbol{\phi}_t, \nu, \tau_\mu, \tau\Big)\pi\Big(\mathcal{T}_t^\star\Big)q\Big(\mathcal{T}_t^\star \to \mathcal{T}_t\Big)}{\pi\Big(\mathbf{R}_t \mid \mathcal{T}_t, \boldsymbol{\phi}_t, \nu, \tau_\mu, \tau\Big)\pi\Big(\mathcal{T}_t\Big)q\Big(\mathcal{T}_t \to \mathcal{T}_t^\star\Big)}\right\}.$$

    for *terminal nodes* ℓ *from* 1 *to* b_t **do**
      Update $\boldsymbol{\psi}_{t\ell}$ via Equation (3.10).
    end
    for *continuous predictors* j *from* 1 *to* $p^\star$ *used in the GPs* **do**
      Update ϕ_{tj} using MH.
    end
  end
  Update τ via Equation (3.11).
end

Output: Samples from $\pi\left((\mathcal{T}_1, \mathbf{G}_1), \ldots, (\mathcal{T}_T, \mathbf{G}_T), \tau \mid \mathbf{y}\right)$.

3.3.2 Prediction in GP-BART

The trees in GP-BART models can provide out-of-sample predictions for a set of $n^\star$ new observations $\mathbf{X}^\star$. For a given terminal node ℓ in tree $\mathcal{T}_t$ for a particular MCMC sample, the joint posterior distribution of the node-level training predictions and the node-level test predictions is given by

$$\begin{pmatrix} \boldsymbol{\psi}_{t\ell} \\ \boldsymbol{\psi}_{t\ell}^\star \end{pmatrix} \Bigg| \ \ldots \sim \text{MVN}\left(\begin{bmatrix} \mathbf{0}_{n_{t\ell}} \\ \mathbf{0}_{n_{t\ell}^\star} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_{t\ell} & \boldsymbol{\Lambda}_{t\ell}^\star \\ \boldsymbol{\Lambda}_{t\ell}^{\star\top} & \boldsymbol{\Lambda}_{t\ell}^{\star\star} \end{bmatrix}\right),$$

with $\boldsymbol{\Lambda}_{t\ell}^\star \in \mathbb{R}^{n_{t\ell}^\star} \times \mathbb{R}^{n_{t\ell}}$ and $\boldsymbol{\Lambda}_{t\ell}^{\star\star} \in \mathbb{R}^{n_{t\ell}^\star} \times \mathbb{R}^{n_{t\ell}^\star}$. Here, $n_{t\ell}$ and $n_{t\ell}^\star$ denote the number of observations assigned to terminal node ℓ of tree $\mathcal{T}_t$ for the training samples and

new data, respectively. This posterior predictive distribution can be conditioned with respect to $\boldsymbol{\psi}_{t\ell}$ to yield

$$\boldsymbol{\psi}^{\star}_{t\ell} \,|\, \boldsymbol{\psi}_{t\ell}, \mathbf{X}_{(t\ell)}, \mathbf{X}^{\star}_{(t\ell)}, \ldots \sim \text{MVN}\left(\boldsymbol{\mu}_{\mathcal{GP}^{\star}_{t\ell}}, \boldsymbol{\Sigma}_{\mathcal{GP}^{\star}_{t\ell}}\right),$$

where $\boldsymbol{\mu}_{\mathcal{GP}^{\star}_{t\ell}} = \boldsymbol{\Lambda}^{\star\top}_{t\ell}\boldsymbol{\Lambda}^{-1}_{t\ell}\boldsymbol{\psi}_{t\ell}$ and $\boldsymbol{\Sigma}_{\mathcal{GP}^{\star}_{t\ell}} = \boldsymbol{\Lambda}^{\star\star}_{t\ell} - \boldsymbol{\Lambda}^{\star\top}_{t\ell}\boldsymbol{\Lambda}^{-1}_{t\ell}\boldsymbol{\Lambda}^{\star}_{t\ell}$.

Ultimately, the function $g^{\star}$ assigns the vector $\boldsymbol{\mu}_{\mathcal{GP}^{\star}_{t\ell}} = \mathbb{E}(\boldsymbol{\psi}^{\star}_{t\ell}| \ldots)$ to the associated new observations $\mathbf{X}^{\star}_{(t\ell)}$ on a *per-iteration* basis, such that the estimates from GP-BART are given by

$$\hat{\mathbf{y}}^{\star(m)} = \text{N}\left(\sum_{t=1}^{T} g^{\star}\left(\mathbf{X}^{\star}; \mathcal{T}^{(m)}_{t}, \mathbf{G}^{(m)}_{t}\right), \hat{\tau}^{-1(m)}\boldsymbol{\mathcal{I}}_{n^{\star}}\right), \tag{3.12}$$

where m indexes the draws from the posterior distribution after the burn-in iterations. The overall prediction $\bar{y}^{\star}_{i}$ for a new observation $\mathbf{x}^{\star}_{i}$ is then given by the average of the estimates $\hat{y}^{\star(1)}_{i}, \ldots, \hat{y}^{\star(M)}_{i}$; i.e., $\bar{y}^{\star}_{i} = \frac{1}{M}\sum_{m=1}^{M}\hat{y}^{\star(m)}_{i}$. Posterior samples from Equation (3.12) can also be used to quantify the uncertainty in the predictions. For instance, with some large number Q of draws *per posterior sample*, the endpoints of a $(1-\alpha)\%$ prediction interval for a predicted value $\bar{y}^{\star}_{i}$ can be obtained from the upper and lower $\alpha/2$ quantiles of $(\hat{y}^{\star(11)}_{i}, \ldots, \hat{y}^{\star(M1)}_{i}), \ldots, (\hat{y}^{\star(1q)}_{i}, \ldots, \hat{y}^{\star(MQ)}_{i})$.

3.4 Simulation studies

In this Section, we present simulation studies to evaluate the performance of GP-BART from several different perspectives. In Section 3.4.1 we primarily aim to assess the efficacy of incorporating the rotated splitting rules and the GPs themselves for data with explicit spatial components, whereas in Section 3.4.2 we first aim to assess the ARD associated with the equiprobable discrete prior on the tree-varying, variable-specific length parameters ϕ_{tj} described in Section 3.2.2.1. An evaluation of the computational burden is also provided in Section 3.4.2.1.

3.4.1 Benchmarking experiments

In these experiments, the simulated data are composed by a summation of trees with two terminal nodes, built using the variables $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$. These covariates are simulated such that each predictor is generated from a uniform grid

between -10 and 10. The values associated with each terminal node follow a multivariate normal distribution with specific mean and covariance parameters. We generate the response variable via

$$\begin{aligned}\mathbf{y} = &\left[(\boldsymbol{\mu}_{11} + \mathbf{s}_{11})_{\mathbb{1}\left(\mathbf{x}^{(1)} \leq \mathbf{x}^{(2)}\right)} + (\boldsymbol{\mu}_{12} + \mathbf{s}_{12})_{\mathbb{1}\left(\mathbf{x}^{(1)} > \mathbf{x}^{(2)}\right)}\right] \\ &+ \left[(\boldsymbol{\mu}_{21} + \mathbf{s}_{21})_{\mathbb{1}\left(\mathbf{x}^{(1)} \leq -\mathbf{x}^{(2)}\right)} + (\boldsymbol{\mu}_{22} + \mathbf{s}_{22})_{\mathbb{1}\left(\mathbf{x}^{(1)} > -\mathbf{x}^{(2)}\right)}\right] \\ &+ \left[(\boldsymbol{\mu}_{31} + \mathbf{s}_{31})_{\mathbb{1}\left(\mathbf{x}^{(1)} \leq 0\right)} + (\boldsymbol{\mu}_{32} + \mathbf{s}_{32})_{\mathbb{1}\left(\mathbf{x}^{(1)} > 0\right)}\right] + \boldsymbol{\varepsilon},\end{aligned} \tag{3.13}$$

with number of trees $T = 3$, each with two terminal nodes. The node-specific mean parameters $\boldsymbol{\mu}_{t\ell}$ are all constant vectors of the form $(\mu_{t\ell}, \ldots, \mu_{t\ell})$, with respective values given by $\mu_{11} = -10$, $\mu_{21} = 0$, $\mu_{31} = 10$, $\mu_{12} = 5$, $\mu_{22} = 20$, and $\mu_{32} = -15$. Multivariate normal spatial terms $\mathbf{s}_{t\ell} \sim \text{MVN}(\mathbf{0}_{n_{t\ell}}, \boldsymbol{\Omega}_{t\ell}(\boldsymbol{\phi}_t = \mathbf{3}_{n_{t\ell}}, \nu = 0.1))$ are added within each terminal node. Residual noise terms $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}_n, \tau^{-1}\boldsymbol{\mathcal{I}}_n)$ are also added. Results obtained with residual precision values of $\tau = \{10, 1, 0.1, 0.01\}$ lead to similar conclusions in that GP-BART consistently shows the best performance in terms of prediction accuracy and uncertainty calibration. For brevity, we show the data and results for $\tau = 10$ here only and defer the other results to Appendix 3.B. Figure 3.2 shows the simulated data surfaces for data sets of size $n = \{100, 500, 1000\}$, respectively, highlighting the different partitioning behaviour and smoothness within each data set.

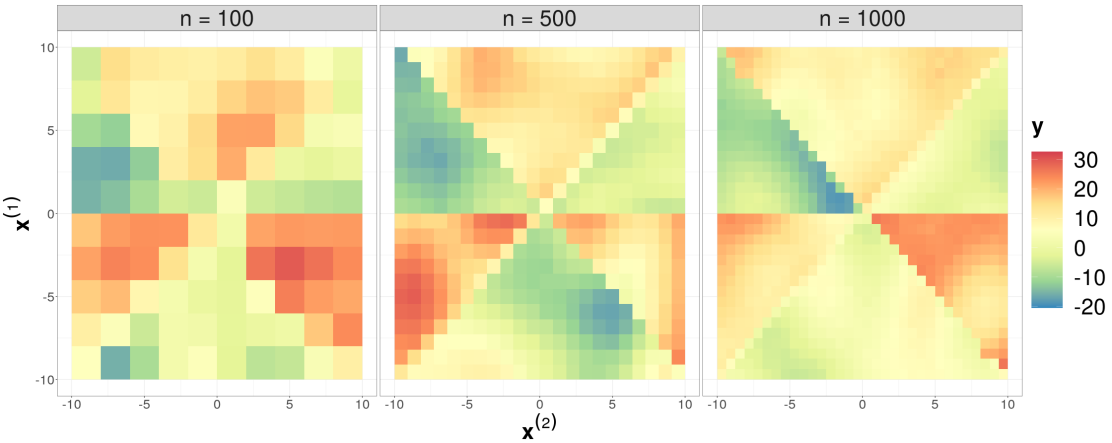


Figure 3.2: Simulated data with $n = \{100, 500, 1000\}$ observations, respectively.

We compare the performance of our GP-BART model to other tree-based methods, namely BART (Chipman et al., 2010), SoftBART (Linero and Yang, 2018), and

tGP (Gramacy and Lee, 2008), as well as the universal kriging model (Cressie, 2015) and latent Gaussian models using integrated nested Laplace approximations (INLA; Lindgren and Rue, 2015). We evaluate the results using 5 repetitions of 5-fold cross-validation; each fold is treated as a test set and prediction accuracy and uncertainty calibration are quantified using the root-mean-square error (RMSE) and the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007), respectively, over all folds within a given repetition.

The models are fitted using the R packages `BART` (Sparapani et al., 2021), `SoftBart` (Linero, 2022b), `tgp` (Gramacy and Taddy, 2010), `fields` (Nychka et al., 2021), and `INLA` (Lindgren and Rue, 2015), with their default settings. All hyperparameters for the GP-BART model were specified using their default values and settings previously described in Sections 3.2 and 3.3. To qualitatively compare the methods, we analyse the prediction surface generated by each algorithm for the data sets of size $n = \{100, 500, 1000\}$, shown in Figure 3.2, using predictions over the test sets in the repeated 5-fold setting. The corresponding plots are provided in Figures 3.3, 3.4, and 3.5. In each case, results from one randomly chosen repetition of the repeated 5-fold cross-validation are used to construct the plots.

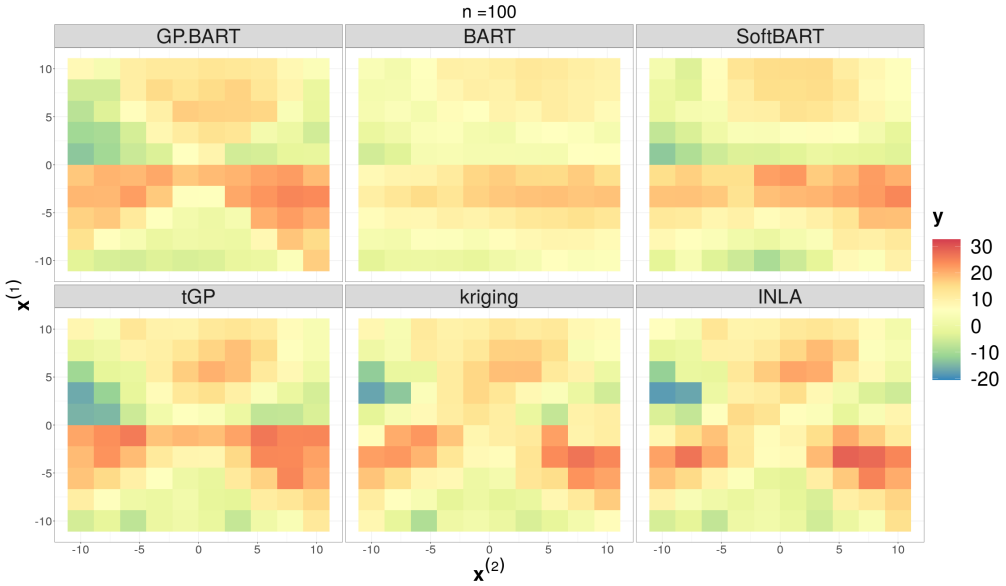


Figure 3.3: Predicted surfaces for the simulated scenario with $n = 100$ observations from the first panel of Figure 3.2 using different methods over one randomly chosen test repetition.

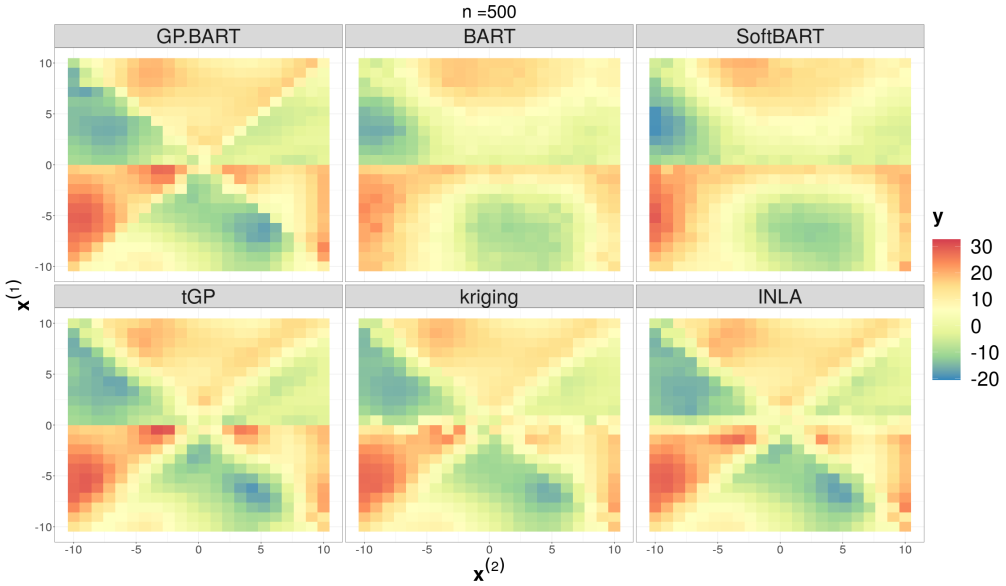


Figure 3.4: Predicted surfaces for the simulated scenario with $n = 500$ observations from the second panel of Figure 3.2 using different methods over one randomly chosen test repetition.

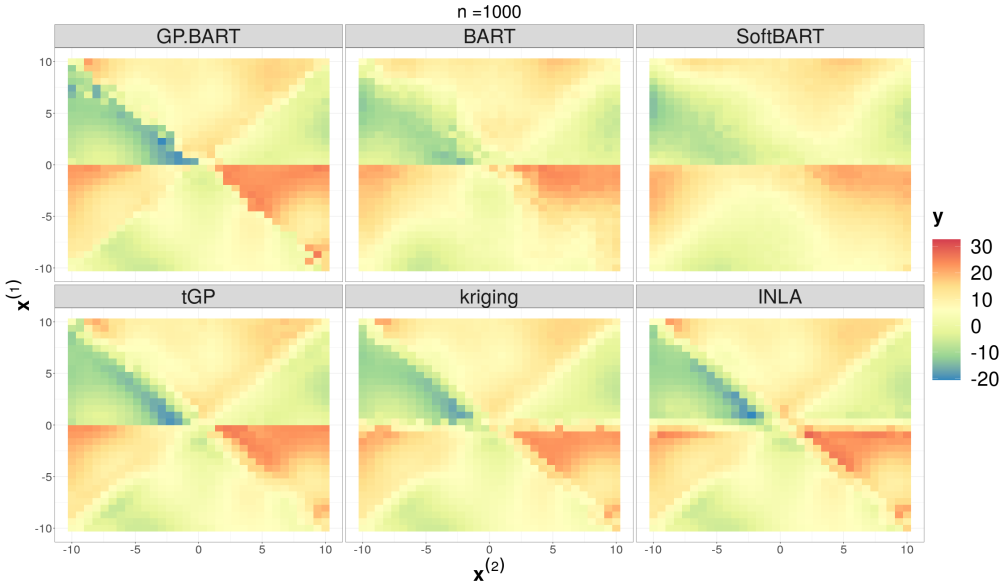


Figure 3.5: Predicted surfaces for the simulated scenario with $n = 1000$ observations from the third panel of Figure 3.2 using different methods over one randomly chosen test repetition.

Though the provided plots indicate clear differences between each model type, each model's behaviour is similar across the sample sizes. GP-BART's prediction surfaces appear most similar to the original data shown in Figure 3.2 in each case. Indeed, GP-BART successfully identifies diagonal partitions due to its rotated splits, while BART, SoftBART, and tGP only produce splits parallel to the axes. Though BART and SoftBART uncover differences among the terminal node regions nonetheless, their predictions are less accurate than their competitors by virtue of spatial dependence not being explicitly accounted for by these two methods. In addition, GP-BART can produce smoother surfaces than BART, as the nature of the original algorithm inherently involves the summation of stepwise-constant functions. The tGP, kriging, and INLA predictions capture the spatial features well, but their failure to identify the partitions results in blurred prediction surfaces in areas where the data splits. Therefore, we emphasise that the proposed model takes advantage of the benefits of rotated splits, explicitly defined spatial dependence assumptions, and the inherent smoothness from the GPs.

A quantitative comparison is shown in the boxplots in Figure 3.6, which reflect the previous qualitative interpretations. Here, GP-BART presents substantially lower RMSE than its competitors, particularly for smaller n. We assess uncertainty calibration by examining boxplots of CRPS scores in Figure 3.7. These results show that the GP-BART model presents the lowest CRPS values among all methods. Thus, considering both metrics jointly, GP-BART's performance in terms of prediction accuracy and uncertainty quantification is superior to the other models considered.

To highlight the effect of the proposed moves and the use of GPs over the terminal nodes, four different, restricted versions of GP-BART are compared:

(**A**) without any rotated moves or GPs (i.e., the standard BART model);

(**B**) without GPs, but with the new rotated 'grow' and 'change' moves;

(**C**) without the new rotated moves, but with GPs;

(**D**) the standard GP-BART with both rotated splitting rules and GPs.

We defer the results for the other sample sizes, which lead to similar conclusions, to Appendix 3.C, along with an evaluation of the acceptance rates for the tree-proposal moves under version (**D**), and consider only the $n = 500$ setting here, for brevity. This comparison is summarised in Figures 3.8 and 3.9, in which the letters above are used to distinguish the model versions. As before, results based on one randomly chosen repetition of the 5-fold cross-validation are used to construct Figure 3.8.

The prediction surface (**A**) in Figure 3.8 suggests BART cannot adequately capture different behaviours in the terminal node regions due to the lack of smoothness and non-linearity compared with GP-BART. Panels (**B**) and (**C**) both compare reasonably well with (**D**), which highlights the benefits of the rotated split rules and use of GPS, respectively. However, there is an apparent lack of smoothness in the terminal node regions of (**B**), and visible blurriness in the areas where the data splits in (**C**). Ultimately, it is evident that combining both innovations in (**D**) yields the best performance.

This conclusion is reinforced by Figure 3.9, which indicates the superior performance of version (**D**). Despite the larger variance in the RMSE, the standard GP-BART obtains the lowest median value of both metrics shown. Versions (**B**) and (**C**), which respectively incorporate rotated splits and GPs, yield similar yet slightly inferior RMSE values to (**D**), but their performance in terms of uncertainty calibration as measured by median CRPS differs more substantially. Notably, the standard BART model (**A**) is unsatisfactory from both points of view.

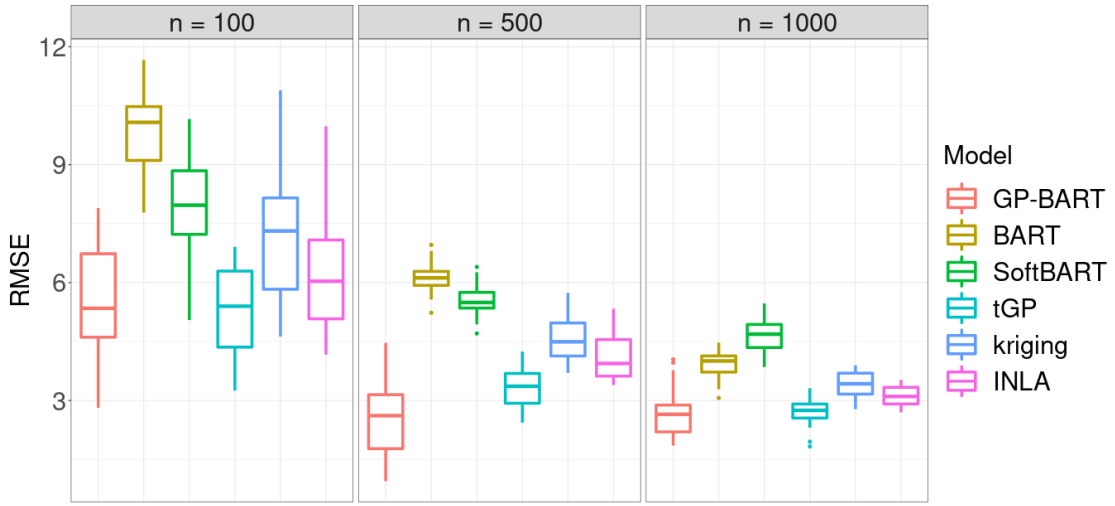


Figure 3.6: Comparisons between the RMSE obtained by the competing models for the simulated data using 10-fold cross validation over different sample sizes. These results show that GP-BART tends to deliver the lowest median RMSE, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

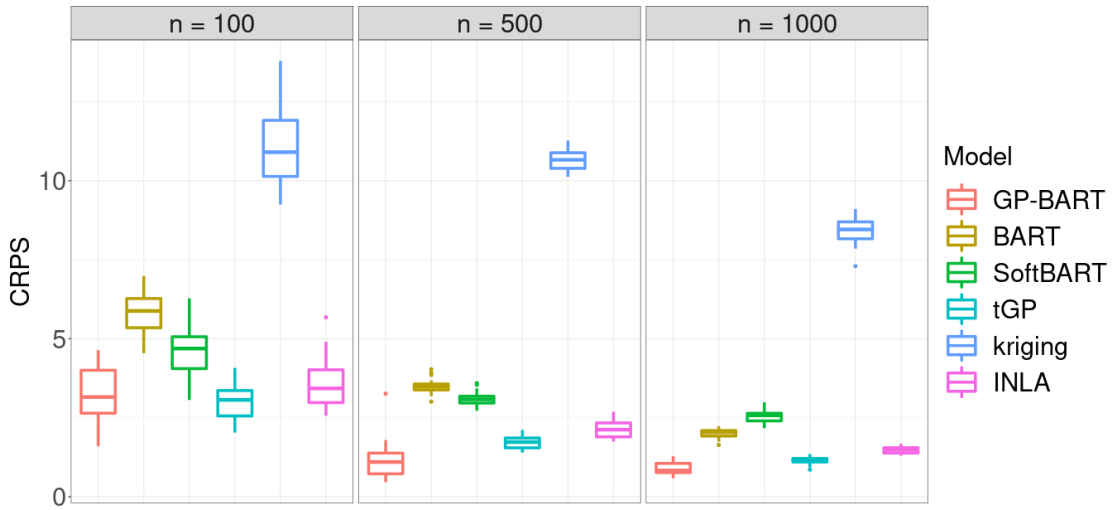


Figure 3.7: Comparisons between the CRPS values obtained by the competing models for the simulated data using 10-fold cross validation over different sample sizes. These results show that GP-BART tends to deliver the lowest median CRPS scores, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

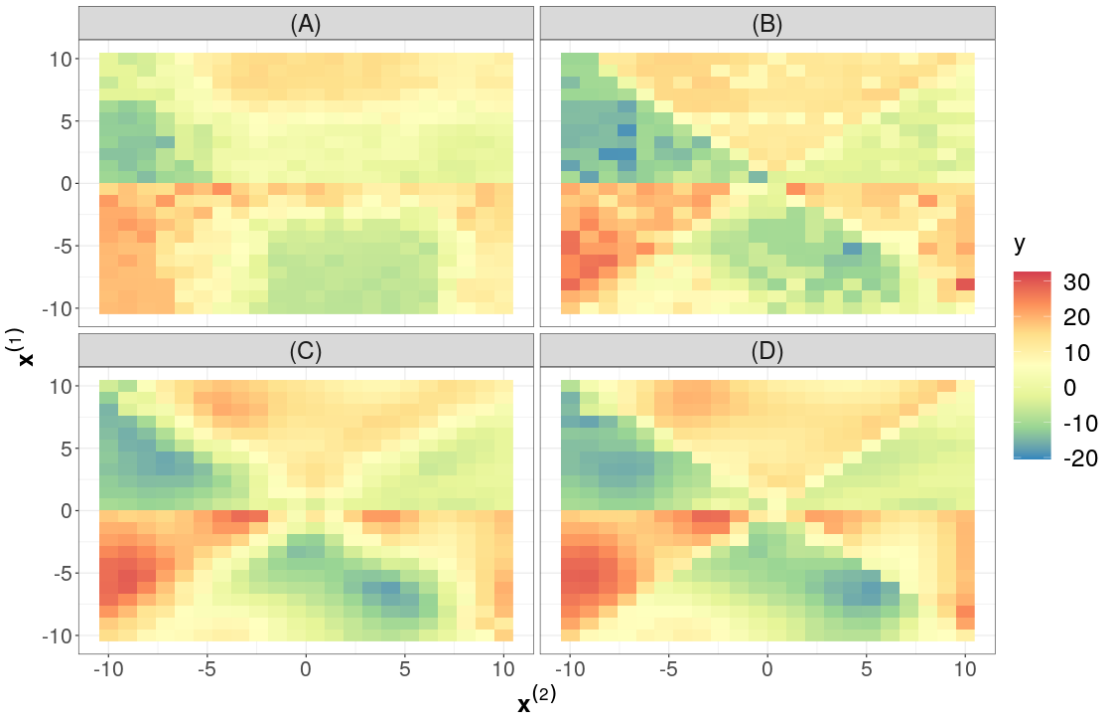


Figure 3.8: Comparison between the predicted surfaces under the different versions of GP-BART for the $n = 500$ simulated data over one randomly chosen test repetition. The surface for (**D**), the standard version of GP-BART, is qualitatively close to the observed data in the second panel of Figure 3.2.

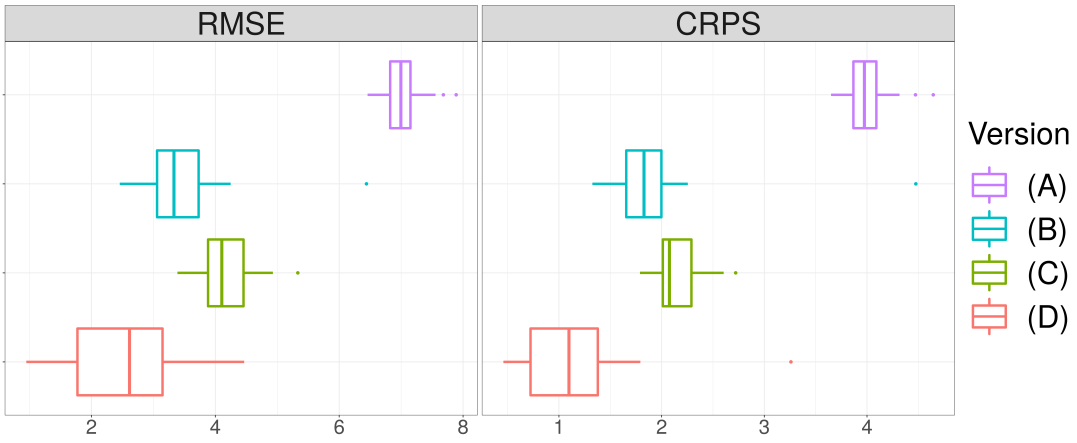


Figure 3.9: Boxplots of the RMSE (left) and CRPS (right) values across the different versions of the GP-BART model for the $n = 500$ simulated data. The standard GP-BART (**D**) has the best performance in terms of both RMSE and calibration.

3.4.2 Friedman data

In this scenario, we consider the Friedman equation (Friedman, 1991):

$$y_i = 10 \sin\left(\pi x_i^{(1)} x_i^{(2)}\right) + 20\left(x_i^{(3)} - 0.5\right)^2 + 10x_i^{(4)} + 5x_i^{(5)} + \epsilon_i,\ i = 1, \ldots, n,$$

where $x_i^{(j)} \sim \text{Unif}(0,1)\ \forall\, j = 1, \ldots, p$ and $\epsilon_i \sim \text{N}(0, \tau^{-1})$. This equation is used for benchmarking tree-based methods using synthetic data, and has been examined in many other papers, e.g., Chipman et al. (2010); Linero and Yang (2018). For these data, we compare GP-BART to its explicitly tree-based competitors, namely BART, SoftBART, and tGP. Though there are no spatial features here, we still anticipate that incorporating GPs and rotated splits will help as there are non-linear smooth interactions in these data.

Here, we specify $\tau = 100$, $n = 500$, and consider two versions of the same data; firstly with $p = 5$ and secondly with $p = 10$ features, of which the first 5 are those from the first scenario. As the Friedman equation uses only 5 covariates to generate the response, the additional five predictors in the second scenario are uninformative noise variables with no effect on y_i. Figure 3.10 shows that GP-BART outperforms the other methods and presents good performance in terms of predictive accuracy and uncertainty calibration, using the RMSE and CRPS metrics as above. Subsequently, Figure 3.11 shows the same comparison, this time with the additional 5 noise variables.

The latter comparison with extra noise variables in Figure 3.11 is also favourable to GP-BART. In particular, these results show that the uninformative variables do not have a detrimental effect on its performance. This can be attributed to the discrete prior assumed for the ϕ_{tj} parameters automatically diminishing their influence on the kernels of the GPs. Conversely, the adverse effects of such variables on the RMSE and CRPS values under BART, SoftBART, and tGP are more readily apparent, when one compares Figure 3.10 and Figure 3.11. The deterioration is especially notable for tGP.

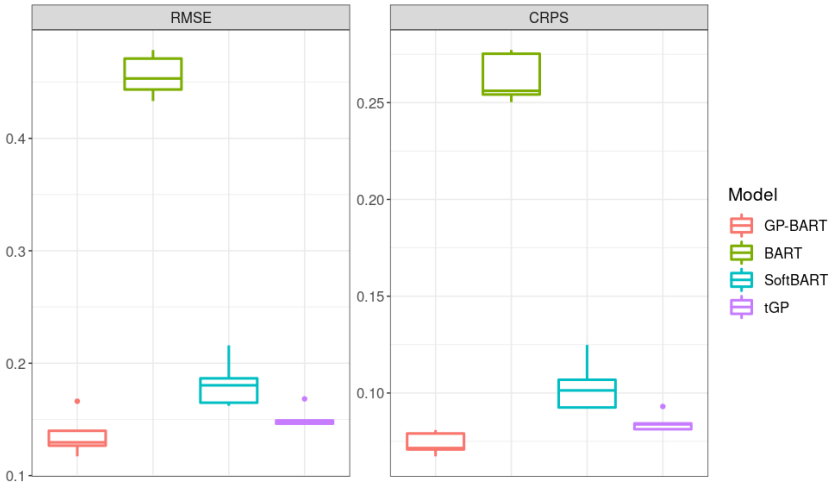


Figure 3.10: Comparison of the RMSE and CRPS over the test set in the 25 folds from 5 repetitions of 5-fold cross-validation for the Friedman data set with $n = 500$ and $p = 5$.

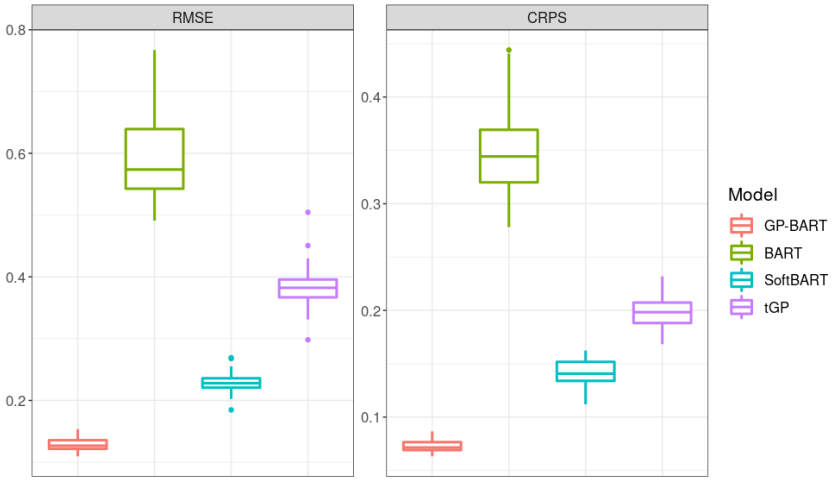


Figure 3.11: Comparison of the RMSE and CRPS over the test set in the 25 folds from 5 repetitions of 5-fold cross-validation for the Friedman data set with $n = 500$ and $p = 10$, i.e. with 5 additional noise variables.

Table 3.1 further demonstrates the effectiveness of the ARD by examining mean values of the minimum values of ϕ_{tj} for each variable (in both the $p = 5$ and $p = 10$ scenarios) over all trees and all accepted MH proposals in the retained posterior samples across each repetition of 5-fold cross-validation. In the first scenario ($p = 5$), the fourth and fifth variables, which are merely related linearly

to the response, are shown to be associated with moderately higher values. In the $p = 10$ scenario, the model selects substantially larger values for the 5 extra noise variables which are unrelated to the response, whereas small values are selected for all of the informative predictors, such that they contribute meaningfully to the GPs. Despite the remarkable performance of GP-BART in the Friedman simulations, it is important to note the extremely high signal-to-noise ratio in these cases. Given the results shown in Appendix 3.B, where the residual precision parameter τ is varied in the benchmarking experiments, we can reasonably expect the performance gap to narrow as this ratio decreases for the Friedman data also.

Table 3.1: Means and standard deviations (in parentheses) of the minimum value for ϕ_{tj} for each variable over all trees and all accepted MH proposals in the retained posterior samples across each repetition of 5-fold cross-validation on the Friedman data sets. The first row shows the $p = 5$ scenario and the subsequent rows show those same 5 variables and the 5 additional noise variables in the $p = 10$ scenario.

Friedman data	**Mean (Standard Deviation)**				
Without noise ($p = 5$)	0.45 (0.13)	0.45 (0.13)	0.20 (0.17)	0.49 (0.08)	0.55 (0.16)
With noise ($p = 10$)	0.46 (0.12)	0.42 (0.16)	0.39 (0.18)	0.50 (0.01)	0.58 (0.18)
	47.4 (10.78)	47.2 (11.09)	47.7 (10.06)	48.2 (8.98)	43.2 (16.93)

3.4.2.1 Computational performance and cost considerations

While GP-BART exhibits superior performance compared to its tree-based competitors, it is important to acknowledge its additional computational costs. This chiefly arises from its composition as a sum of GPs, which incurs a computational complexity of $\mathcal{O}(n_{t\ell}^3)$ within each terminal node. Although the model shows favorable outcomes in the simulations presented thus far, it is important to weigh this against computational efficiency. While the the aforementioned $\mathcal{O}(n_{t\ell}^3)$ costs can be reduced by encouraging deeper trees *a priori*, the relevant hyperparameters of Equation (3.4) should be handled with care. We continue to adopt the default values of $\alpha = 0.95$ and $\beta = 2$ as modifying them can decrease run times but comes at the expense of worse predictive performance. See Appendix 3.D for more details.

To asses the algorithm's computational demands, the `microbenchmark` R package (Mersmann, 2021) was used to obtain accurate measurements of the run times for GP-BART, tGP, and SoftBART, with five replications for each method. These competing treed models were specifically chosen due to their substantial computational requirements, and both were applied using their default settings. For GP-BART, our own R package based on C++ code was used. All computations were performed using R version 4.2.1 on a MacBookPro laptop, equipped with a 2.3 GHz Dual-Core Intel Core i5 processor and 8GB of RAM. The experiments were conducted on the Friedman data set with noise variables (i.e., $p = 10$), while varying the training sample size (n_{tr}) among $\{50, 100, 500\}$ and keeping the testing sample size fixed at $n_{te} = 50$.

The findings are summarized in Table 3.2, which shows that both GP-BART and tGP experience a rapid escalation in computational time as the training sample size (n_{tr}) increases. Notably, GP-BART exhibits the highest computational burden in the comparison. Indeed, the run times with $n_{tr} = 500$ suggest that GP-BART would need to be run on a dedicated machine or server for feasible modelling of larger datasets. However, it is noteworthy that despite the greater run times required by GP-BART, its timings remain comparable to those of tGP, particularly when considering the ratio of GP-BART's timings to the number of trees ($T = 20$ in its default setting).

Table 3.2: Computational time statistics for the $p = 10$ Friedman data in seconds, across five runs of each implementation for GP-BART and two tree-based competitors.

Method	**Metric**	$\mathbf{n_{tr} = 50}$	$\mathbf{n_{tr} = 100}$	$\mathbf{n_{tr} = 500}$
GP-BART	Min.	150.8	442.9	37633.9
	Mean	167.1	458.4	39360.5
	Max.	176.4	482.2	40187.2
tGP	Min.	5.1	20.6	2062.8
	Mean	6.2	21.6	2119.9
	Max.	6.5	23.8	2177.0
SoftBART	Min.	10.7	12.5	43.2
	Mean	12.8	15.8	44.0
	Max.	13.6	21.7	44.8

3.5 Applications

In this Section, we appraise the predictive performance of GP-BART compared to BART, SoftBART, tGP, kriging, and INLA on diverse real data sets, as a larger and more challenging test of GP-BART's capabilities. For illustration, we use four public data sets containing spatial features; i.e., with inherent dependence over the observations. These data sets are:

1. The *Auckland* data; consisting of 166 observations describing infant mortality in Auckland, with two spatial covariates and the target variable (Bivand and Wong, 2018).

2. The *Baltimore* data; comprising 221 observations of house sales prices, two spatial features, and 13 other covariates, not all of which are continuous (Bivand and Wong, 2018).

3. The *Boston* data; containing 506 observations of the median values of owner-occupied suburban homes, two spatial features, and 13 other covariates, not all of which are continuous. We model a corrected version (Gilley and Pace, 1996) of the original data (Harrison and Rubinfeld, 1978).

4. *Swmud*; a data set of seabed mud content in the southwest Australia Exclusive Economic Zone with 177 observations of two sets of spatial coordinates and mud content as the target variable (Li et al., 2011).

Our implementations of each algorithm follow their respective default settings, including those previously described in Sections 3.2 and 3.3 for GP-BART. As before, 5 repetitions of 5-fold cross-validation are used to evaluate performance. Categorical features cannot be formally accommodated in the GPs under the present parameterisation of GP-BART's kernel function. Hence, for the *Baltimore* and *Boston* data sets, we restrict the GPs to include the continuous and integer-valued covariates only. However, categorical features are still used to form splitting rules for GP-BART, as described in Section 3.2.1. All other methods accommodate categorical features using dummy variable representations. In each case, the strictly spatial continuous features represent the exact coordinates of the instances.

The results are summarised in Figure 3.12 and Figure 3.13, which show the RMSE and CRPS, respectively, for each data set over all folds. According to Figure 3.12, GP-BART presents the lowest median RMSE for the *Auckland*, *Boston*, and *Swmud* data sets. The difference is most pronounced for the *Boston* data, for which kriging and INLA perform notably worse than all tree-based methods. For the *Baltimore* data, it ranks second among all methods.

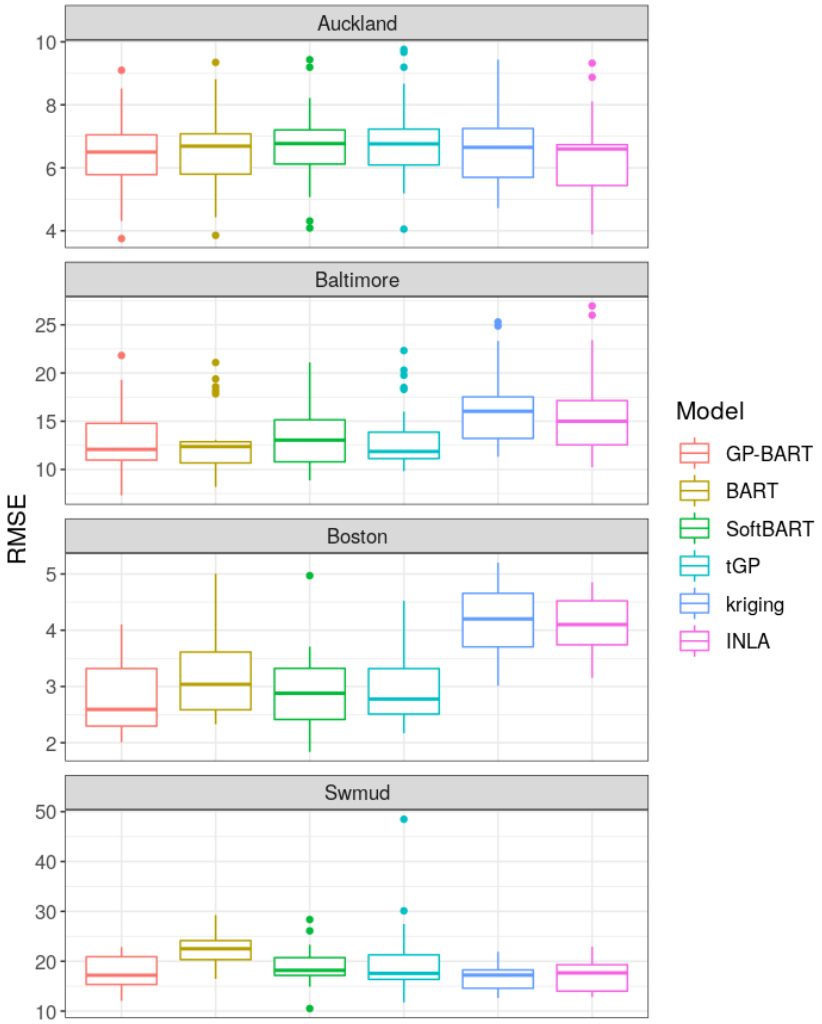


Figure 3.12: Comparison between the RMSE values for the benchmarking data sets across the six competing methods using 5 repetitions of 5-fold cross-validation.

Figure 3.13 shows that the CRPS values produced by GP-BART are similarly favourable when compared with the performance of the other algorithms, with GP-BART having the lowest or second-lowest median CRPS values for all but the *Baltimore* data set. Note that boxplots of the CRPS values for kriging are omitted from Figure 3.13 for the sake of visual clarity, as they are well outside the range of those for the other models in the comparison. Jointly considering both the predictive accuracy and the uncertainty calibration, GP-BART was able to consistently yield superior or competitive predictions.

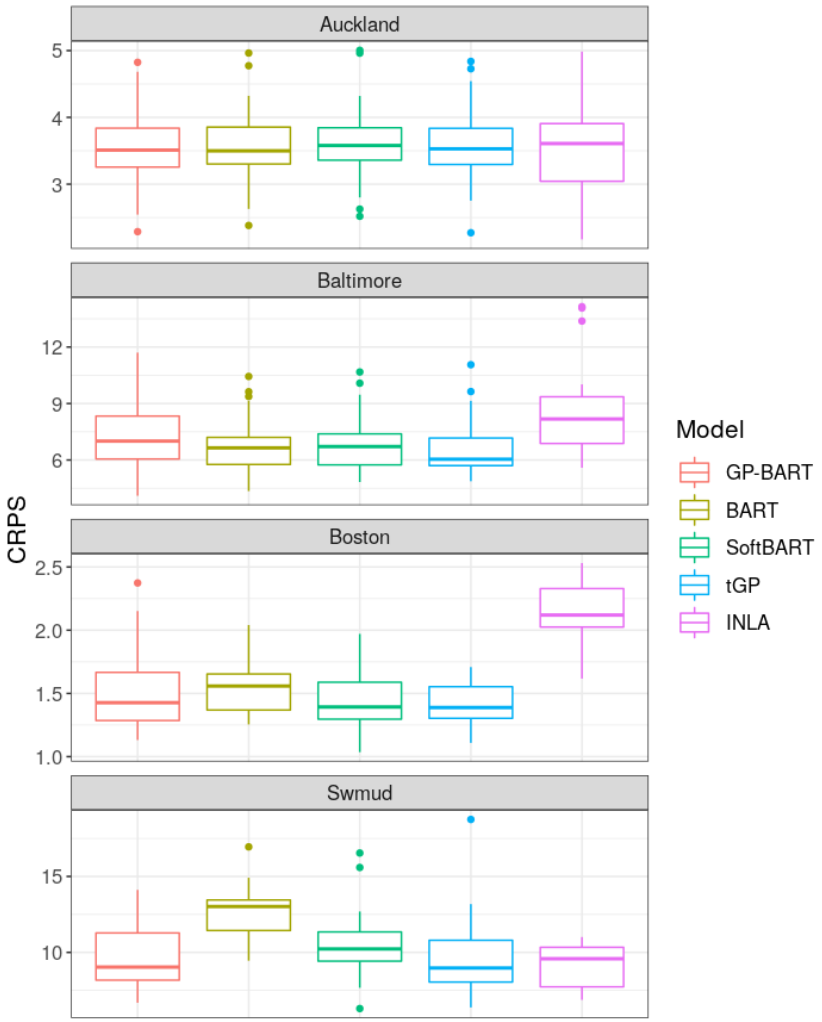


Figure 3.13: Comparison between CRPS values for the benchmarking data sets across five of the six competing methods using 5 repetitions of 5-fold cross-validation.

Given the variability in these boxplots, another aspect of performance evaluation for each model and data set is illustrated in Figure 3.14, which presents the average RMSE rank for each of the 25 test partitions from the repeated cross-validations. Ranks are defined here such that the model yielding the lowest mean RMSE is given a rank of 1, while the one with the worst prediction performance is given the highest possible rank of 6, for each test partition. From Figure 3.14, we can see that GP-BART has the lowest average RMSE rank for the *Boston* data set, particularly compared to the standard BART model. For the *Auckland* data, INLA's performance in this regard is also the best followed right after by GP-BART, where both jointly outperform the other methods. For the *Swmud* data, GP-BART presents the lowest average ranking among treed methods, losing only to the traditional spatial methods. Finally, GP-BART's performance on the *Baltimore* data is competitive with respect to other methods based on trees and GPs and superior to the traditional spatial methods.

Figure 3.15 also relies on average ranks, though here using CRPS as the metric of comparison in order to evaluate uncertainty quantification. As per Figure 3.14, GP-BART performs best among the treed methods for the *Auckland* data and performs better than the traditional spatial methods for the *Baltimore* data. For the *Boston* data, SoftBART and tGP surpass all other methods, but GP-BART achieves the next-lowest mean rank. Finally, GP-BART remains competitive for the *Swmud* data, notably outperforming the standard BART. Following its omission from Figure 3.13, kriging's CRPS performance is by far the worst for each data set.

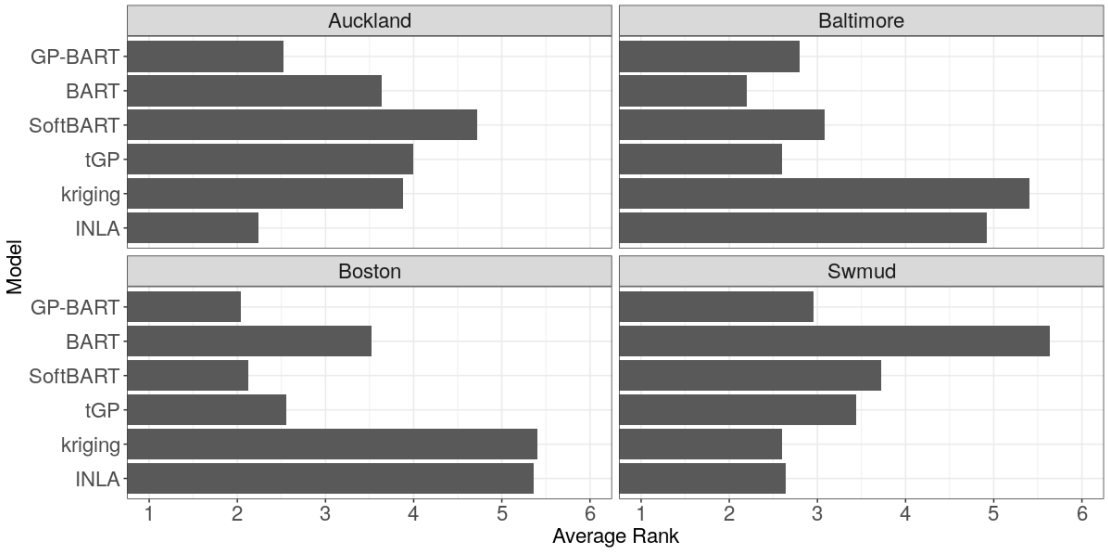


Figure 3.14: RMSE ranks for all six competing models over the four benchmark data sets, averaged over all five repetitions of the 5-fold cross validation. The ranks range from 1 to 6, with lower ranks being associated with lower mean RMSE values.

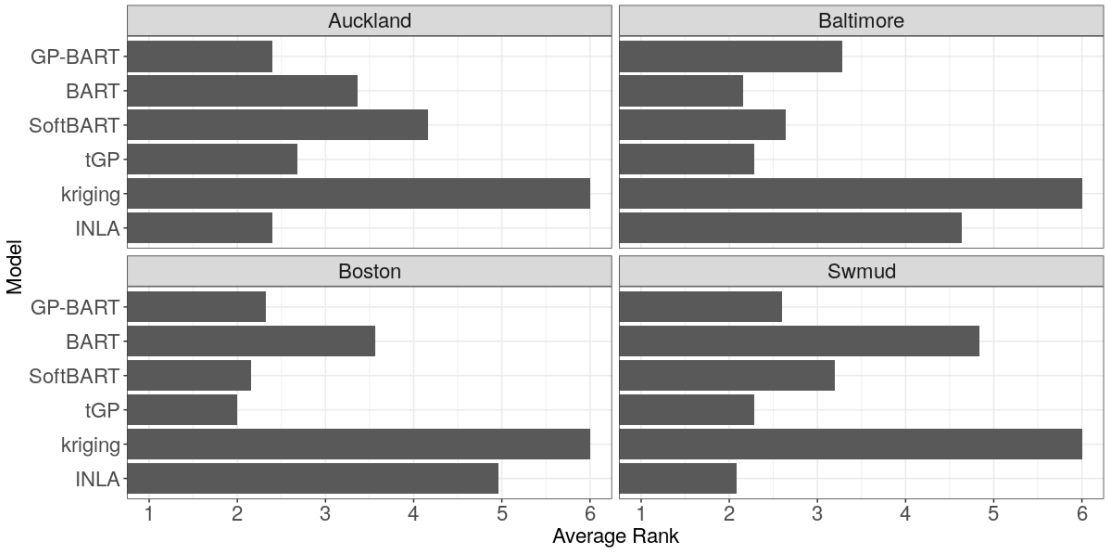


Figure 3.15: CRPS ranks for all six competing models over the four benchmark data sets, averaged over all five repetitions of the 5-fold cross validation. The ranks range from 1 to 6, with lower ranks being associated with lower mean CRPS values.

3.6 Discussion

In this chapter, we proposed GP-BART as an extension to the standard BART model. We used Gaussian processes (GPs) to make observation-specific predictions at the terminal node level, and thus are able to capture non-linear relations and spatial dependence through the covariance structure of the GPs. In addition, our novel model allows the use of rotated splitting rules to build rotated partitions, which enable more flexibility in the tree representations.

The performance of GP-BART was evaluated over a number of simulated scenarios, where the model outperformed BART, restricted versions of GP-BART itself without the use of GPs and/or novel rotated splitting rules, and another unrelated BART extension. Our benchmarking studies also highlighted GP-BART's superior performance relative to some spatial models, namely basic kriging and INLA. Our second simulation setting, using data generated according to the well-known Friedman equation, without explicit spatial components, was also favourable to GP-BART over other tree-based methods. In particular, these results demonstrated GP-BART's insensitivity to the inclusion of noise variables through the use of ARD.

When tested on real applications, using out-of-sample data via 5 repetitions of 5-fold cross-validation, GP-BART displayed competitive predictive capabilities, beating many of the established methods. We also compared the calibration properties of our method using CRPS; again, GP-BART performed as well or better than competing methodologies. Overall, in terms of predictive accuracy and uncertainty quantification, GP-BART consistently showed promising performance from both perspectives.

There are several potential issues remaining with the model and the sampling algorithm, which may provide opportunities for future research and further performance improvements:

- Careful choices have been made regarding the specification of prior distributions for the model parameters because the trees and the GPs can compete to explain the variability in the data. We have endeavoured to set sensible

default parameters throughout. However, a more substantial study might suggest general rules as to how these parameters might be elicited in light of certain data set properties. In some simpler scenarios, reparameterising the kernel functions to specify the length parameters at the tree-level only (i.e., no longer adopting variable-specific ϕ_{tj}) may be appropriate, and would significantly speed-up the algorithm by reducing the number of likelihood evaluations involved in learning these parameters via MH. However, predictive performance may deteriorate as a result of this simplification in the presence of uninformative variables or other cases where variables contribute unequally to the GPs. Alternatively, block updates of ϕ_{tj} would also reduce the computational burden, though designing an efficient proposal distribution for simultaneously sampling an adequate set of parameter values is not a trivial task.

- The model can be computationally challenging to fit for larger data sets, since the calculation of each terminal node's contribution to the overall likelihood involves inverting each associated covariance matrix, though the cost is reduced from $\mathcal{O}(n^3)$ under a single GP to $\mathcal{O}(n_{t\ell}^3)$ *per node*, given the partitioning introduced by the tree structure. Marginalising the GP mean parameters also speeds up the algorithm. Potential strategies for further speeding up the algorithm fall into two categories.

    1. Regarding the necessary matrix computations, scalable, sparse, greedy approximations for GPs — e.g., the Nyström method (Williams et al., 2002) or the methods of Quiñonero-Candela et al. (2007), Rahimi and Recht (2007), and Wilson et al. (2020) — may also be advantageous in future work. However, such approximations may compromise model performance compared to our present MCMC implementation.

    2. Incorporating warm-up procedures to initialise GP-BART could be another viable strategy. For instance, XBART (He and Hahn, 2023) employs recursive partitioning and other modifications to the standard BART to rapidly find large trees which fit the data well; by ensuring that its draws are in high-probability regions of the BART posterior, this approach greatly reduces burn-in times. Seeding GP-BART in a

similar fashion would allow convergence to be achieved more rapidly. However, it would be crucial to carefully design this initialisation process to align with the specific GP-BART setting, as a faster initialisation may potentially result in local minima or wasteful iterations and trees in high-probability regions of the BART posterior may not be well-suited to GP-BART.

- In general, determining variable importance in GP-BART is difficult as variables may contribute to both the GPs and/or the splits. Though the ARD appears to adequately capture relevant variables and account for irrelevant variables in the applications considered herein, there is further scope for re-calibrating the discrete prior and proposal distributions for the length parameters in cases where there is prior knowledge about the relative importance of specific predictors, as well as scope for exploiting variable-selection from the BART component. At present, all continuous predictors used to construct the trees are used in the GPs, which need not be the case. It may be beneficial to restrict the GPs only to the variables used to define splits along the given branch, though this would come with significant additional computational costs.

- In the applications herein, we have focused on the use of GP-BART for spatial data sets, but there is nothing to prohibit the model being used in generic machine learning tasks. However, we have restricted the GPs to be covariance-stationary through our use of anisotropic exponentiated-quadratic kernels, which are governed only by scalar rate and tree-level, variable-specific length parameters. A superior approach may introduce non-stationarity to the autocovariance and hence produce more flexible GP surfaces. Relatedly, recall that tGP incorporates non-stationarity in its single 'treed-GP'. Doing so for GP-BART may result in our model demonstrating even further performance improvements over tGP in the applications, but it would come with more computational challenges.

  Indeed, though the model outperforms its competitors in all simulation experiments and on most of the real data sets analysed above, the underlying exponentiated-quadratic kernel functions used in our parameterisation of the

GP components may be inappropriate in other settings. Investigating alternative kernel functions to further improve GP-BART's performance is of great interest for future work. For instance, kernels capable of accommodating the non-continuous features we discarded in our analysis of the *Baltimore* and *Boston* data sets would also be of particular interest. However, this would not be immediately straightforward, given that changing the kernel necessitates specifying priors appropriately and deriving posterior distributions from scratch for sampling parameters with each new kernel and that more sophisticated kernels may further increase the computational burden.

- An advantage of Bayesian additive tree ensembles is their faster convergence compared to Bayesian CART models (Chib and Greenberg, 1998). While more trees generally lead to quicker convergence, the computational cost associated with the GPs may outweigh the benefits due to the increased burden of adjusting more trees, especially given the aforementioned costs of the required node-specific matrix inversion operations. The improved convergence of ensembles of trees relative to models with only a single tree may also explain how GP-BART is capable of effectively exploring the $\phi_{t\ell}$ parameter space and avoiding local minima. A more extensive comparison on how the number of trees can affect the convergence and how to optimise this aspect of the proposed ensemble while balancing computational considerations may furtherimprove the algorithm's performance.

We hope to report on these developments as part of our future research plans.

Appendix

3.A Tree likelihood

In general terms, following the initial formulation of the GP-BART model described in Section 3.2.2, the posterior distribution of the residuals for a terminal node ℓ in tree t is given by

$$\mathbf{R}_{t\ell} \mid \mathcal{T}_t, \mu_{t\ell}, \boldsymbol{\phi}_t, \nu, \tau \sim \text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, \tau^{-1}\boldsymbol{\mathcal{I}}_{n_{t\ell}} + \boldsymbol{\Omega}_{t\ell}\right).$$

However, in writing this likelihood, we can marginalise out the terminal-node mean parameters $\mu_{t\ell} \mid \mathcal{T}_t, \tau_\mu \sim \text{N}(0, \tau_\mu^{-1})$ as follows

$$\begin{aligned}
\pi\left(\mathbf{R}_{t\ell} \mid \mathcal{T}_t, \boldsymbol{\phi}_t, \nu, \tau\right) &= \int \pi\left(\mathbf{R}_{t\ell} \mid \mu_{t\ell}, \boldsymbol{\phi}_t, \nu, \tau\right) \pi\left(\mu_{t\ell}\right) \mathrm{d}\mu_{t\ell} \\
&\propto \left|\boldsymbol{\Gamma}_{t\ell}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{R}_{t\ell} - \boldsymbol{\mu}_{t\ell}\right)^\top \boldsymbol{\Gamma}_{t\ell}^{-1}\left(\mathbf{R}_{t\ell} - \boldsymbol{\mu}_{t\ell}\right)\right\} \times \\
&\quad \tau_\mu^{-1/2} \exp\left\{-\frac{\tau_\mu}{2}\mu_{t\ell}^2\right\},
\end{aligned}$$

where $\boldsymbol{\Gamma}_{t\ell} = \tau^{-1}\boldsymbol{\mathcal{I}}_{n_{t\ell}} + \boldsymbol{\Omega}_{t\ell}$. After further calculations, letting

$$v_{t\ell} = \mathbf{1}_{n_{t\ell}}^\top \boldsymbol{\Gamma}_{t\ell}^{-1}\mathbf{1}_{n_{t\ell}} + \tau_\mu,$$

applying the log, and then summing over the terminal nodes, we obtain

$$\begin{aligned}
\log \pi\left(\mathbf{R}_t \mid \mathcal{T}_t, \boldsymbol{\phi}_t, \nu, \tau\right) = \log \mathcal{C} &- \frac{1}{2}\sum_{\ell}^{b_t} \log v_{t\ell} - \frac{1}{2}\sum_{\ell}^{b_t} \log\left|\boldsymbol{\Gamma}_{t\ell}\right| \\
&- \frac{1}{2}\sum_{\ell}^{b_t} \mathbf{R}_{t\ell}^\top \boldsymbol{\Gamma}_{t\ell}^{-1}\mathbf{R}_{t\ell} + \frac{1}{2}\sum_{\ell}^{b_t} v_{t\ell}^{-1}\mathbf{1}_{n_{t\ell}}^\top \boldsymbol{\Gamma}_{t\ell}^{-1}\mathbf{R}_{t\ell}\mathbf{R}_{t\ell}^\top \boldsymbol{\Gamma}_{t\ell}^{-1}\mathbf{1}_{n_{t\ell}},
\end{aligned}$$

where $\mathcal{C}$ is a constant of proportionality. Recalling $\boldsymbol{\Lambda}_{t\ell} = \tau_\mu^{-1}\mathbf{1}_{n_{t\ell}}\mathbf{1}_{n_{t\ell}}^\top + \boldsymbol{\Omega}_{t\ell}$, this expression can be further simplified with the constant $\mu_{t\ell}$ parameters explicitly absorbed into the kernel of the GP. This yields the following distribution for the partial residuals

$$\mathbf{R}_{t\ell} \mid \mathcal{T}_t, \boldsymbol{\phi}_t, \nu, \tau \sim \text{MVN}\left(\mathbf{0}_{n_{t\ell}}, \tau^{-1}\boldsymbol{\mathcal{I}}_{t\ell} + \boldsymbol{\Lambda}_{t\ell}\right),$$

which bypasses the need to sample the $\mu_{t\ell}$ parameters and leads to better mixing.

3.B Performance evaluation with varying residual precision on the benchmarking experiments

To assess the model's performance across different levels of noise, we replicated the experiments from Section 3.4.1, varying the residual precision parameter τ at three levels — specifically $\tau = \{1, 0.1, 0.01\}$ — and compared GP-BART with its competitors in each case, as before. Recall that the results shown throughout Section 3.4.1 are based on $\tau = 10$ only. The results now indicate that even with increasing noise (i.e., lower precision), GP-BART maintains consistent performance and continues to exhibit the lowest median RMSE and CRPS values, though the variability of both metrics does increase as τ decreases.

The results for $\tau = 0.1$, $\tau = 0.1$, and $\tau = 0.01$ are presented in Appendices 3.B-i, 3.B-ii, and 3.B-iii, respectively. As per Section 3.4.1, we show in each case the simulated data surface for the given τ value with sample sizes of $n = 100$, $n = 500$, and $n = 1000$ and then show the predicted surfaces according to GP-BART and its competitors BART, SoftBART, tGP, kriging, and INLA at each sample size. Finally, we show boxplots of the RMSE and CRPS values obtained by the competing methods on the data generated with the respective τ value.

3.B-i Residual precision $\tau = 1$

The simulated data surfaces considering the residual precision $\tau = 1$ for different samples sizes $n = \{100, 500, 1000\}$ are shown in Figure 3.B.1. Figures 3.B.2–3.B.4

show the corresponding predicted surfaces from one randomly chosen repetition of the repeated 5-fold cross-validation for each respective sample size. As before, GP-BART's predicted surfaces more closely resemble the signal from the original data depicted in Figure 3.B.1 in every instance, when compared with its competitors. The quantitative comparison is summarised via boxplots of RMSE and CRPS values in Figure 3.B.5 and Figure 3.B.6, respectively. These boxplots reflect the conclusions draw from previous plots where, in general, GP-BART presents the lowest median values for RMSE and CRPS across all scenarios.

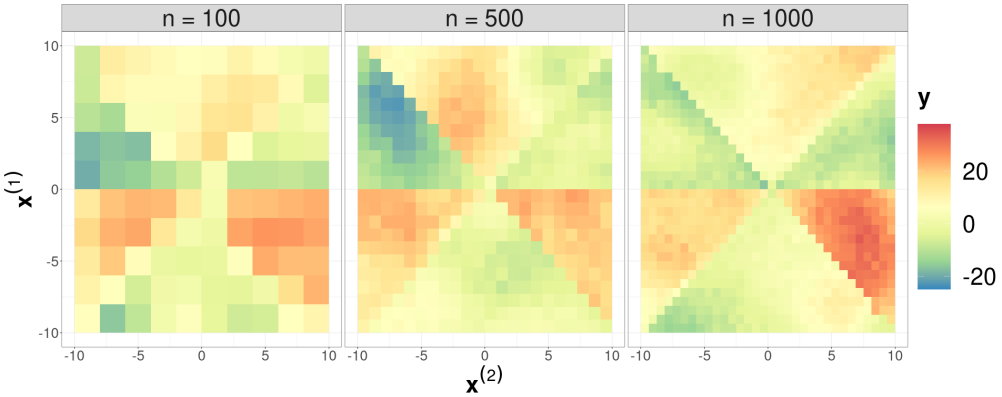


Figure 3.B.1: Simulated data with $n = \{100, 500, 1000\}$ observations, respectively, and residual precision of $\tau = 1$.

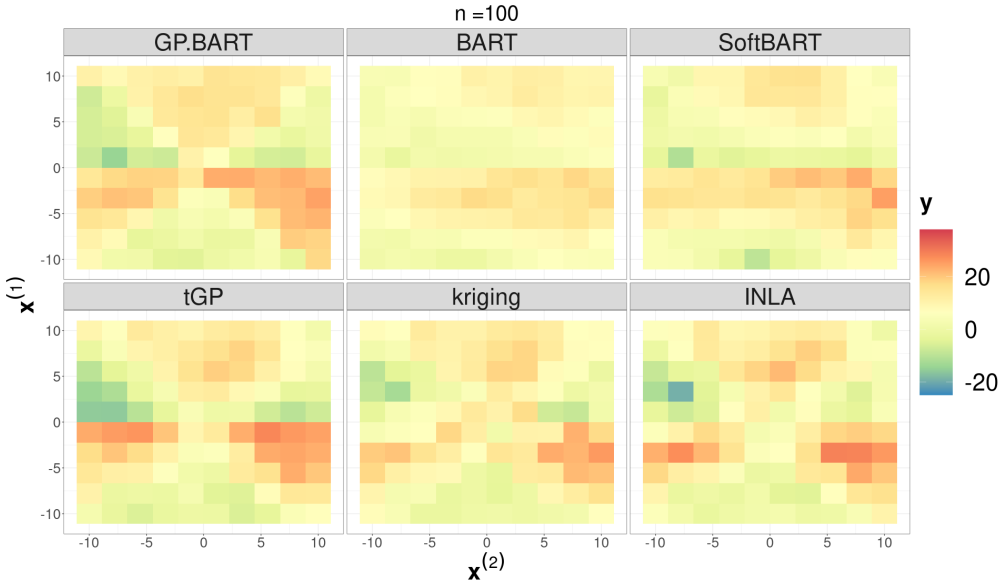


Figure 3.B.2: Predicted surfaces for the simulated scenario with $n = 100$ observations from the first panel of Figure 3.B.1 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 1$.

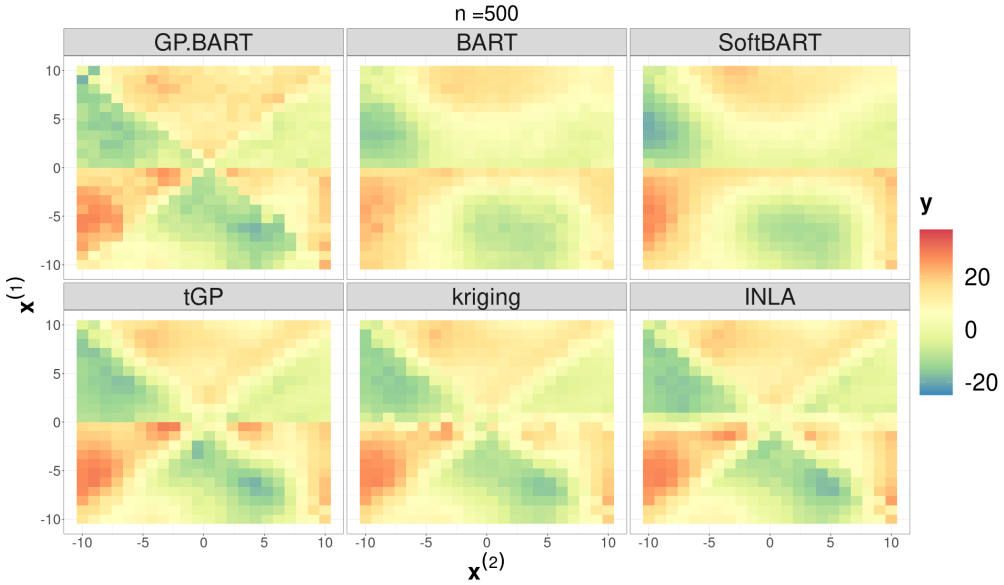


Figure 3.B.3: Predicted surfaces for the simulated scenario with $n = 500$ observations from the second panel of Figure 3.B.1 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 1$.

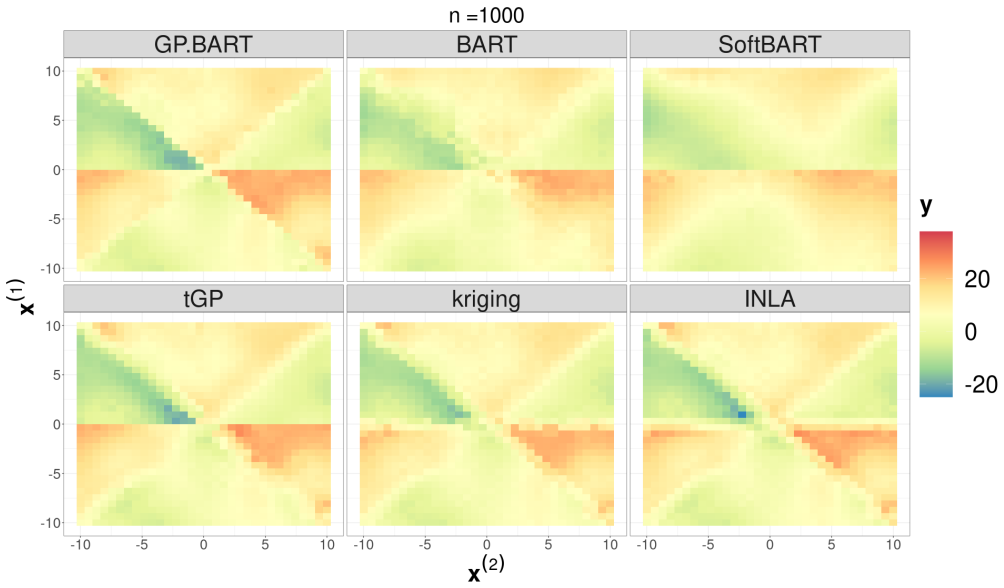


Figure 3.B.4: Predicted surfaces for the simulated scenario with $n = 1000$ observations from the third panel of Figure 3.B.1 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 1$.

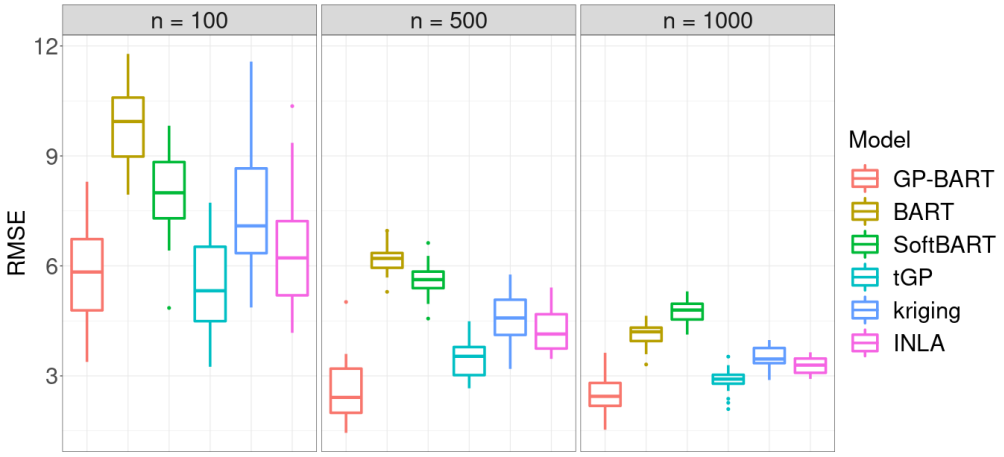


Figure 3.B.5: Comparisons between the RMSE obtained by the competing models for the simulated data using 5 repeated 5-fold cross validation over different sample sizes, and $\tau = 1$. Based on the results, it is evident that GP-BART consistently delivers the best performance on average, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

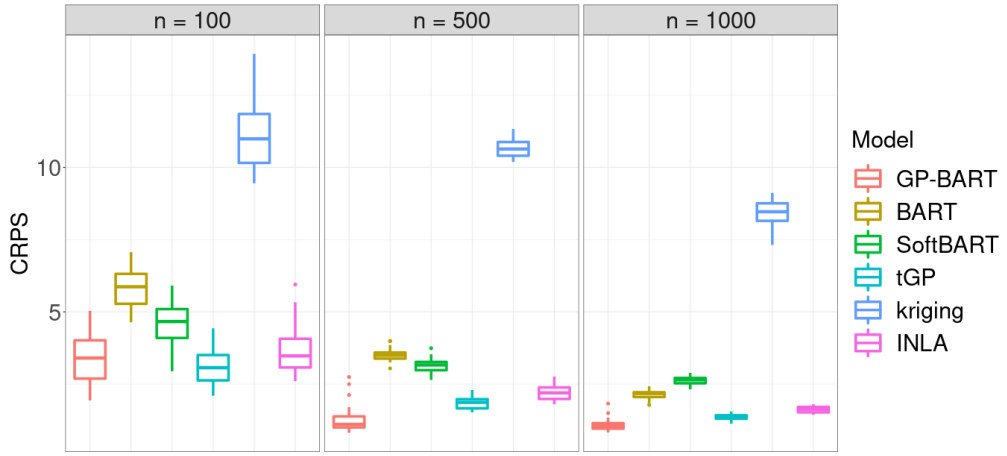


Figure 3.B.6: Comparisons between the CRPS values obtained by the competing models for the simulated data using 5 repeated 5-fold cross validation over different sample sizes, and $\tau = 1$. Based on the results, it is evident that GP-BART consistently delivers the best performance on average, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

3.B-ii Residual precision $\tau = 0.1$

The simulated data surfaces considering the residual precision $\tau = 0.1$ for different samples sizes $n = \{100, 500, 1000\}$ are shown in Figure 3.B.7. Figures 3.B.8–3.B.10 show the corresponding predicted surfaces from one randomly chosen repetition of the repeated 5-fold cross-validation for each respective sample size. As before, GP-BART's predicted surfaces more closely resemble the signal from the original data depicted in Figure 3.B.7 in every instance, when compared with its competitors. The quantitative comparison is summarised via boxplots of RMSE and CRPS values in Figure 3.B.11 and Figure 3.B.12, respectively. These boxplots reflect the conclusions draw from previous plots where, in general, GP-BART presents the lowest median values for RMSE and CRPS across all scenarios.

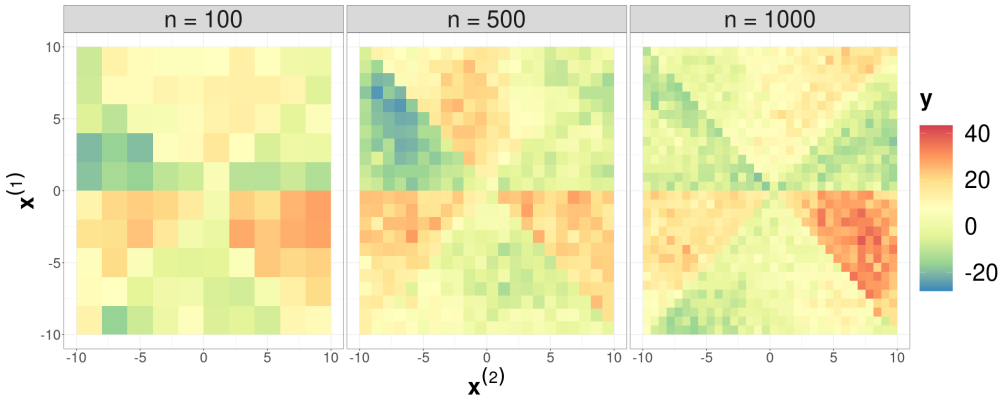


Figure 3.B.7: Simulated data with $n = \{100, 500, 1000\}$ observations, respectively, and residual precision of $\tau = 0.1$.

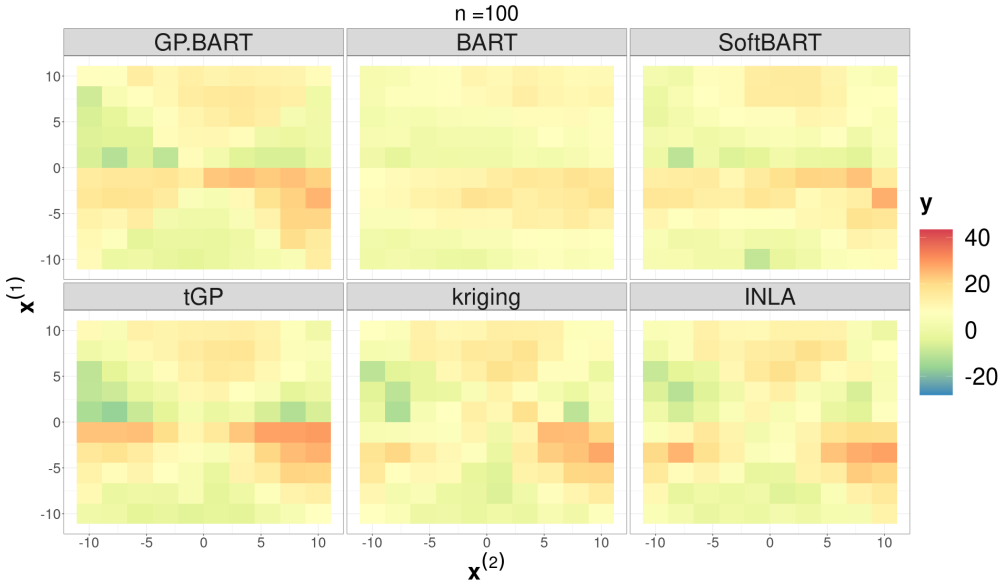


Figure 3.B.8: Predicted surfaces for the simulated scenario with $n = 100$ observations from the first panel of Figure 3.B.7 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 0.1$.

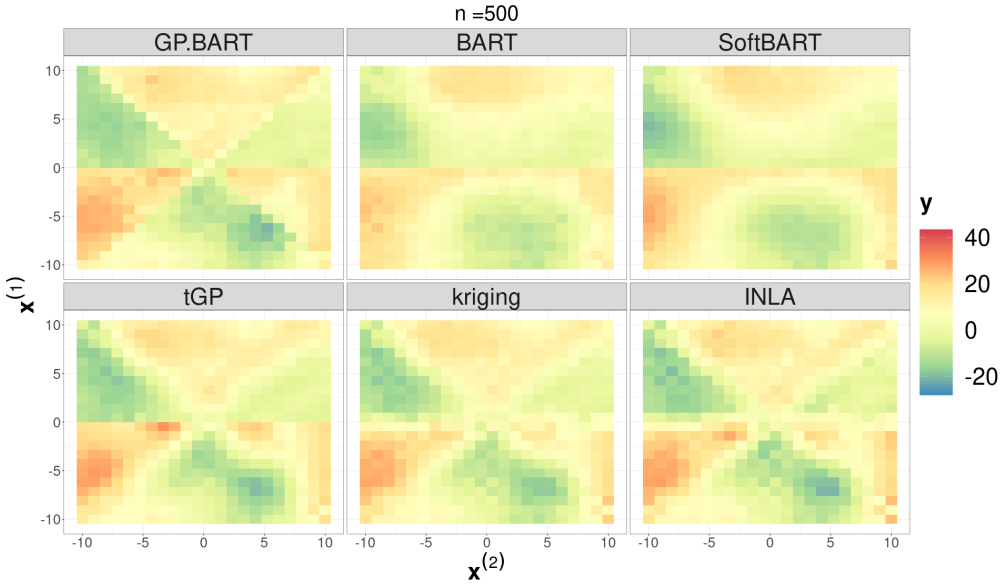


Figure 3.B.9: Predicted surfaces for the simulated scenario with $n = 500$ observations from the second panel of Figure 3.B.7 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 0.1$.

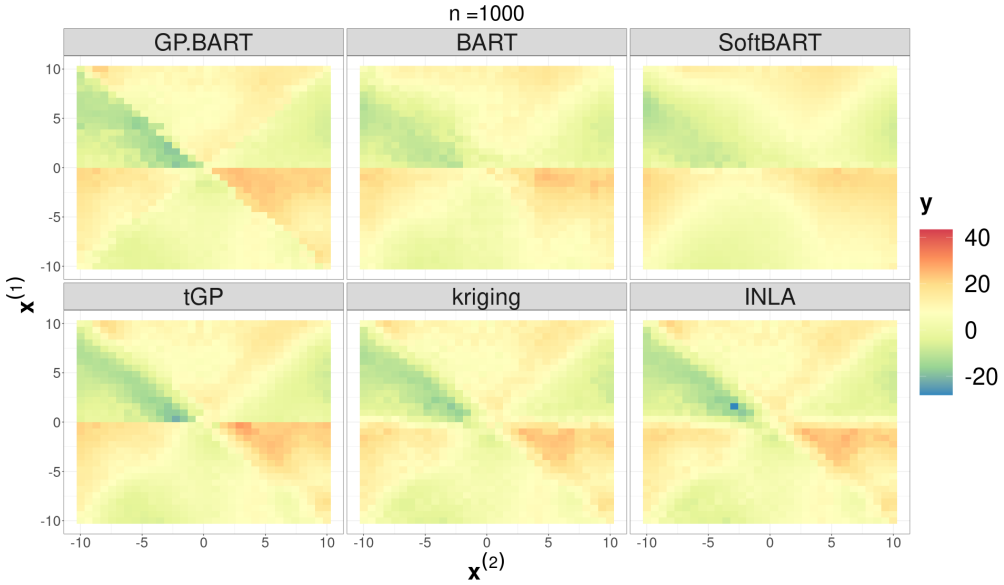


Figure 3.B.10: Predicted surfaces for the simulated scenario with $n = 1000$ observations from the third panel of Figure 3.B.7 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 0.1$.

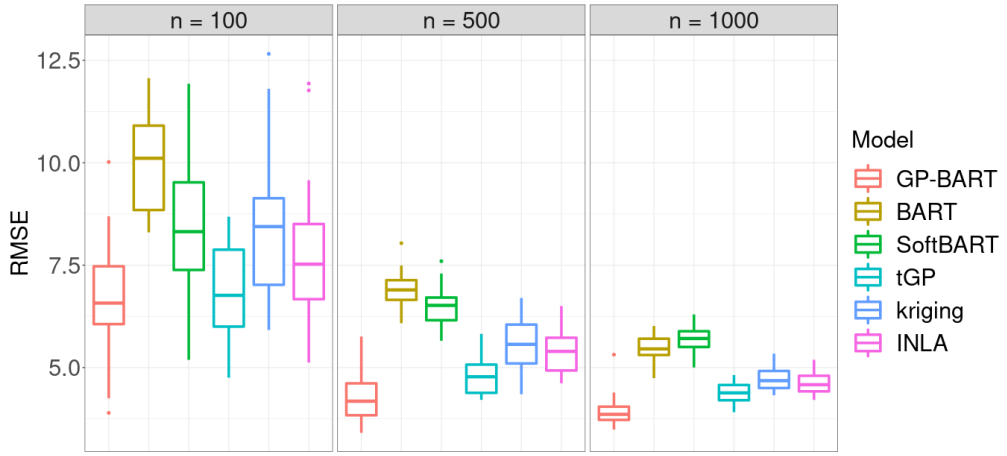


Figure 3.B.11: Comparisons between the RMSE obtained by the competing models for the simulated data using 5 repeated 5-fold cross validation over different sample sizes, and $\tau = 0.1$. Based on the results, it is evident that GP-BART consistently delivers the best performance on average, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

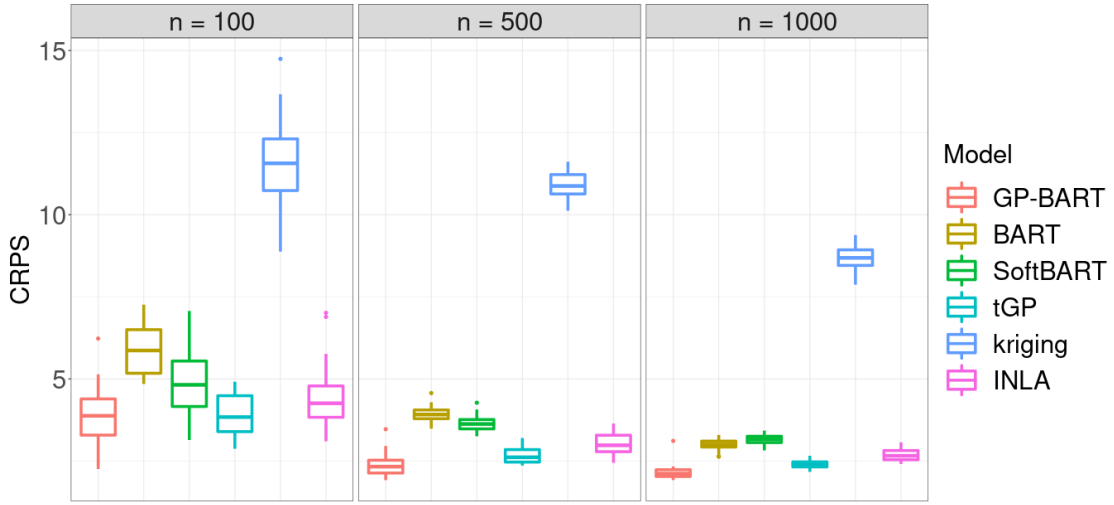


Figure 3.B.12: Comparisons between the CRPS values obtained by the competing models for the simulated data using 5 repeated 5-fold cross validation over different sample sizes, and $\tau = 0.1$. Based on the results, it is evident that GP-BART consistently delivers the best performance on average, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

3.B-iii Residual precision $\tau = 0.01$

The simulated data surfaces considering the residual precision $\tau = 0.01$ for different samples sizes $n = \{100, 500, 1000\}$ are shown in Figure 3.B.13. Figures 3.B.14–3.B.16 show the corresponding predicted surfaces from one randomly chosen repetition of the repeated 5-fold cross-validation for each respective sample size. As before, GP-BART's predicted surfaces more closely resemble the signal from the original data depicted in Figure 3.B.13 in every instance, when compared with its competitors. The quantitative comparison is summarised via boxplots of RMSE and CRPS values in Figure 3.B.17 and Figure 3.B.18, respectively. These boxplots reflect the conclusions draw from previous plots where, in general, GP-BART presents the lowest median values for RMSE and CRPS across all scenarios.

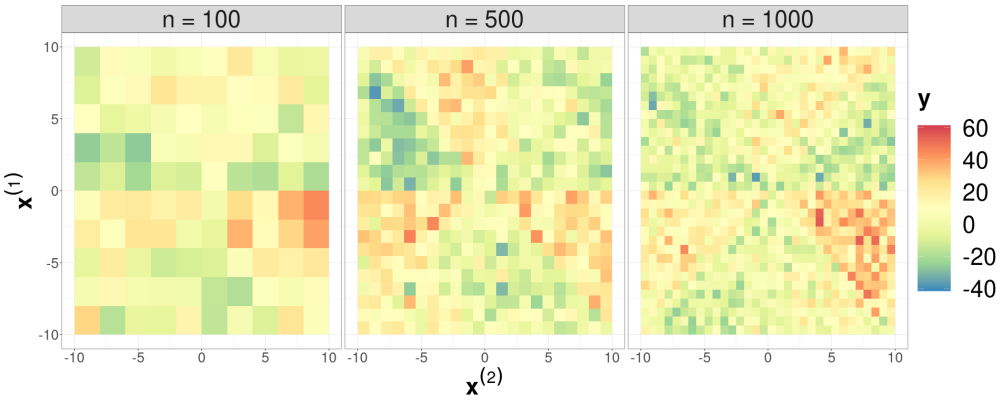


Figure 3.B.13: Simulated data with $n = \{100, 500, 1000\}$ observations, respectively, and residual precision of $\tau = 0.01$.

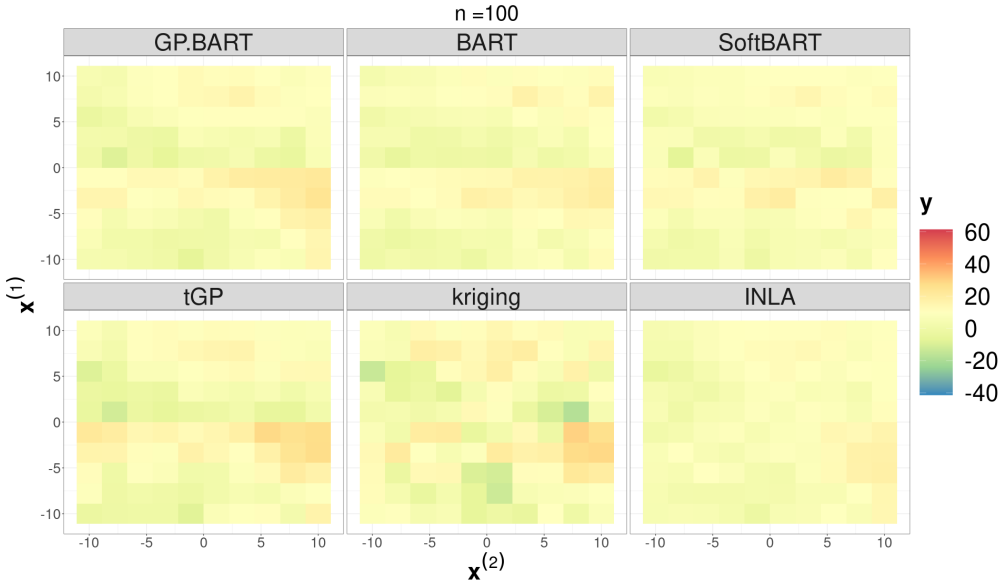


Figure 3.B.14: Predicted surfaces for the simulated scenario with $n = 100$ observations from the first panel of Figure 3.B.13 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 0.01$.

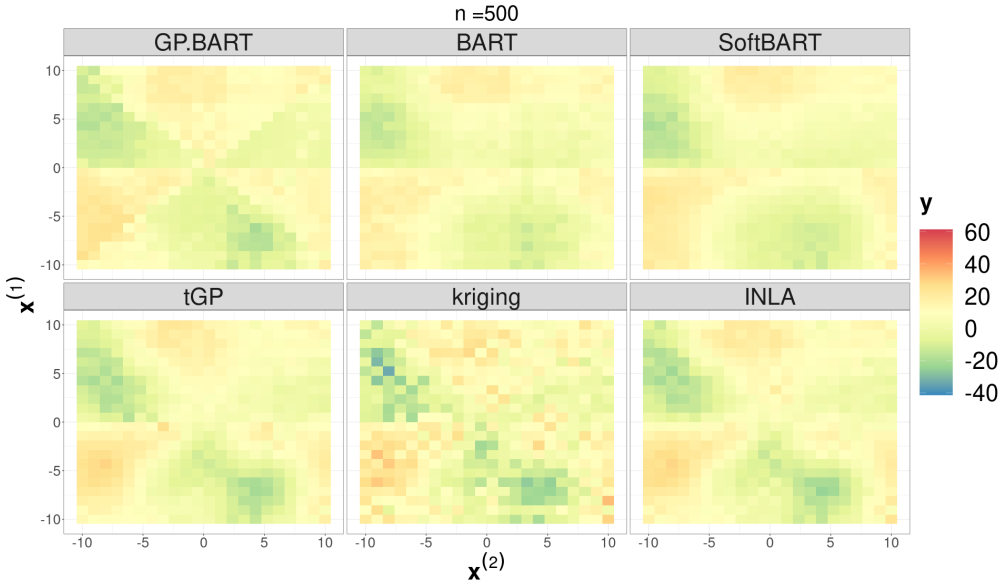


Figure 3.B.15: Predicted surfaces for the simulated scenario with $n = 500$ observations from the second panel of Figure 3.B.13 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 0.01$.

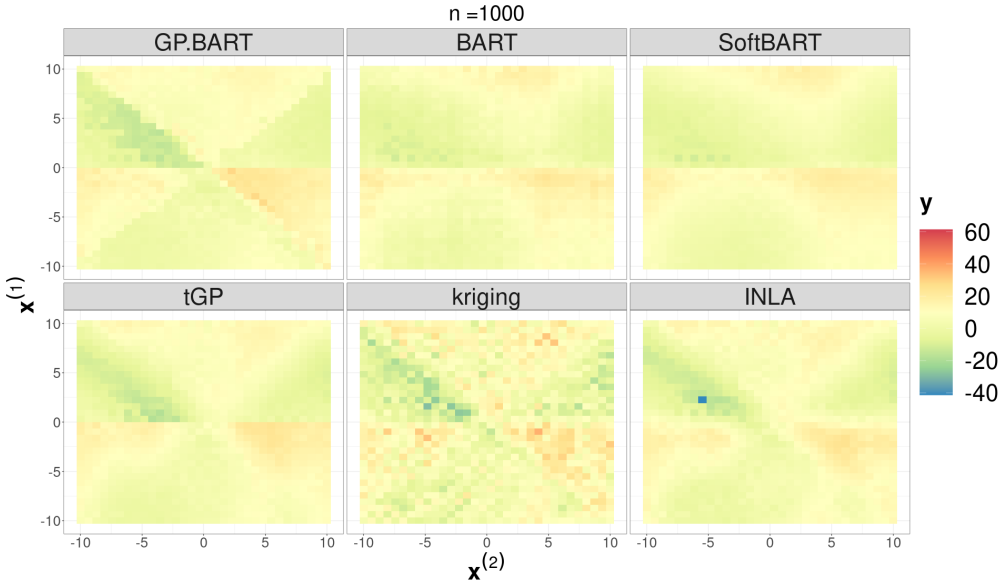


Figure 3.B.16: Predicted surfaces for the simulated scenario with $n = 1000$ observations from the third panel of Figure 3.B.13 using different methods over one randomly chosen test repetition. The residual precision is $\tau = 0.01$.

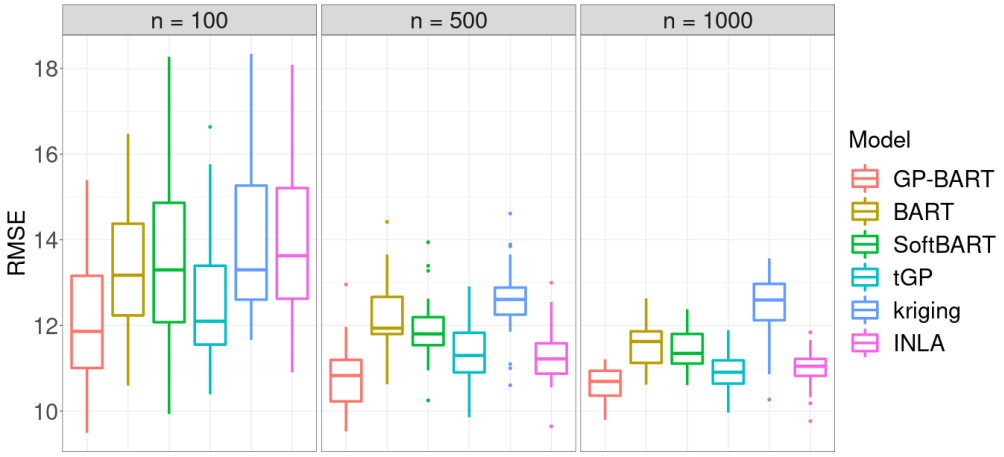


Figure 3.B.17: Comparisons between the RMSE obtained by the competing models for the simulated data using 5 repeated 5-fold cross validation over different sample sizes, and $\tau = 0.01$. Based on the results, it is evident that GP-BART consistently delivers the best performance on average, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

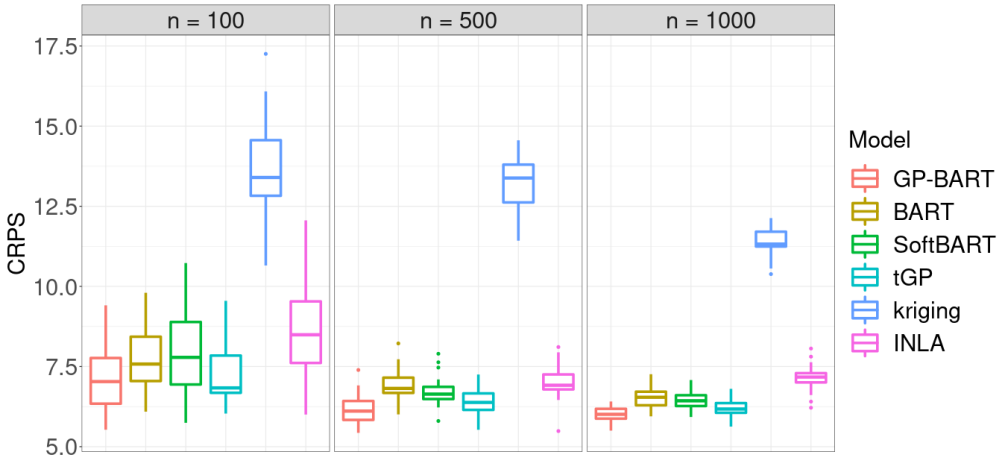


Figure 3.B.18: Comparisons between the CRPS values obtained by the competing models for the simulated data using 5 repeated 5-fold cross validation over different sample sizes, and $\tau = 0.01$. Based on the results, it is evident that GP-BART consistently delivers the best performance on average, as it encompasses assumptions of spatial dependence, smoothness, and allows for rotated splits.

3.C Performance evaluation for restricted versions of GP-BART

The results of a comparison between different versions of GP-BART for simulated data with $n = 500$ are illustrated in Figure 3.8, showing predicted surfaces, and Figure 3.9, showing boxplots of the RMSE and CRPS values. For completeness, we provide here the analogous plots for the other sample sizes considered in the simulation study, with predicted surfaces and boxplots for the $n = 100$ data in Figures 3.C.1 and 3.C.2, respectively, and equivalent plots for the $n = 1000$ data in Figures 3.C.3 and 3.C.4. Recall that the restricted versions of GP-BART evaluated here are: (**A**) without any projection moves or GPs (equivalent to the standard BART model); (**B**) without GPs, but with the addition of the new rotation moves; (**C**) without the new moves, but with GPs; and (**D**) the standard GP-BART with both rotated split rules and GPs. Finally, numerical summaries of the median RMSE and CRPS values for all sample sizes across all four versions are summarised in Table 3.C.1 and the acceptance rates for the tree-proposal moves under the full GP-BART are summarised in Table 3.C.2.

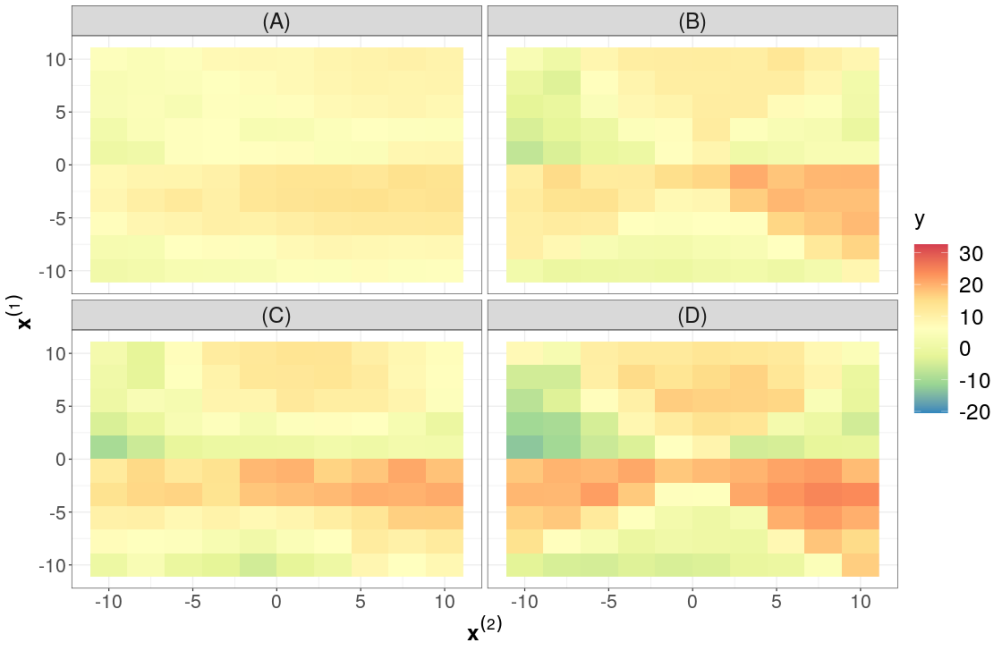


Figure 3.C.1: Comparison between the predicted surfaces under the different versions of GP-BART for the $n = 100$ simulated data over one randomly chosen repetition. The surface for (**D**), the standard version of GP-BART, is qualitatively close to the observed data in the first panel of Figure 3.2.

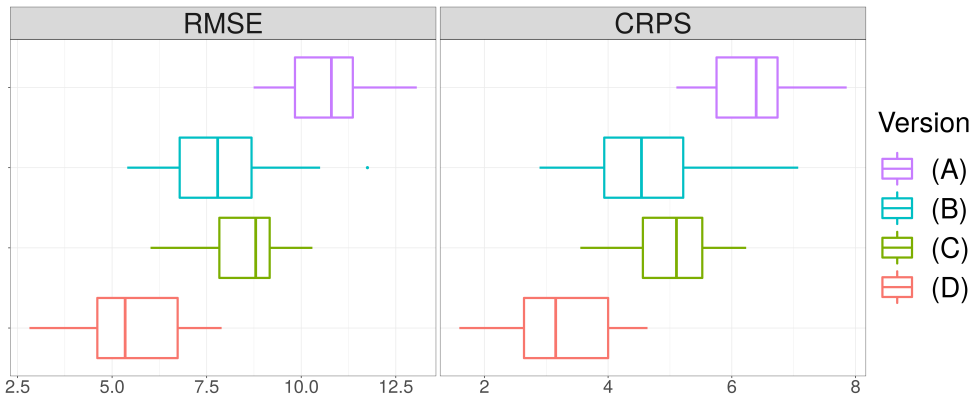


Figure 3.C.2: Boxplots of the RMSE (left) and CRPS (right) values across the different versions of the GP-BART model for the $n = 100$ simulated data. The standard GP-BART (**D**) has the best performance in terms of both RMSE and calibration.

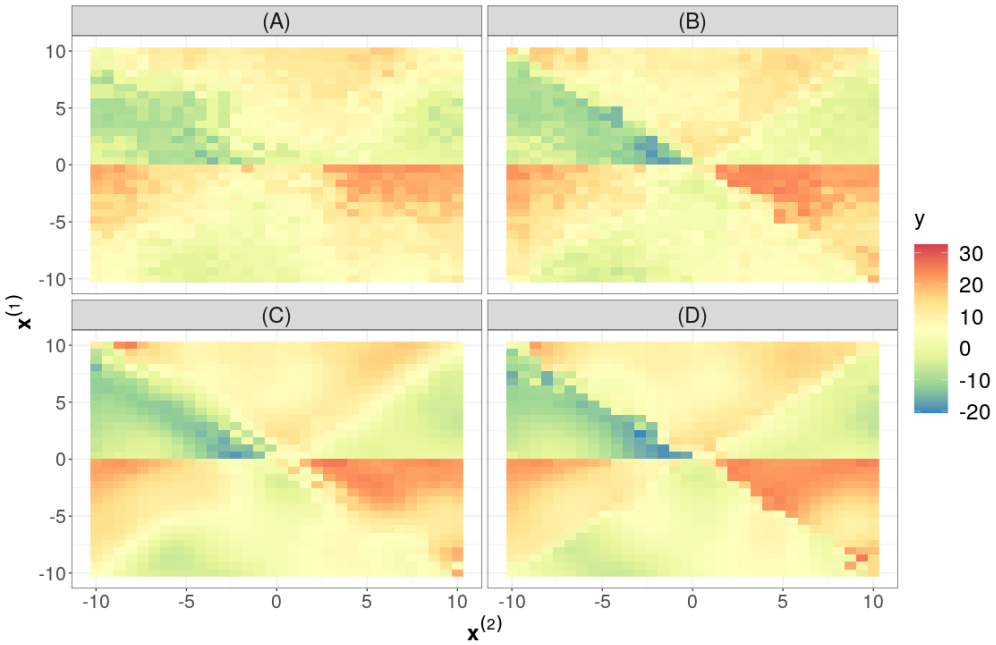


Figure 3.C.3: Comparison between the predicted surfaces under the different versions of GP-BART for the $n = 1000$ simulated data over one randomly chosen test repetition. The surface for (**D**), the standard version of GP-BART, is qualitatively close to the observed data in the third panel of Figure 3.2.

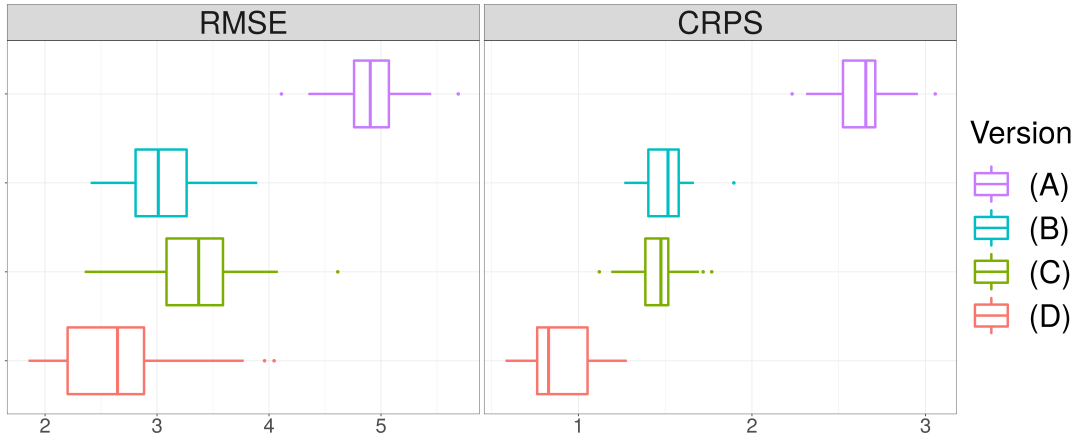


Figure 3.C.4: Boxplots of the RMSE (left) CRPS (right) values across the different versions of the GP-BART model for the $n = 1000$ simulated data. The standard GP-BART (**D**) has the best performance in terms of both RMSE and calibration.

The behaviour of versions (**B**) and (**C**) in Figures 3.C.2 and 3.C.4 is consistent with the corresponding Figure 3.9 for the $n = 500$ benchmarking experiment in Section 3.4.1. These versions clearly demonstrate the efficacy of the novel grow-rotate and change-rotate moves and the use of GP priors over terminal nodes, in that they show improved performance relative to the standard BART according to both metrics, but incorporating both innovations under GP-BART (**D**) yields the best performance. Regarding Figures 3.C.1 and 3.C.3, the predicted surface under GP-BART is the one which is closest to the observed data in each case. To provide further clarity, Table 3.C.1 numerically summarises the median lines of the boxplots from Figure 3.9, Figure 3.C.2, and Figure 3.C.4. All versions present lower values of both metrics as the sample size increases. While (**B**) and (**C**) improve on the standard BART (**A**) in each case, GP-BART remains the superior method from both perspectives at each value of n. Though the difference between it and its competitors in terms of RMSE and CRPS becomes less pronounced as n increases, GP-BART remains the best from the points of view of prediction accuracy and uncertainty calibration. Interestingly, there is no unanimous tendency for version (**B**), which adds rotated split rules only, or version (**C**), which adds GPs only, to be second best; when jointly considering both RMSE and CRPS, (**C**) outperforms (**B**) in terms of CRPS at $n = 1000$. This reaffirms that combining both innovations is necessary to achieve the best performance.

Table 3.C.1: Summaries of the median RMSE and CRPS values over the 5 repetitions of 5-fold cross-validations for the $n = \{100, 500, 1000\}$ simulated data sets from the benchmarking experiments in Section 3.4.1.

Version	$n = 100$		$n = 500$		$n = 1000$	
	RMSE	CRPS	RMSE	CRPS	RMSE	CRPS
(**A**)	10.80	6.39	7.00	3.97	4.90	2.66
(**B**)	7.79	4.54	3.33	1.83	3.01	1.52
(**C**)	8.80	5.11	4.10	2.08	3.37	1.48
(**D**)	5.35	3.15	2.65	1.10	2.61	0.83

Finally, we present the MH acceptance rates of the newly proposed moves used for learning the tree structures under the standard GP-BART (**D**). Table 3.C.2 shows the proportion of new trees that were accepted after the burn-in phase using each

of the three available moves for each simulated data set, over all 25 folds in total. The acceptance rates of the novel grow-rotate and change-rotate moves highlight their effectiveness.

Table 3.C.2: Acceptance rates for the tree-proposal moves available under GP-BART for the three simulated data sets, obtained by dividing the number of times the given move was accepted by the total number of trees across all 25 folds in all retained posterior samples.

Move	$n = 100$	$n = 500$	$n = 1000$
grow-rotate	0.107	0.052	0.037
change-rotate	0.218	0.053	0.031
prune	0.109	0.056	0.038

3.D Examining the effects of the hyperparameters of the tree prior

The choice of the tree hyperparameters α and β from the tree prior in Equation (3.4) controls the depth of the trees which compose the ensemble. The default choice is $\alpha = 0.95$ and $\beta = 2$, which tends to favour shallow trees. In the GP-BART context, it would appear to be of interest to consider alternative hyperparameter specifications, in order to encourage deeper trees with fewer observations in each terminal node, given the computational complexity of $\mathcal{O}(n_{t\ell}^3)$ per node. However, we show here that doing so comes at the expense of worse predictive performance.

To evaluate the joint effect of alternative specifications of α and β on the computational cost and the accuracy of the predictions, we conducted an experiment using data generated via the Friedman equation (Friedman, 1991); specifically, we use the same data from Section 3.4.2 with $p = 10$ predictors, of which five are additional noise variables, as an example. In this case, GP-BART was trained with $n_{\text{train}} = 500$ and evaluated with $n_{\text{test}} = 500$. The tree parameters were evaluated over a discrete grid of $\alpha = \{0.1, 0.5, 0.95, 0.99\}$ and $\beta = \{1, 2, 5\}$. All possible combinations of these parameters were evaluated, constituting a total of 12 different scenarios. All other parameters were set to their default values. The outcomes are summarised in Figure 3.D.1, in the form of relative run times and RMSE values.

The run time of each setting is given relative to the time taken under the defaults of $\alpha = 0.95$ and $\beta = 2$.

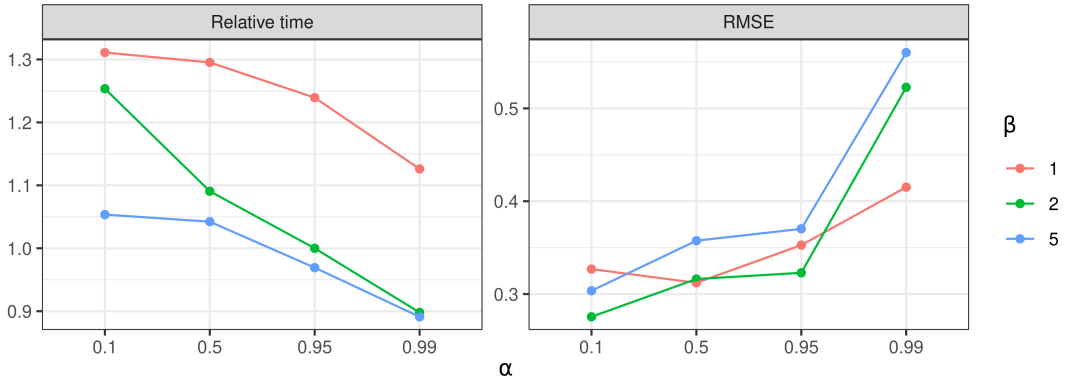


Figure 3.D.1: Performance assessment for Friedman data with noise variables and $n = 500$, over a range of α and β values in the tree prior, in terms of run time (relative to the default parameterisation of $\alpha = 0.95$ and $\beta = 2$) and RMSE.

From these results, it is evident that setting priors which favour more splits can reduce the computational cost of the model. As the cost of each matrix inversion is at the scale of $\mathcal{O}(n_{t\ell}^3)$, deeper trees with fewer observations in each terminal node reduces the burden of matrix inversion. However, the predictive performance diminishes due to forcing splits that should not exist. Conversely, the few settings which slightly improve the RMSE are substantially slower. Therefore, it remains sensible to adopt the default values for α and β from the standard BART are as the default for GP-BART. Indeed, we do so throughout the main body of the paper and note that changing these settings to increase the speed of computations should be done with caution as it can significantly harm predictive performance.

CHAPTER 4

Seemingly unrelated BART for cost-effectiveness analyses in healthcare

In recent years, theoretical results and simulation evidence have shown Bayesian additive regression trees to be a highly-effective method for nonparametric regression. Motivated by cost-effectiveness analyses in health economics, where interest lies in jointly modelling the costs of healthcare treatments and the associated health-related quality of life experienced by a patient, we propose a multivariate extension of BART applicable in regression and classification analyses with several correlated outcome variables. Our framework overcomes some key limitations of existing multivariate BART models by allowing each individual response to be associated with different ensembles of trees, while still handling dependencies between the outcomes. In the case of continuous outcomes, our model is essentially a nonparametric version of seemingly unrelated regression. Likewise, our proposal for binary outcomes is a nonparametric generalisation of the multivariate probit model. We give suggestions for easily interpretable prior distributions, which allow specification of both informative and uninformative priors. We provide detailed discussions of MCMC sampling methods to conduct posterior inference. Our methods are implemented in the R package `suBART`*. We showcase their performance through extensive simulations and an application to an empirical case study from health economics. By also accommodating propensity scores in a manner befitting a causal analysis, we find substantial evidence for a novel trauma care intervention's cost-effectiveness.*

4.1 Introduction

Many research questions in health economics are concerned with trading off the costs and benefits of a medical intervention. The most prominent examples are in *cost-effectiveness analysis* (CEA), where we wish to decide whether a new innovative treatment is worth the associated increase in costs. We therefore need to estimate the average treatment effects on both costs and health. However, in order to get coherent measures of uncertainty, the two treatment effects must be estimated jointly in order to account for the correlation between them (Baio, 2012). This point is elaborated in Section 4.2. If the CEA is performed with observational data, where the treatment assignment is not randomised, we additionally have to adjust for confounding bias in the analysis.

In this chapter, we aim to estimate the cost-effectiveness of a novel treatment for physical trauma rehabilitation, called the *transmural trauma care model* (TTCM), using data gathered under a study by Wiertsema et al. (2019) which we will henceforth refer to as the *TTCM data*. The treatment assignment is not randomised and the number of potential confounders is large relative to the sample size. Of the multiple cost and effectiveness outcomes the authors investigated, we focus on healthcare-related costs and health-related quality of life. It is of interest to jointly estimate both outcomes, and reasonable to assume both that the outcomes are non-linearly related to the available predictors and that each outcome may depend on different subsets of predictors, which may in turn interact in complex ways. The challenge of CEA under these circumstances motivated the development of our novel methodology, though we also anticipate its use in other CEA studies and broader healthcare settings.

In the causal inference literature, there is wide agreement that flexible nonparametric methods, which do not impose strong parametric assumptions on the regression functions, are the best tools for estimating treatment effects with observational data (Dorie et al., 2019; Rudolph et al., 2023). It is not straightforward, however, to apply this knowledge in the context of CEAs. Seemingly unrelated regression models (SUR; Zellner, 1962), the most recommended statistical method for CEAs (Willan et al., 2004; El Alili et al., 2022), impose strong linearity assumptions,

which can bias the inferences if there are strong non-linear relationships between the variables of interest. On the other hand, there is a distinct lack of nonparametric regression methods which can handle multivariate outcomes. We adopt a Bayesian perspective and fill this gap by developing a nonparametric version of SUR. The idea is to replace the linear predictors in the SUR model by sums of regression trees. In the univariate case, this regression method has become known as *Bayesian additive regression trees* (BART). BART has already demonstrated competitive performance for univariate responses (Dorie et al., 2019; Rudolph et al., 2023) and it seems plausible that this efficacy will extend to situations with multiple outcomes of interest. However, by embedding BART in the SUR framework, we also seek to overcome some limitations of existing multivariate BART extensions.

Chipman et al. (2010) introduced the BART method as an ensemble method, where each learner is a tree following the Bayesian CART approach previously proposed by the same authors (Chipman et al., 1998). Considering a univariate response vector $\mathbf{y} \in \mathbb{R}^n$ and a set of predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$, which may be of mixed type, one of the main objectives of regression modelling is to estimate the conditional expectation $\mathbb{E}[y_i \mid \mathbf{x}_i] = f(\mathbf{x}_i)$. As a nonparametric model, BART offers high flexibility in estimating this conditional expectation. However, the propensity of decision trees to overfit is mitigated by the Bayesian underpinnings of the framework allowing informative prior distributions to impose regularisation in a principled and transparent fashion, as well as the additive nature of the ensemble. These features facilitate better generalisation, in a similar vein to the gradient boosting approach of Friedman (2001). The success of BART is widely reported in the literature across a broad spectrum of applications (Janizadeh et al., 2021; Sarti et al., 2023; Yee and Deshpande, 2023). In addition, theoretical work has demonstrated the frequentist optimality of BART under certain conditions (Linero and Yang, 2018; Ročková and Saha, 2019; Ročková and Van der Pas, 2020; Rocková, 2020).

The BART model was originally developed for univariate responses but subsequent formulations emerged to adapt BART to multivariate responses. Examples include the Bayesian additive vector autoregressive tree (Huber and Rossini, 2022) for multivariate time-series analysis and formulations by Peruzzi and Dunson (2022) for multivariate spatial data. Other applied work involves an extension of BART to

forecast the tails of multivariate responses (Clark et al., 2023). Um et al. (2023) adapted BART to cover not only multivariate responses but also the assumption of a skew-normal distribution; throughout this chapter, we refer to the variant without skewness as mvBART. Furthermore, McJames et al. (2023) proposed an extension of Bayesian causal forests (BCF; Hahn et al., 2020)) for multivariate responses. All aforementioned approaches share the limitation that the tree structure must be identical for each component of the outcome vector. It is easy to envisage situations for which this is inappropriate: a specific covariate may be strongly associated with one outcome but independent of another. For example, factors which govern costs may be unrelated to quality of life, and *vice versa*.

In this study, we propose a novel variant of BART termed seemingly unrelated BART (suBART) which is designed to handle multivariate continuous responses and address this key limitation. Chipman et al. (2010) previously drew parallels to SUR models (Zellner, 1962) and alluded to the potential extension of BART in this direction. Our framework differs from the aforementioned multivariate BART extensions, which assume a single set of trees with correlated multivariate Gaussian distributions in the terminal nodes. Instead, we jointly fit individual ensembles of trees and model the interdependence of the outcomes through correlated error terms. Thus, suBART also differs from merely applying entirely separate univariate BART models to each outcome. Motivated by our investigation into the cost-effectiveness of the TTCM intervention, we further extend suBART to incorporate propensity scores, in the spirit of Hahn et al. (2020), as befits causal analyses. Beyond CEA settings, we also develop probit suBART, an extension of suBART to accommodate multivariate binary outcomes. We envision this version of the model being useful in economic applications with correlated binary outcomes; see Ramful and Zhao (2009) as an example. Our approach is similar to that of Chakraborty (2016), who presented a version of seemingly unrelated BART for exclusively continuous outcomes with an adaptive number of trees. However, it is not specifically tailored to causal inference objectives typical of CEAs and lacks an available open-source software implementation. We address these gaps by providing a comprehensive framework, which covers either continuous or binary outcomes, and a practical implementation through an R named `suBART`, which is available at `https://github.com/MateusMaiaDS/suBART`.

This chapter proceeds as follows: Section 4.2 provides background theory on CEA and motivates the development of the suBART model in the context of the application to the TTCM data from Wiertsema et al. (2019). Section 4.3 then elaborates on the theoretical underpinnings of the suBART methodology and Section 4.4 discusses the posterior inference for both the multivariate continuous and multivariate binary outcome settings. Section 4.5 considers different simulation scenarios and discusses the performance of both suBART models compared with standard competitors. The empirical findings of our application of suBART to the TTCM data are presented in Section 4.6. Finally, Section 4.7 summarises the proposed methodologies, highlighting both their limitations and potential for further extension, and presents conclusions regarding the health economic application. Additional results, comparisons, and findings are included in Appendix 4.A.

4.2 CEA and the suBART model

We now provide more detail about the CEA setting which inspired the suBART model and review some relevant ideas from health economics and causal inference. Detailed treatments can be found in Gabrio et al. (2019) and Li et al. (2023). We defer a description of the specific TTCM data to which we apply suBART to Section 4.6.

A major motivation to develop the suBART method was its potential applicability in cost-effectiveness analyses of healthcare treatments. Such analyses are usually performed with data from clinical trials, but there is increasing interest in the analysis of observational data, where the treatment assignment is not randomised. There is broad consensus among epidemiologists that simple parametric models often lead to severe bias and that flexible nonparametric models are preferable for the analysis of observational data (Hernán and Robins, 2024). It seems reasonable to assume that this would extend to the setting of cost-effectiveness analysis, where we want to infer two treatment effects simultaneously. There is, however, a lack of statistical methods fit for these purposes. Given that BART has proven to be very useful for causal inference in the univariate setting (Hill, 2011; Dorie et al., 2019; Rudolph et al., 2023), we consider it a promising method for the multivariate cost-effectiveness setting.

The fundamental problem of cost-effectiveness analyses in health economics is to determine which of two competing healthcare treatments — usually, but not always, for the same disease — should be implemented. Often one is both more effective and more expensive than the other, which raises the question whether the increase in health is worth the added expenses. Henceforth, we let c_i and q_i respectively denote the healthcare costs and the health-related quality of life associated with a patient. We also suppose that there are two different treatments of interest, $t = 0$ and $t = 1$. Using the usual potential outcomes notation, we let $c_i(t)$ denote the costs associated with patient i, had they received treatment t. We do likewise for $q_i(t)$. We furthermore suppose that there is some vector of baseline characteristics $\mathbf{x}_i$ which have an effect on both the outcomes c_i and q_i, as well as the treatment indicator t.

Given a sample of n observations, we wish to estimate the mixed average treatment effect (MATE)[1] on the costs in this sample, which we define as

$$\Delta_c \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[c_i(1) \mid \mathbf{x}_i\right] - \mathbb{E}\left[c_i(0) \mid \mathbf{x}_i\right]. \tag{4.1}$$

We now make the assumption of *ignorability*, which means that conditional on the baseline covariates $\mathbf{x}_i$, the treatment t is independent of the potential outcomes $c_i(0)$ and $c_i(1)$. Under this assumption, we may rewrite our treatment effect as

$$\Delta_c = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[c_i \mid t = 1, \mathbf{x}_i\right] - \mathbb{E}\left[c_i \mid t = 0, \mathbf{x}_i\right].$$

It follows that Δ_c is completely specified by the conditional expectations $\mathbb{E}[c_i|t, \mathbf{x}_i]$. We proceed in the same manner for Δ_q, the MATE on the patient's quality of life. It is then customary to combine the two treatment effects into a utility function, the incremental net benefit (INB):

$$\text{INB}_\lambda \coloneqq \lambda\Delta_q - \Delta_c.$$

[1]The MATE is closely related to the population average treatment effect (PATE), although the terminology for treatment effects is not consistent across the literature. We elect to use the terminology of Li et al. (2023), according to whom the PATE requires a generative model for the covariates $\mathbf{x}_i$ and thus necessitates additional assumptions. We work with the MATE in Equation (4.1) to simplify the analyses. Note that this common approach actually corresponds to what is called the PATE by Imbens and Rubin (2015) and other authors.

The scalar parameter λ is called the *willingness-to-pay*. Roughly speaking, λ quantifies how much cost (in the given currency) a decision-maker is willing to trade for a one-unit increase in healthcare-related quality of life for one patient. The decision rule is then simple: if the INB is at most zero, we say that treatment 1 is not cost-effective, and treatment 0 should be implemented. If the INB is larger than zero, we consider treatment 1 to be cost-effective and worthy of being implemented.

To illustrate the importance of modelling the two outcomes c and q jointly, let us assume for simplicity that the joint distribution of Δ_c and Δ_q is bivariate normal. Then

$$\begin{aligned}\Pr\left(\mathrm{INB}_\lambda > 0\right) &= \Phi\left(\frac{\mathbb{E}\left[\mathrm{INB}_\lambda\right]}{\sqrt{\mathrm{Var}\left[\mathrm{INB}_\lambda\right]}}\right) \\ &= \Phi\left(\frac{\lambda\mathbb{E}\left[\Delta_q\right] - \mathbb{E}\left[\Delta_c\right]}{\left(\lambda^2\mathrm{Var}\left[\Delta_q\right] + \mathrm{Var}\left[\Delta_c\right] - 2\lambda\mathrm{Cov}\left[\Delta_q, \Delta_c\right]\right)^{1/2}}\right).\end{aligned}$$

Consequently, the probability of cost-effectiveness depends on the covariance of Δ_c and Δ_q. Without the normality assumption, this probability can usually not be found explicitly, but the same principle applies nonetheless: the probability of cost-effectiveness depends on the joint distribution of Δ_c and Δ_q (Löthgren and Zethraeus, 2000; Gabrio et al., 2019). It follows that we must model c and q jointly, as modelling them separately would enforce the unrealistic prior belief that the treatment effects Δ_c and Δ_q are independent. This belief is seldom appropriate, since empirical cost and health data are often strongly correlated (Willan et al., 2004).

We hence use the suBART model developed below to jointly estimate the conditional expectations $\mathbb{E}[c \mid t, \mathbf{x}_i]$ and $\mathbb{E}[q \mid t, \mathbf{x}_i]$. The treatment effects and INB can then be obtained as functions of these estimates. Our approach mirrors that of Hahn et al. (2020): we first estimate propensity scores (using probit BART), and then condition the suBART model on all covariates, the treatment indicator, and the estimated propensity scores.

4.3 The suBART models for continuous and binary responses

We start by reviewing the original BART model in the univariate setting in Section 4.3.1, in order to provide context for what is to follow. We then present our novel extensions to the multivariate continuous outcome setting in Section 4.3.2 and the multivariate binary outcome setting in Section 4.3.3. Specific details regarding posterior inference for the suBART models are deferred to Section 4.4.

4.3.1 A review of univariate BART

BART was designed to solve the classic regression problem of the form

$$y_i = \mathbb{E}\left[y_i \mid \mathbf{x}_i\right] + \varepsilon_i,$$

where y_i is a univariate response variable for observation $i = 1, \ldots, n$, $\mathbf{x}_i$ is a $p-$dimensional predictor, and $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. The idea is to find a flexible approximation for the conditional expectation $\mathbb{E}[y_i \mid \mathbf{x}_i]$ by expressing it as a sum of regression trees. A regression tree consists of two components:

1. A binary tree $\mathcal{T}$, which defines a finite partition $\{\mathcal{A}_1, \ldots, \mathcal{A}_h\}$ of $\mathbb{R}^p$ based on the feature space of $\mathbf{X}$, using the available predictors or a subset thereof to form splitting rules. In other words, $\mathcal{A}_1, \ldots, \mathcal{A}_h$ are subsets of $\mathbb{R}^p$ such that any $\mathbf{x}_i \in \mathbb{R}^p$ is contained in exactly one $\mathcal{A}_\ell$.

2. A collection of scalar parameters $\mathcal{M} = (\mu_1, \ldots, \mu_h)$, called leaf nodes, with each component being associated with the corresponding subset in the partition.

We now define a function $\mathbf{x}_i, \mathcal{T}, \mathcal{M} \mapsto g(\mathbf{x}_i, \mathcal{T}, \mathcal{M})$ as follows: $\mathbf{x}_i \in \mathcal{A}_\ell$ for exactly one ℓ; then $g(\mathbf{x}_i, \mathcal{T}, \mathcal{M}) \coloneqq \mu_\ell$. In the case of a single regression tree, we may then define the regression function $\mathbf{x}_i \mapsto \mathbb{E}[y_i \mid \mathbf{x}_i]$ by $\mathbb{E}[y_i \mid \mathbf{x}_i] \coloneqq g(\mathbf{x}_i, \mathcal{T}, \mathcal{M})$. Figure 4.3.1 shows an illustration of a simple regression tree, including the regression function it implies, for the case $p = 1$.

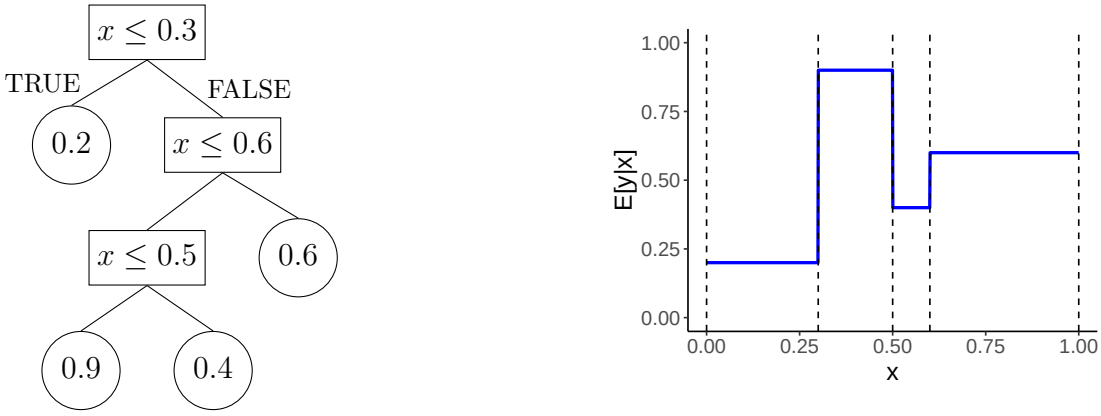


Figure 4.3.1: Regression tree (left) and implied regression function (right).

To extend this idea to *additive* regression trees, we consider not just one tree, but multiple trees $\mathcal{T}_1, \ldots, \mathcal{T}_m$, each with their own corresponding partitions and leaf nodes. Then we let $\mathbb{E}[y_i \,|\, \mathbf{x}_i] \coloneqq \sum_{t=1}^{m} g(\mathbf{x}_i, \mathcal{T}_t, \mathcal{M}_t)$. The definition of the statistical model now becomes

$$y_i = \sum_{t=1}^{m} g\left(\mathbf{x}_i, \mathcal{T}_t, \mathcal{M}_t\right) + \varepsilon_i,$$

which is the basic BART model presented in Chipman et al. (2010). The model is typically not identified, since different sets of trees can lead to the same regression function. However, this is not a problem, since the individual trees are rarely of direct interest.

The sum of trees framework can also be used to model the conditional expectation of a binary response $\mathbf{y}$, which takes values in $\{0, 1\}$. This is the probit BART model, again proposed originally in Chipman et al. (1998). The model is easier to present and analyse when it is cast in terms of a continuous latent variable. Suppose $\mathbf{z}$ is such that

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

As before, we model z_i as

$$z_i = \sum_{t=1}^{m} g\left(\mathbf{x}_i, \mathcal{T}_t, \mathcal{M}_t\right) + \varepsilon_i$$

with $\varepsilon_i \sim \mathrm{N}(0, 1)$. This then implies that

$$\mathbb{E}\left[y_i \mid \mathbf{x}_i\right] = \Pr\left(y_i = 1 \mid \mathbf{x}_i\right) = \Pr\left(z_i > 0 \mid \mathbf{x}\right) = \Phi\left(\sum_{t=1}^{m} g\left(\mathbf{x}_i, \mathcal{T}_t, \mathcal{M}_t\right)\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

For both types of outcome, given a sample of size n and a univariate outcome vector $\mathbf{y} \in \mathbb{R}^n$ associated with a predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, we are interested in sampling from the joint posterior distribution $\pi(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}} \,|\, \mathbf{y}, \mathbf{X})$ where $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_m)$ denotes the collection of all trees and $\boldsymbol{\mathcal{M}} = (\mathcal{M}_1, \ldots, \mathcal{M}_m)$ denotes their corresponding mean parameters. To obtain the posterior, it is necessary to define priors for both the trees and the terminal node parameters. Assuming the independence of leaf parameters conditional on the tree structures, Chipman et al. (1998) defines the joint prior distribution as

$$\begin{aligned}
\pi\left(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}}, \sigma^2\right) &= \left[\prod_{t=1}^{m} \pi\left(\mathcal{T}_t, \mathcal{M}_t\right)\right] \times \pi\left(\sigma^2\right) \\
&= \left[\prod_{t=1}^{m} \pi\left(\mathcal{M}_t \mid \mathcal{T}_t\right) \times \pi\left(\mathcal{T}_t\right)\right] \times \pi\left(\sigma^2\right) \\
&= \left[\prod_{t=1}^{m} \prod_{\ell=1}^{b_m} \pi\left(\mu_{t\ell} \mid \mathcal{T}_t\right) \times \pi\left(\mathcal{T}_t\right)\right] \times \pi\left(\sigma^2\right).
\end{aligned}$$

To achieve conjugacy, it is typically assumed that the residual variance parameter σ^2 follows an inverse-gamma distribution and $\mu_{t\ell} \sim N(0, \sigma_\mu^2)$. with $\sigma_\mu^2 = \frac{0.25}{\kappa^2 m}$ being proportional to the number of trees in order to regularise the contribution of each tree. The definition of $\pi(T_t)$ includes specifying the probability of a non-terminal node as $\alpha(1 + \gamma_{t\ell})^{-\beta}$ where $\gamma_{t\ell}$ denotes the depth of that node. The hyperparameters α and β take the default values suggested in Chipman et al. (2010) of 0.95 and 2, respectively, to favour shallow trees.

Once the prior is defined, a sampler for the aforementioned posterior distribution can be obtained. Referring to $\boldsymbol{\mathcal{T}}_{(-t)} \coloneqq \boldsymbol{\mathcal{T}} \setminus \{\mathcal{T}_t\}$ and $\boldsymbol{\mathcal{M}}_{(-t)} \coloneqq \boldsymbol{\mathcal{M}} \setminus \{\mathcal{M}_t\}$, an MCMC sampler can be built by sequentially sampling $\pi(\mathcal{T}_t | \boldsymbol{\mathcal{T}}_{(-t)}, \boldsymbol{\mathcal{M}}, \mathbf{y}, \mathbf{X}, \sigma^2)$ and $\pi(\mathcal{M}_t \,|\, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}}_{(-t)}, \mathbf{y}, \mathbf{X}, \sigma^2)$. It can be shown that $\mathcal{T}_t$ and $\mathcal{M}_t$ depend on $(\boldsymbol{\mathcal{T}}_{(-t)}, \mathbf{y})$ only through the partial residuals $\mathbf{r}_t \coloneqq \mathbf{y} - \sum_{j \neq t}^{m} g(\mathbf{X}, \mathcal{T}_j, \mathcal{M}_j)$. This fact can be

used to construct a Bayesian back-fitting algorithm (Hastie and Tibshirani, 2000). Therefore, the successive draws become

$$\pi\left(\mathcal{T}_t \mid \mathbf{r}_t, \sigma^2\right)$$
$$\pi\left(\mathcal{M}_t \mid \mathcal{T}_t, \mathbf{r}_t, \sigma^2\right),$$

where new trees and splitting rules are sampled through a Metropolis-Hastings step calculated using the integrated-likelihood for the tree $\mathcal{T}_t$ over the leaf parameters $\mathcal{M}_t$. See Chipman et al. (2010) and Kapelner and Bleich (2016) for further details of the model and additional information the algorithmic implementation on which that of the suBART models is based.

4.3.2 The suBART model for continuous outcomes

We consider a regression problem of the form

$$\begin{pmatrix} y_i^{(1)} \\ \vdots \\ y_i^{(d)} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[y_i^{(1)} \mid \mathbf{x}_i\right] \\ \vdots \\ \mathbb{E}\left[y_i^{(d)} \mid \mathbf{x}_i\right] \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{(1)} \\ \vdots \\ \varepsilon_i^{(d)} \end{pmatrix} \tag{4.3}$$

where $\mathbf{y}^{(j)}$ represents the j-th component of a d-variate outcome, and $(\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(d)})^\top \sim \text{MVN}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$. In principle, different models can be used for each conditional expectation in Equation (4.3). If the conditional expectations are all assumed to be linear in $\mathbf{x}_i$, we obtain the classic SUR model (Zellner, 1962). We instead want to allow the possibility that the conditional expectations are non-linear. Given that the BART model has been shown to be a viable model in the one-dimensional setting, it seems reasonable to expect this viability to extend to the multivariate case. We therefore proceed by assigning an ensemble of regression trees to each $\mathbb{E}[y_i^{(j)} \mid \mathbf{x}_i]$ as follows

$$\begin{pmatrix} y_i^{(1)} \\ \vdots \\ y_i^{(d)} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^m g\left(\mathbf{x}_i, \mathcal{T}_t^{(1)}, \mathcal{M}_t^{(1)}\right) \\ \vdots \\ \sum_{t=1}^m g\left(\mathbf{x}_i, \mathcal{T}_t^{(d)}, \mathcal{M}_t^{(d)}\right) \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{(1)} \\ \vdots \\ \varepsilon_i^{(d)} \end{pmatrix}, \tag{4.4}$$

where

- $\mathcal{T}_t^{(j)}$ is a binary tree which defines a finite partition $\{\mathcal{A}_{t\ell}^{(j)} : 1 \leq \ell \leq h_t^{(l)}\}$ of $\mathbb{R}^p$. Note that $h_t^{(j)}$ is the number of leaf nodes of the tree $\mathcal{T}_t^{(j)}$. The collection of all trees pertaining to the j-th outcome is denoted by $\boldsymbol{\mathcal{T}}^{(j)} = (\mathcal{T}_1^{(j)}, \ldots, \mathcal{T}_m^{(j)})$.
- $\mathcal{M}_t^{(j)} = (\mu_{t1}^{(j)}, \ldots, \mu_{th_t^{(j)}}^{(j)})$ is the vector of leaf parameters associated with the tree $\mathcal{T}_t^{(j)}$. Similarly, the collection of all leaf parameters associated with the j-th outcome is denoted by $\boldsymbol{\mathcal{M}}^{(j)}$.
- $(\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(d)})^\top \sim \text{MVN}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ being a $d \times d$ covariance matrix. We write $\sigma_j^2 \coloneqq \Sigma_{jj}$ for the diagonal elements, such that σ_j^2 is the variance of the error term $\boldsymbol{\varepsilon}^{(j)}$. We further write $\rho_{jk} \coloneqq \text{Cor}(\boldsymbol{\varepsilon}^{(j)}, \boldsymbol{\varepsilon}^{(k)}) \, \forall \, j \neq k$. Note that for any j, k, we have

$$\rho_{jk} = \frac{\text{Cov}\left[\boldsymbol{\varepsilon}^{(j)}, \boldsymbol{\varepsilon}^{(k)}\right]}{\sqrt{\text{Var}\left[\boldsymbol{\varepsilon}^{(j)}\right] \text{Var}\left[\boldsymbol{\varepsilon}^{(k)}\right]}} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}} = \frac{\Sigma_{jk}}{\sigma_j \sigma_k}.$$

The model setup is straightforward, and indeed similar to the original BART model in Chipman et al. (2010). Our model is comprised of d univariate BART models, which are linked through the correlated error terms $\boldsymbol{\varepsilon}^{(j)}$. For simplicity, we assume that the number of trees m is common across all d outcomes. However, we stress that each submodel has its *own* separate collection of m trees. Thus, there are $d \times m$ trees in total. This is a key difference compared to other multivariate BART versions McJames et al. (2023); Um et al. (2023), which assume that the trees are the same for all outcomes, i.e., that there are only m trees in which the leaf parameters associated with each tree are vectors $\boldsymbol{\mu}_{t\ell}$.

A consequence of this assumption in existing multivariate BART versions is the implication that the p-dimensional predictor $\mathbf{x}_i$ is the same for all d outcomes. However, in systems of equations such as (4.3) and (4.4), the set of predictors $\mathbf{x}_i$ need not be of common dimension p for each outcome. Indeed, our `suBART` software implementation allows different subsets of the predictors to be used for each constituent univariate BART model. Though we will henceforth assume that the predictors are the same for all outcomes, for simplicity, it is important to note that having the same set of predictors $\mathbf{x}_i$ available as candidates to form splitting rules in each set of trees does *not* imply that the trees for different outcomes will

split on the same components of $\mathbf{x}_i$ at the same cutoff values. Unlike the standard SUR, imposing restrictions on the $\mathbf{x}_i$ for different outcomes is not required to ensure that different outcomes depend on different covariates under suBART, as different sets of trees will tend to form different partitions on different subsets of the feature space $\mathbf{X}$ anyway, by the inherent nature of the tree-generating process in BART. Although practitioners can impose such restrictions nonetheless, to strictly guarantee that the trees for a given outcome do not depend on certain predictors, this is an appealing property, in the sense that suBART minimises the need to pre-specify the parametric form of the models for each conditional expectation.

We assume that the m regression trees for each outcome are all independent of each other and of the covariance matrix *a priori*, as per Chipman et al. (2010); i.e,

$$\begin{aligned}\pi\left(\left(\boldsymbol{\mathcal{T}}^{(1)}, \boldsymbol{\mathcal{M}}^{(1)}\right), \ldots, \left(\boldsymbol{\mathcal{T}}^{(d)}, \boldsymbol{\mathcal{M}}^{(d)}\right), \boldsymbol{\Sigma}\right) &= \left(\prod_{t=1}^{m}\prod_{j=1}^{d} \pi\left(\mathcal{T}_t^{(j)}, \mathcal{M}_t^{(j)}\right)\right) \times \pi\left(\boldsymbol{\Sigma}\right) \\ &= \left(\prod_{t=1}^{m}\prod_{j=1}^{d} \pi\left(\mathcal{M}_t^{(j)} \mid \mathcal{T}_t^{(j)}\right) \pi\left(\mathcal{T}_t^{(j)}\right)\right) \times \pi\left(\boldsymbol{\Sigma}\right).\end{aligned}$$

We further assume that the leaves of a tree are conditionally independent, given the tree structure, i.e.,

$$\pi\left(\mathcal{M}_t^{(j)} \mid \mathcal{T}_t^{(j)}\right) = \prod_{\ell=1}^{h_t^{(d)}} \pi\left(\mu_{t\ell}^{(j)} \mid \mathcal{T}_j^{(d)}\right).$$

With this setup, prior distributions for $\mathcal{T}_t^{(j)}, \mu_{t\ell}^{(j)}$, and $\boldsymbol{\Sigma}$ are sufficient to specify the joint prior distribution for all model parameters. The tree structure $\mathcal{T}_t^{(j)}$ is assigned the same prior as in the original work by Chipman et al. (2010), with default hyperparameters $\alpha = 0.95$ and $\beta = 2$ to favour shallow trees and avoid over-fitting.

The prior used for the leaf node parameters is also aligned with the approach of standard BART. As noted earlier, this prior is formulated conditionally on the tree $\mathcal{T}_t^{(j)}$. We assume that each outcome component $\mathbf{y}^{(j)}$ is re-scaled such that $y_i^{(j)} \in [-0.5, 0.5]$. This enables the model to specify, with defined probability, that the implicit prior for $\mathbb{E}[y_i^{(j)} \,|\, \mathbf{x}_i]$ lies within the rescaled interval. Consequently, the prior is then

$$\mu_{t\ell}^{(j)} \mid \mathcal{T}_t^{(j)} \sim \mathrm{N}\left(0, {\sigma_\mu^{(j)}}^2\right), \tag{4.5}$$

where ${\sigma_\mu^{(j)}}^2 = \frac{0.25}{\kappa^2 m}$, as before. We suggest $\kappa = 2$ as a default choice, which assigns a prior probability of 0.95 to the event $\{\mathbb{E}[y_i^{(j)} \mid \mathbf{x}_i] \in [-0.5, 0.5]\}$.

In the univariate BART model, the error variance is assigned an inverse-gamma prior, which is conditionally conjugate and can thus be easily incorporated into a Gibbs sampler. Furthermore, by choosing the hyperparameters accordingly, it is possible to put an informative prior on the error variance. The usual approach is as follows: suppose that for the outcome $\mathbf{y}$, we have a 'data-based overestimate' $\hat{\sigma}^2$ of the error variance σ^2 (for example, the sample variance of the observed $\mathbf{y}$ values). Presumably, the true value of σ^2 is smaller than $\hat{\sigma}^2$, since the variation of $\mathbf{y}$ is partly explained by the covariates $\mathbf{X}$. Therefore, we would like to assign a large prior probability to the event $\{\sigma^2 < \hat{\sigma}^2\}$ (for example, 0.95).

We now wish to generalise this idea to multivariate settings: for all components j of the outcome vector $\left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(d)}\right)$, we have an overestimate $\hat{\sigma}_j^2$ of σ_j^2, and want to assign some probability p to the event $\{\sigma_j^2 < \hat{\sigma}_j^2\}$. Additionally, we also would like to control the prior on the correlations ρ_{jk}, where $j \neq k$. Since there is usually not strong prior information about these correlations, it is preferable for the prior to not be too informative in any direction. Additionally, the chosen solution should be computationally tractable and readily incorporated into MCMC samplers.

The inverse Wishart distribution is often used as a prior for covariance matrices. As it is a straightforward multivariate generalisation of the inverse-gamma distribution, it facilitates easy computations. Unfortunately, the inverse-Wishart prior is not well-suited to meet our aforementioned goals; among other problems, it imposes a strong prior dependency between the variances and the correlations. Consequently, it is generally not feasible to choose the hyperparameters such that the prior has the desired properties for both the variances and the correlations. See Alvarez et al. (2014) for a detailed study of this issue.

We thus instead adapt an approach by Huang and Wand (2013) and parameterise the covariance matrix $\mathbf{\Sigma}$ as follows:

- $a_j \sim \text{Inv-Gamma}(1/2, 1/A_j^2)$, where $A_j > 0$ is a fixed hyperparameter.
- $\boldsymbol{\Sigma}|a_1, \ldots, a_d \sim \text{Inv-Wishart}_d(\nu+d-1, \mathbf{S}_0)$, where $\mathbf{S}_0 \coloneqq 2\nu \times \text{diag}\,(1/a_1, \ldots, 1/a_d)$.

The implied prior distribution for the correlations can be derived (Huang and Wand, 2013) and is given by

$$\pi\left(\rho_{jk}\right) = \left(1 - \rho_{jk}^2\right)^{\frac{\nu}{2}-1}, \quad \rho_{jk} \in (-1, 1)\,. \tag{4.6}$$

Crucially, the prior does not depend on $A_1, \ldots, A_d$. It is uniform if and only if $\nu = 2$. For higher values, the prior increasingly concentrates around zero. We consider $\nu = 2$ a reasonable default choice, since we usually do not have any strong prior information on the correlations. In simulation tests, it was found that this uniform prior can sometimes lead to an ill-identified posterior, consequently causing problems with the MCMC sampling. This seems to occur primarily in situations where the sample size is small, the dimension of $\mathbf{X}$ is large, and the variability of the multivariate outcome is almost entirely explained by $\mathbf{X}$. In such cases, we have found it useful to increase ν to improve sampling. It is also worth recalling that the response vector is bivariate in the motivating CEA application in Section 4.2, such that there is only one such correlation parameter.

From the previous definitions, the prior for the standard deviations is given by $\sigma_j \sim \text{Half-}t(\nu, A_j)$; see Wand et al. (2011) for more details. Since the priors for the correlations are independent of A_j, the choice of A_j remains arbitrary. We can thus tweak it to enforce the prior probability $\Pr(\sigma_j < \hat{\sigma}_j) = \alpha_\sigma$, in accordance with the standard BART approach. To do this, we set up the following equation

$$\alpha_\sigma = 2\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi A_j^2}} \int_0^{\hat{\sigma}_j} \left(1 + \frac{x^2}{\nu A_j^2}\right)^{-\frac{\nu+1}{2}} \mathrm{d}x \tag{4.7}$$

and solve it for A_j, which can be done through numerical root-finding. This expression is the cumulative distribution function of a Half-t-distributed random variable with degrees of freedom ν, scale parameter A_j, and support on $[0, \infty)$. By rewriting Equation (4.7) in terms of regularised incomplete beta functions, it can be shown that this expression is continuous as well as strictly decreasing in A_j and approaches 1 and 0 as A_j approaches 0 and ∞, respectively. Thus, the solution for A_j exists and is unique.

4.3.3 Probit suBART

While the previous model was designed to jointly model multiple continuous outcome variables, we now turn our attention to binary outcomes. We will present a generalisation of the linear multivariate probit model (Chib and Greenberg, 1998), where the linear predictors are replaced by sums of regression trees. Alternatively, it can also be seen as a multivariate generalisation of the probit BART model.

Suppose that we have some predictor variables $\mathbf{X}$, and a binary outcome vector $\mathbf{y}_i^{(j)} \in \{0,1\}\ \forall\ j = 1,\dots,d$, whose dependence on $\mathbf{x}_i$ we want to model. In particular, we do not want to assume that the components of $\mathbf{y}_i$ are conditionally independent, given $\mathbf{x}_i$; there may be some leftover correlation which is not explained by $\mathbf{x}_i$. As in the basic probit BART model, we cast the multivariate version in terms of latent variables $\mathbf{z}_i = (z_i^{(1)},\dots,z_i^{(d)})$; the construction is exactly as per Equation (4.2) for each outcome j. Then, the probit suBART model is given by

$$\begin{pmatrix} z_i^{(1)} \\ \vdots \\ z_i^{(d)} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^m g\left(\mathbf{x}_i, \mathcal{T}_t^{(1)}, \mathcal{M}_t^{(1)}\right) \\ \vdots \\ \sum_{t=1}^m g\left(\mathbf{x}_i, \mathcal{T}_t^{(d)}, \mathcal{M}_t^{(d)}\right) \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{(1)} \\ \vdots \\ \varepsilon_i^{(d)} \end{pmatrix}, \tag{4.8}$$

where

- $g(\cdot)$, $\mathcal{T}_t^{(j)}$, and $\mathcal{M}_t^{(j)}$ are defined as before.
- $(\varepsilon_i^{(1)},\dots,\varepsilon_i^{(d)})^\top \sim \mathrm{MVN}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ being a $d \times d$ correlation matrix. Again writing $\rho_{jk} \coloneqq \mathrm{Corr}(\boldsymbol{\varepsilon}^{(j)}, \boldsymbol{\varepsilon}^{(k)})$ for $j \neq k$, we have

$$\boldsymbol{\Sigma}_{jk} = \begin{cases} 1 & \text{if } j = k \\ \rho_{jk} & \text{if } j \neq k. \end{cases}$$

Conditional on the latent variables $\mathbf{z}_i$, the model is essentially the same as the suBART model presented earlier. The only difference is that for each error term $\boldsymbol{\varepsilon}^{(j)}$, we fix the variance at 1. Without doing this, the model would be unidentified. This issue is not specific to probit suBART, but also arises in the linear multivariate probit model. See Chib and Greenberg (1998) for details. It follows that $\boldsymbol{\Sigma}$, the covariance matrix of the error terms, is equal to 1 in each diagonal entry, and hence must be a correlation matrix.

The priors for the trees are exactly same as those for the suBART model. The dependence structure of the trees, leaves, and covariance matrix are also the same. The other priors are broadly similar, but there are some implications that are worth briefly highlighting. The prior for the terminal node parameters $\mu_{t\ell}^{(j)}$ again follows Equation (4.5), though the calibration of the variance hyperparameter requires more care. Taking ${\sigma_\mu^{(j)}}^2 = \frac{q_z^2}{\kappa^2 m}$, with $\kappa = 2$ as a default choice, we assign a prior probability of 0.95 to the event $\{\mathbb{E}[z_i^{(j)} \,|\, \mathbf{x}_i] \in [-q_z, q_z]\}$. On the probability scale, this means that $\{\Pr(y_i^{(j)} = 1 \mid \mathbf{x}_i) \in [\Phi(-q_z), \Phi(q_z)]\}$. For example, when taking $q_z = 3$ as per Chipman et al. (2010), we assign a prior probability of 0.95 to the event $\{\Pr(y_i^{(j)} = 1 \mid \mathbf{x}_i) \in [0.0013, 0.9987]\}$. This is reasonable for many applications, since extremely small or large probabilities are uncommon.

In the probit setting, $\boldsymbol{\Sigma}$ is a correlation matrix and hence must be positive definite, as well as having all diagonal entries equal to 1. These restrictions make it difficult to choose a prior for $\boldsymbol{\Sigma}$ which has desirable properties *and* facilitates easy sampling. Chib and Greenberg (1998) present a prior (and related sampling strategy) which we found to be extremely inefficient in our application. We thus instead adapt an approach by Zhang (2020) (see also Barnard et al. (2000), where some of the following results originate, and **?**). We introduce an auxiliary parameter $\mathbf{D}$, which is a $d \times d$ diagonal matrix. We then define $\mathbf{W} \coloneqq \mathbf{D}^{\frac{1}{2}} \boldsymbol{\Sigma} \mathbf{D}^{\frac{1}{2}}$, and assume the prior $\mathbf{W} \sim \text{Inv-Wishart}_d(\nu + d - 1, \mathbf{I}_d)$, where $\mathbf{I}_d$ is the d-dimensional identity matrix.

Given $\mathbf{W}$, we can recover $\mathbf{D}$ and $\boldsymbol{\Sigma}$ thanks to the identities $\mathbf{D} = \text{diag}(\mathbf{W})$ and $\boldsymbol{\Sigma} = \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}$. The induced marginal prior density of $\boldsymbol{\Sigma}$ is

$$\pi(\boldsymbol{\Sigma}) \propto (\det \boldsymbol{\Sigma})^{\frac{1}{2}(\nu+d-1)(d-1)-1} \left(\prod_{j=1}^{d} \det [\boldsymbol{\Sigma}]_{jj} \right)^{-\frac{\nu+d-1}{2}},$$

where $[\boldsymbol{\Sigma}]_{jj}$ is the j-th principle submatrix of $\boldsymbol{\Sigma}$. It can be shown that the marginal prior density for the correlations is again given by Equation (4.6), as per the suBART model for continuous outcomes above, despite the different priors assumed for $\boldsymbol{\Sigma}$. The hyperparameter ν plays a similar role as it did before; in most situations, we again consider $\nu = 2$ a reasonable default choice but reiterate that higher values of ν may lead to more stable and efficient sampling in specific scenarios.

4.4 Posterior inference

This section describes strategies and algorithmic details for conducting posterior inference under suBART and probit suBART for multivariate continuous and multivariate binary outcomes, respectively. For both frameworks, sampling is performed using a Metropolis-within-Gibbs sampler based on their respective priors and model specifications. Given an observed sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$, with $\mathbf{y}_i \in \mathbb{R}^d$, all computations are carried out conditionally on the covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$. Thus, for the sake of readability, we will consider $\mathbf{X}$ fixed and not condition on it explicitly. Additionally, the expression $\pi(\theta \mid \Theta \setminus \{\ldots\})$ denotes the conditional distribution of θ with respect to all parameters *except* θ itself and the ones listed within the braces. For example, $\pi(\mathcal{T}_t^{(j)} \mid \Theta \setminus \{\mathcal{M}_t^{(j)}\})$ refers to the distribution of tree $\mathcal{T}_t^{(j)}$ given all parameters except $\mathcal{T}_t^{(j)}$ and $\mathcal{M}_t^{(j)}$. This slightly unusual notation reduces the complexity of the expressions which follow. As we routinely condition on all but two parameters, it is clearer to highlight what is *not* being conditioned on.

4.4.1 suBART continuous

For brevity, we define the estimate for $y_i^{(j)}$ as $\hat{y}_i^{(j)} \coloneqq \sum_{t=1}^m g(\mathbf{x}_i, \mathcal{T}_t^{(j)}, \mathcal{M}_t^{(j)})$ and $\hat{\mathbf{y}}_i = (\hat{y}_i^{(1)}, \ldots, y_i^{(d)})^\top$, since there are now multiple components of $\mathbf{y}_i$. Analogously, we define the residuals from a tree t associated with the j-th component as $\mathbf{r}_t^{(j)} \coloneqq \{r_{t1}^{(j)}, \ldots, r_{tn}^{(j)}\}$, where $r_{ti}^{(j)} \coloneqq y_i^{(j)} - \sum_{k \neq t}^m g(\mathbf{x}_i, \mathcal{T}_k^{(j)}, \mathcal{M}_k^{(j)})$. As per the standard BART, we are interested in sampling from the posterior distribution

$$\pi(\Theta \mid \mathbf{Y}) = \pi\left(\left(\boldsymbol{\mathcal{T}}^{(1)}, \boldsymbol{\mathcal{M}}^{(1)}\right), \ldots, \left(\boldsymbol{\mathcal{T}}^{(d)}, \boldsymbol{\mathcal{M}}^{(d)}\right), \boldsymbol{\Sigma}, a_1, \ldots, a_d \mid \mathbf{Y}\right),$$

which, due to the back-fitting algorithm (Hastie and Tibshirani, 2000) and properties of the multivariate normal distribution, can be obtained from sequential draws from a collection of conditional distributions. In the multivariate continuous outcomes setting, we have that $\mathbf{y}_i \sim \text{MVN}_d(\hat{\mathbf{y}}_i, \boldsymbol{\Sigma})$. For the following, we will also need the conditional distribution of any component $y_i^{(j)}$ given all other components $\mathbf{y}_i^{(-j)}$. Using a well-known result (see e.g., Baldi (2024), Section 4.4), this can be found in closed form:

$$y_i^{(j)} \mid \mathbf{y}_i^{(-j)}, \Theta \sim \text{N}\left(\hat{y}_i^{(j)} + \boldsymbol{\Sigma}_{j(-j)} \boldsymbol{\Sigma}_{(-j)(-j)}^{-1} \left(\mathbf{y}_i^{(-j)} - \hat{\mathbf{y}}_i^{(-j)}\right), \Sigma_{jj} - \boldsymbol{\Sigma}_{j(-j)} \boldsymbol{\Sigma}_{(-j)(-j)}^{-1} \boldsymbol{\Sigma}_{(-j)j}\right), \tag{4.9}$$

where $\mathbf{\Sigma}_{(-j)(-j)}$ is the submatrix obtained by excluding the j-th row and column. Analogously, $\mathbf{\Sigma}_{j(-j)}$ denotes the vector obtained by selecting the j-th row and excluding the j-th column from $\mathbf{\Sigma}$. Using the result from Equation (4.9), the posterior distribution $\pi(\mathcal{T}_t^{(j)} \mid \mathbf{r}^{(j)}, \Theta \setminus \{\mathcal{M}_t^{(j)}\})$ can also be obtained in closed-form, up to a normalising constant, as the conditional distribution of the residual component $r_i^{(j)}$ given $\hat{\mathbf{y}}_i^{(-j)}$ is known. Then, as described in Section 4.3.1, the sampler for the joint posterior distribution of the trees $\mathcal{T}_t^{(j)}$ and their parameters $\mathcal{M}_t^{(j)}$, for the j-th component of $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(d)}$, is given by successive draws from

$$\mathcal{T}_t^{(j)} \mid \mathbf{r}_t^{(j)}, \Theta \setminus \left\{\mathcal{M}_t^{(j)}\right\}$$
$$\mathcal{M}_t^{(j)} \mid \mathbf{r}_t^{(j)}, \Theta.$$

Notably, the algorithm reduces to the standard BART approach when the dimension d is equal to one. Indeed, each draw above can also be viewed as univariate BART — albeit with distinct mean and variance parameters — since it is conditioned on the values of all other components in $\mathbf{Y}$, as illustrated by Equation (4.9). The full structure of the suBART sampler is given in Algorithm 4.1, but we first describe the remaining required posterior conditional distributions.

The posterior distribution for $\mu_{t\ell}^{(j)}$ is given by

$$\mu_{t\ell}^{(j)} \mid \Theta \sim \mathrm{N}\left(\left(\frac{{\sigma_\mu^{(j)}}^2}{v^{(j)} + n_{t\ell}^{(j)} {\sigma_\mu^{(j)}}^2}\right) \times \left(\sum_{i=1}^{n_{t\ell}^{(j)}} r_i^{(j)} - \sum_{i=1}^{n_{t\ell}^{(j)}} u_i^{(j)}\right), \frac{v^{(j)} {\sigma_\mu^{(j)}}^2}{v^{(j)} + n_{t\ell}^{(j)} {\sigma_\mu^{(j)}}^2}\right), \quad (4.10)$$

where $u_i^{(j)} \coloneqq \mathbf{\Sigma}_{j(-j)} \mathbf{\Sigma}_{(-j)(-j)}^{-1} (\mathbf{y}_i^{(-j)} - \hat{\mathbf{y}}_i^{(-j)})$, $v^{(j)} \coloneqq \Sigma_{jj} - \mathbf{\Sigma}_{j(-j)} \mathbf{\Sigma}_{(-j)(-j)}^{-1} \mathbf{\Sigma}_{(-j)j}$, and $n_{t\ell}^{(j)}$ denotes the number of observations in the given terminal node. It is evident from Equation (4.10) that the term $u_i^{(j)}$ vanishes and $v^{(j)} = \sigma_\mu^2$ when $d = 1$, yielding in an expression identical to the original BART formulation. Due to the conditional conjugacy in the construction of Huang and Wand (2013), the conditional posteriors of the auxiliary parameters $a_1, \ldots, a_d$ and $\mathbf{\Sigma}$ take simple forms. We have

$$a_j \mid \Theta \sim \text{Inv-Gamma}\left(\frac{\nu + n}{2}, \frac{1}{A_j^2} + \nu \left(\mathbf{\Sigma}^{-1}\right)_{jj}\right), \quad (4.11)$$

where $(\mathbf{\Sigma}^{-1})_{jj}$ denotes the j-th entry along the diagonal of $\mathbf{\Sigma}^{-1}$, and

$$\boldsymbol{\Sigma} \mid \Theta \sim \text{Inv-Wishart}_d \left(\nu + d - 1 + n, \mathbf{S}_0 + \mathbf{S}\right), \tag{4.12}$$

where $\mathbf{S} = \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^\top$.

Algorithm 4.1: suBART sampling algorithm.

Input: $\mathbf{X}$, $\mathbf{Y}$, m, N_{MCMC}, $N_{\text{burn-in}}$, and all hyper-parameters of the priors.

Initialise: $\boldsymbol{\mathcal{T}}^{(1)}, \ldots, \boldsymbol{\mathcal{T}}^{(d)}$ tree stumps, $\boldsymbol{\Sigma}$, $\mu_{t\ell}^{(j)} = 0 \; \forall \, (t, j)$.

for *iterations* h *from* 1 *to* N_{MCMC} **do**
  for *dimension* j *from* 1 *to* d **do**
    for *trees* t *from* 1 *to* m **do**
      Calculate the partial residuals $\mathbf{r}_t^{(j)}$;
      Propose a new tree $\mathcal{T}_t^{(j)\star}$ by a grow, prune, or change move[a];
      Accept and update $\mathcal{T}_t^{(j)} = \mathcal{T}_t^{(j)\star}$ with probability

$$\gamma^\star\left(\mathcal{T}_t^{(j)}, \mathcal{T}_t^{(j)\star}\right) = \min\left\{1, \frac{\pi\left(\mathbf{r}_t^{(j)} \mid \mathcal{T}_t^{(j)\star}, \Theta \setminus \left\{M_t^{(j)}\right\}\right) \pi\left(\mathcal{T}_t^{(j)\star}\right) q\left(\mathcal{T}_t^{(j)\star} \to \mathcal{T}_t^{(j)}\right)}{\pi\left(\mathbf{r}_t^{(j)} \mid \mathcal{T}_t^{(j)}, \Theta \setminus \left\{M_t^{(j)}\right\}\right) \pi\left(\mathcal{T}_t^{(j)}\right) q\left(\mathcal{T}_t^{(j)} \to \mathcal{T}_t^{(j)\star}\right)}\right\}.$$

      for *terminal nodes* ℓ *from* 1 *to* $b_t^{(j)}$ **do**
        Update $\mu_{t\ell}^{(j)} \mid \mathbf{r}_t^{(j)}, \Theta$ using Equation (4.10).
      end
    end
  end
  for j *from* 1 *to* d **do**
    Update $a_j \mid \Theta$ using Equation (4.11).
  end
  Update $\boldsymbol{\Sigma} \mid \Theta$ using Equation (4.12).
end

[a] See Kapelner and Bleich (2016) for further details on these tree proposal steps and transition probabilities $q(\cdot)$.

4.4.2 Probit suBART

In multivariate binary settings, the goal is to sample from the similar posterior

$$\pi(\Theta \mid \mathbf{Y}) = \pi\left(\left(\boldsymbol{\mathcal{T}}^{(1)}, \boldsymbol{\mathcal{M}}^{(1)}\right), \ldots, \left(\boldsymbol{\mathcal{T}}^{(d)}, \boldsymbol{\mathcal{M}}^{(d)}\right), \boldsymbol{\Sigma}, \mathbf{D} \mid \mathbf{Y}\right).$$

Note that $\mathbf{W}$ is a deterministic function of $\boldsymbol{\Sigma}$ and $\mathbf{D}$, and is hence omitted from the above distribution. However, it will be more convenient to work with the joint posterior of the parameters and the latent variables

$$\pi\left(\Theta, \mathbf{Z} \mid \mathbf{Y}\right) = \pi\left(\left(\boldsymbol{\mathcal{T}}^{(1)}, \boldsymbol{\mathcal{M}}^{(1)}\right), \ldots, \left(\boldsymbol{\mathcal{T}}^{(d)}, \boldsymbol{\mathcal{M}}^{(d)}\right), \boldsymbol{\Sigma}, \mathbf{D}, \mathbf{Z} \mid \mathbf{Y}\right),$$

for which the sampling algorithm is very similar to the previously presented Algorithm 4.1 in the continuous setting. For brevity, we discuss only the required modifications to Algorithm 4.1 without presenting a new algorithm in full.

The updates for the trees and leaf nodes stay essentially the same, with the one difference being that the latent variables $\mathbf{Z}$ replace the data $\mathbf{Y}$. The updates for the a_j parameters are of course dropped, since they do not apply to the probit model. An important additional step is that the latent variables for each component j should be updated after line 9 in Algorithm 4.1. In a similar manner to Equation (4.9), the marginal distribution of $z_i^{(j)}$ can be obtained as follows

$$\begin{aligned} z_i^{(j)} \mid \mathbf{z}_i^{(-j)}, \Theta \sim \mathrm{N}\Big(&\hat{z}_i^{(j)} + \boldsymbol{\Sigma}_{j(-j)} \boldsymbol{\Sigma}_{(-j)(-j)}^{-1} \left(\mathbf{z}_i^{(-j)} - \hat{\mathbf{z}}_i^{(-j)}\right), \\ &\Sigma_{jj} - \boldsymbol{\Sigma}_{j(-j)} \boldsymbol{\Sigma}_{(-j)(-j)}^{-1} \boldsymbol{\Sigma}_{(-j)j}\Big). \end{aligned} \tag{4.13}$$

However, sampling the latent variables also requires conditioning on $\mathbf{Y}$. In doing so, we find that $\pi(z_i^{(j)} \mid \mathbf{z}_i^{(-j)}, \mathbf{Y}, \Theta)$ follows a truncated normal distribution with the same location and scale parameters as Equation (4.13). However, we encounter a case distinction for the support of this conditional posterior distribution based on the values of the associated response. If $y_i^{(j)} = 0$, which implies that $z_i^{(j)} \leq 0$, the support is truncated to $(-\infty, 0]$. Conversely, if $y_i^{(j)} = 1$, which implies that $z_i^{(j)} > 0$, the support is truncated to $(0, \infty)$. In each case, we draw the sample through the method proposed by Robert (1995).

Finally, the other major difference for the probit suBART sampler is the update of $\boldsymbol{\Sigma}$ and the auxiliary parameter $\mathbf{D}$. We write out the conditional posterior as

$$\begin{aligned} \pi\left(\boldsymbol{\Sigma}, \mathbf{D} \mid \Theta, \mathbf{Z}\right) &\propto \pi\left(\boldsymbol{\Sigma}, \mathbf{D}\right) \left(\det \mathbf{D}\right)^{\frac{d-1}{2}} \pi\left(\mathbf{Z} \mid \Theta\right) \\ &\propto \pi\left(\boldsymbol{\Sigma}, \mathbf{D}\right) \left(\det \mathbf{D}\right)^{\frac{d-1}{2}} \left(\det \boldsymbol{\Sigma}\right)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{z}_i - \hat{\mathbf{z}}_i\right)^\top \boldsymbol{\Sigma}^{-1} \left(\mathbf{z}_i - \hat{\mathbf{z}}_i\right)\right), \end{aligned}$$

where $\pi(\boldsymbol{\Sigma}, \mathbf{D}) = \pi(\mathbf{W})$ is the aforementioned inverse-Wishart prior on $\mathbf{W}$. The term $\left(\det \mathbf{D}\right)^{\frac{d-1}{2}}$ is the Jacobian determinant which arises due to the change of

variables $\mathbf{W} \mapsto \boldsymbol{\Sigma}, \mathbf{D}$. This distribution is not of known form, and can hence cannot be sampled from directly. We instead proceed using the parameter-expanded Metropolis-Hastings (PX-MH) algorithm of Zhang (2020). This defines a proposal for $\mathbf{W}^{(k+1)} | \mathbf{W}^{(k)}, \nu_{\text{prop}} \sim \text{Inv-Wishart}_d \left(\nu_{\text{prop}}, \mathbf{W}^{(k)} \right)$ where k is the current MCMC iteration and ν_{prop} is a tuning parameter.

4.5 Simulation studies

In this section, we evaluate the efficacy of the proposed models through experiments with simulated data. Section 4.5.1 and Section 4.5.2 are devoted to simulation designs in which the responses $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(d)}$ are all continuous and each outcome $\mathbf{y}^{(j)} \in \{0, 1\}$ is binary, respectively. We undertake a comparative analysis of suBART against benchmark models, including the standard BART model applied independently to each response, the multivariate BART (mvBART) model, and a Bayesian linear seemingly unrelated regression (SUR) model. We also consider probit versions of each model, where available. This comparative study aims to explore various aspects of the models, including predictive performance and their ability to accommodate assumptions regarding correlation among responses and/or assumptions of linearity. Furthermore, our simulations aim to elucidate the primary distinctions between suBART and mvBART. For example, the splits generated by trees under the mvBART framework entail a splitting rule in all components of $\mathbf{Y}$, potentially deviating from an accurate representation of the true function $f(\mathbf{X})$ for some scenarios. Consequently, each response variable in our experiments is generated using a different subset of covariates. Additionally, a significant improvement in predictive performance capacity is anticipated when compared with the linear SUR model as, for the most part, the responses in our experiments are almost all assumed to be non-linearly related to the covariates.

These assumptions were tested over 100 replications of each simulation scenario, using different sample sizes of $n_{\text{train}} = n_{\text{test}} = \{250, 500, 1000\}$ for training and test samples respectively. The metrics employed to assess differences in model performance included the root mean squared error (RMSE), the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007), and the prediction interval (PI) coverage for the multivariate regression cases, while the logarithmic loss, the

accuracy (ACC), and the credible interval coverage of the probabilities from $\Phi(z_i^{(j)})$ were used for the multivariate probit scenarios. The 50% posterior intervals are computed using the 25-th and 75-th percentiles over the posterior samples from $\hat{Y}_i^{(j)}$, while the posterior means for each $\hat{Y}_i^{(j)}$, σ_j, and ρ_{jk} are obtained by averaging the posterior replications.

Throughout all experiments, the default choice for the inverse-Wishart hyperparameter is $\nu = 2$, reflecting our lack of prior information about the correlation structure (Huang and Wand, 2013). The selection of the proposal degrees of freedom ν_{prop} for the PX-MH algorithm can be fine-tuned to adjust its acceptance rate, as outlined in prior studies (Zhang et al., 2006). Consistent with existing literature (Zhang et al., 2015), we adopt the default value of $\nu_{\text{prop}} = n_{\text{train}}$, which appears to ensure a sufficiently well-behaved sampler. The number of trees for each component j was fixed at $m = 50$. For the MCMC settings, we set a total of $N_{\text{MCMC}} = 3000$ iterations, of which $N_{\text{burn-in}} = 1000$ samples are discarded as burn-in. Adjustments to the number of MCMC samples and other hyperparameters such as ν, ν_{prop}, and m could be made to enhance convergence and predictive performance, though the model does not seem to be overly sensitive to such choices.

Lastly, we note the software implementations for each model included in the comparison. The suBART models are fitted using our own `suBART` implementation and the BART models are fitted using the `dbarts` package Dorie et al. (2024), while the linear Bayesian SUR models (henceforth BayesSUR) are fitted using the probabilistic programming language `Stan` Stan Development Team (2024b), through the `rstan` package Stan Development Team (2024a) which provides an R interface for this library. The mvBART model was evaluated using the `skewBART` implementation provided by Um et al. (2023), specifically by setting the argument `do_skew=FALSE` of the main `MultiskewBART()` function. However, it is worth noting that its current implementation is limited to continuous scenarios with two dimensions, thereby results were constrained to such cases. The default arguments were retained for all competing models with the exception of the number of trees for the tree-based models, which were set to the same value ($m = 50$) as suBART.

4.5.1 Continuous response experiments

The two simulation scenarios described by the systems of equations below were created to accommodate different types of complexity. In these experiments, the values of the response are non-linear functions modified from examples described in Friedman (1991) and Breiman (1996) for a multivariate response scenario. In the first scenario, the third response is exceptional in the sense that the generating function is purely linear. Note that correlated noise $(\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(d)})^\top \sim \text{MVN}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$ is subsequently added to each scenario's d-dimensional response.

Friedman #1:

$$
\begin{aligned}
x_i^{(1)}, \ldots, x_i^{(10)} &\overset{\text{iid}}{\sim} \text{Uniform}(0,1) \\
y_i^{(1)} &= 10 \sin\left(x_i^{(1)} x_i^{(2)} \pi\right) + 20\left(x_i^{(3)} - 0.5\right)^2 \\
y_i^{(2)} &= 8 x_i^{(4)} + 20 \sin\left(x_i^{(1)} \pi\right) \\
y_i^{(3)} &= 10 x_i^{(5)} - 5 x_i^{(2)} - 5 x_i^{(4)}
\end{aligned}
$$

Friedman #2:

$$
\begin{aligned}
x_i^{(1)}, \ldots, x_i^{(5)}, x_i^{(8)}, x_i^{(10)} &\overset{\text{iid}}{\sim} \text{Uniform}(0,1) \\
x_i^{(6)} &\sim \text{Uniform}(0, 100) \\
x_i^{(7)} &\sim \text{Uniform}(40\pi, 560\pi) \\
x_i^{(9)} &\sim \text{Uniform}(1, 11) \\
y_i^{(1)} &= 10 \sin\left(x_i^{(1)} x_i^{(2)} \pi\right) + 20\left(x_i^{(3)} - 0.5\right)^2 + 10 x_i^{(4)} + 5 x_i^{(5)} \\
y_i^{(2)} &= \sqrt{{x_i^{(6)}}^2 + \left(x_i^{(7)} x_i^{(8)} - \frac{1}{x_i^{(7)} x_i^{(9)}}\right)^2} \\
y_i^{(3)} &= \text{atan}\left(\frac{x_i^{(7)} x_i^{(8)} - \frac{1}{x_i^{(7)} x_i^{(9)}}}{x_i^{(6)}}\right)
\end{aligned}
$$

The $p = 10$ predictors are generated from a uniform distribution in each scenario. It is notable that not all predictors are used to build the responses. However, all $p = 10$ predictors are used for model fitting in each case. It is anticipated that the tree-based models will be able to identify these uninformative noise variables. It is also essential to emphasise that each component of the outcome vector $\mathbf{Y}$ is

derived from a distinct set of predictors in both scenarios. However, no restrictions are imposed on which predictors are associated with which response during model fitting. As previously mentioned, it is anticipated that mvBART may encounter challenges in accurately approximating the true generating functions in such cases. For each tree, the partitioning of the covariate space is reflected across all responses, which may not hold true, particularly when examining the responses of the Friedman #2 scenario where $y^{(1)}$ and $y^{(2)}$ do not share any predictors.

In each simulated scenario, we varied the dimension of the covariance error matrix within $d = \{2, 3\}$ and defined $\mathbf{\Sigma}$ accordingly, with specific values assigned to each parameter σ_j and each correlation parameter ρ_{jk}, for all $j \neq k$. In both Friedman scenarios, the error covariance parameters were set as detailed in Table 4.5.1 and it is the first two responses $y_i^{(1)}$ and $y_i^{(2)}$ which comprise the $d = 2$ settings. The restriction to $d = 2$ enables consideration of the mvBART model in the comparisons, owing to the aforementioned limitation of the `skewBART` software to bivariate outcome settings.

Table 4.5.1: True parameters of $\mathbf{\Sigma}$ used for each simulation scenario.

	d	σ_1	σ_2	σ_3	ρ_{12}	ρ_{13}	ρ_{23}
Friedman #1	2	1.00	10.00	—	0.75	—	—
	3	1.00	2.50	5.00	0.80	0.50	0.25
Friedman #2	2	1.00	125.00	—	0.75	—	—
	3	1.00	125.00	0.10	0.80	0.50	0.25

A comparison of results is depicted in the boxplots in Figure 4.5.1 and Figure 4.5.2 which confirm previous assumptions about suBART performance. These figures illustrate the results for *Friedman #1* with $n_{\text{train}} = n_{\text{test}} = 1000$. In general, suBART exhibits either slightly superior or competitive predictive performance when compared to BART and mvBART, as evidenced by small average values of RMSE and CRPS over the test samples. Furthermore, when compared with BayesSUR, all tree-based methods exhibit a clear superiority in estimating the non-linear responses. The primary discrepancy occurs when $j = 3$, where BayesSUR has the best performance owing to the linearity of this response.

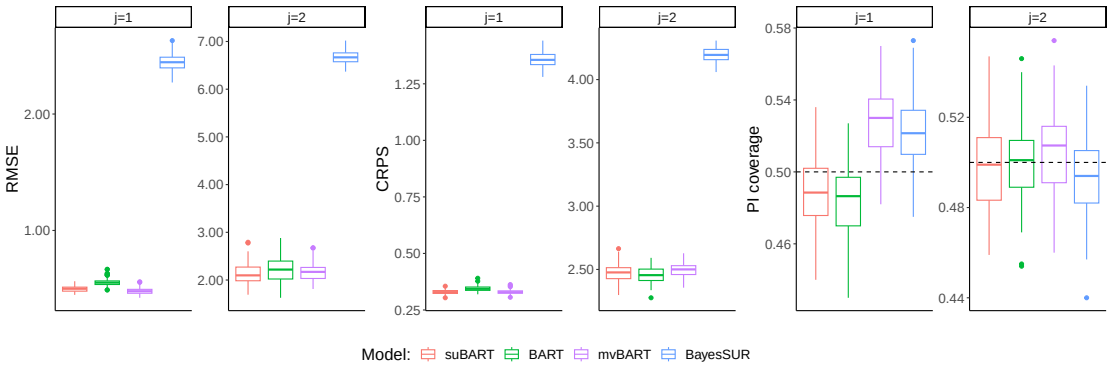


Figure 4.5.1: Simulation results for continuous outcomes for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 2$.

In terms of uncertainty estimation, Figure 4.5.1 illustrates that all methods exhibit reasonable coverage ratios when $d = 2$, except for the first component where both mvBART and linear SUR displayed higher coverage ratios for the prediction intervals, indicating that σ_1^2 was overestimated. Figure 4.5.2 corroborates these findings, with the divergence observed only when the response is solely dictated by a linear function, as illustrated by panels with $j = 3$ in Figure 4.5.2.

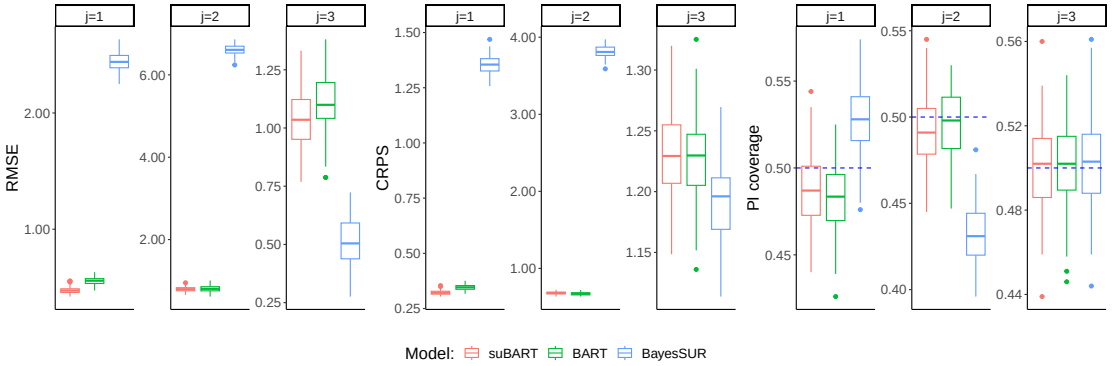


Figure 4.5.2: Simulation results for continuous outcomes for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 3$.

For the Friedman #2 scenario, the results are summarised in Figure 4.5.3 where $d = 2$ and $n_{\text{train}} = n_{\text{test}} = 1000$. In this case, suBART consistently outperforms its competitors in all aspects. The deteriorated performance of mvBART can be explained by the particular nature of the simulation setting, where each outcome relates to an entirely distinct set of predictors, while the tree splits assume the opposite. Additionally, the calibration of the suBART estimations remains con-

sistent, as evidenced by the boxplot for the PI coverage, which mostly covers the correct value. Equivalent figures summarising the results for the remaining scenarios, with $d = 3$ and/or different sample sizes, yield the same conclusions as above and have been omitted for brevity; they can be found in Appendix 4.A.

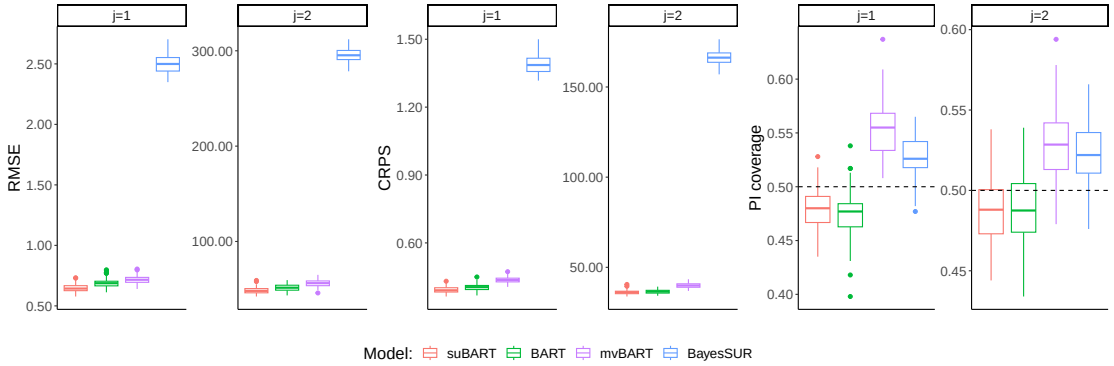


Figure 4.5.3: Simulation results for continuous outcomes for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 2$.

Ultimately, for proper uncertainty quantification, it is essential to correctly estimate the correlation values from the covariance matrix $\boldsymbol{\Sigma}$. Table 4.5.1 displays the RMSE and coverage ratio of a 50% credible interval (CI) for the correlation parameters ρ_{jk} for all $j \neq k$ provided by suBART, mvBART, and BayesSUR. Notably, correlation values for the BART model are not provided as it assumes independence among multiple responses (i.e., $\hat{\rho}_{jk} = 0 \, \forall \, j \neq k$), and the mvBART estimations are restricted when $d = 2$ due to limitations of the `skewBART` package. From the results, it is clear that suBART outperforms mvBART and BayesSUR in terms of coverage, demonstrating its superior ability to estimate correlation structures. The coverage values of zero for BayesSUR are particularly notable and suggest an inability to accurately estimate correlations when assuming linear regressions for non-linear responses. In terms of RMSE, suBART is superior to BayesSUR and comparable to mvBART, albeit only in the $d = 2$ setting.

Table 4.5.2: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 1000$ for continuous outcomes.

	RMSE			**CI coverage**		
	suBART	**mvBART**	**BayesSUR**	**suBART**	**mvBART**	**BayesSUR**
$d = 2$						
ρ_{12}	0.02	0.02	0.33	0.50	0.05	0.00
$d = 3$						
ρ_{12}	0.02	—	0.37	0.37	—	0.00
ρ_{13}	0.03	—	0.31	0.41	—	0.00
ρ_{23}	0.03	—	0.18	0.52	—	0.00

4.5.2 Binary response experiments

The experiments for binary responses are aligned with those from Section 4.5.1, wherein different training and test sample sizes of $n_{\text{train}} = n_{\text{test}} = \{250, 500, 1000\}$ are used. The simulation of the latent variables $z^{(j)}$ is described by the system of equations below. Other than when $j = 3$, the values of the latent variables are non-linear functions. Recall that correlated noise $(\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(d)})^\top \sim \text{MVN}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$ is subsequently added to the d-dimensional latent variable, where the true $\boldsymbol{\Sigma}$ parameters are set to the same values as Table 4.5.1, and that the generating process for each binary response follows Equation (4.2) thereafter.

Friedman #3:

$$
\begin{aligned}
x_i^{(1)}, \ldots, x_i^{(10)} &\overset{\text{iid}}{\sim} \text{Uniform}(0, 1) \\
z_i^{(1)} &= \sin\left(x_i^{(1)} x_i^{(2)} \pi\right) + {x_i^{(3)}}^3 \\
z_i^{(2)} &= -1 + 2x_i^{(1)} x_i^{(4)} + e^{x_i^{(5)}} \\
z_i^{(3)} &= 0.5\left(x_i^{(2)} + x_i^{(4)}\right) + x_i^{(5)}
\end{aligned}
$$

The models evaluated for these experiments are the probit suBART and probit extensions of the standard BART and BayesSUR. Despite the complete unavailability of a probit version of mvBART in the `skewBART` software, for any dimensionality, we persist in evaluating settings with varying dimension $d = \{2, 3\}$ with the $d = 2$ setting again comprising the first two responses. The results are summarised in Figure 4.5.4 and Figure 4.5.5, which are consistent with the findings from Section 4.5.1. When logarithm loss and ACC are considered as metrics for evaluating

predictive performance, suBART either exhibits superior results or comparable averages. Both tree-based models outperform Bayesian SUR, with the exception of the linear third response in the $d = 3$ setting, as per the continuous simulation studies in Section 4.5.1.

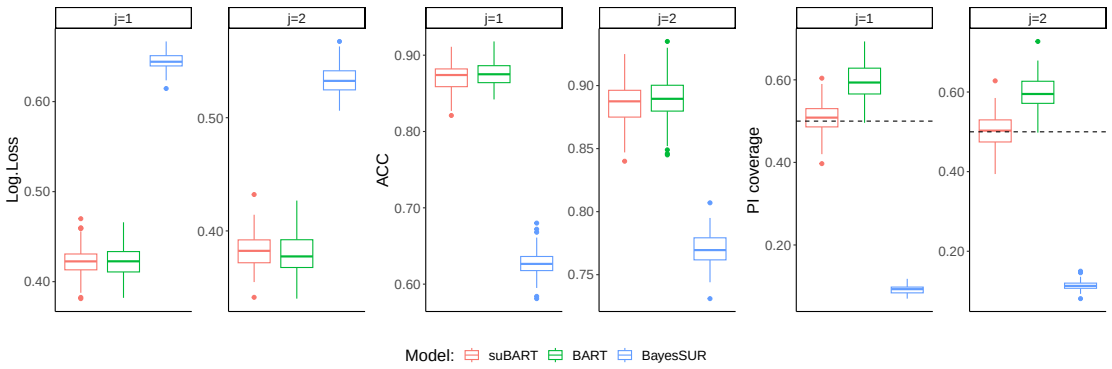


Figure 4.5.4: Simulation results for binary outcomes for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 2$.

Regarding calibration, it is evident that suBART outperforms BART across all scenarios, showing coverage ratios that closely approximate the true values. On the other hand, due to the inherent linearity of BayesSUR, its calibration performance is notably poorer, though the third linear response is again an exception in this regard, as per Section 4.5.1. However, even for this linear response, suBART appears to exhibit superior performance in uncertainty quantification when compared to standard BART. For brevity, the results for remaining sample sizes are presented in Appendix 4.A, as they lead to similar conclusions.

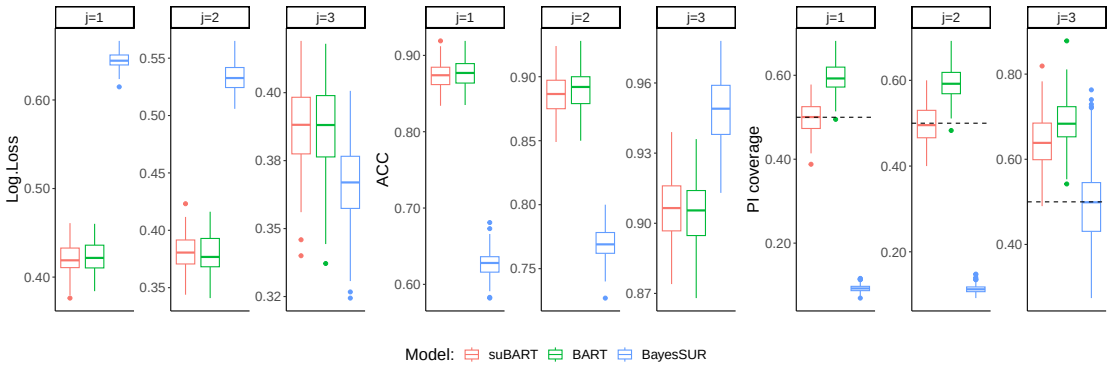


Figure 4.5.5: Simulation results for binary outcomes for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 3$.

The results regarding the estimation of the correlation parameters are presented in Table 4.5.3, which includes the RMSE and CI coverage for the correlation parameters ρ_{jk} associated with binary responses when $n_{\text{train}} = 1000$. These results are consistent with Table 4.5.2 in clearly demonstrating superior performance of suBART with respect to prediction accuracy and estimating correlation structures.

Table 4.5.3: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 1000$ for binary outcomes.

	RMSE		**CI coverage**	
	suBART	**BayesSUR**	**suBART**	**BayesSUR**
$d = 2$				
ρ_{12}	0.25	0.51	0.42	0.00
$d = 3$				
ρ_{12}	0.04	0.27	0.46	0.00
ρ_{13}	0.05	0.11	0.47	0.06
ρ_{23}	0.06	0.08	0.47	0.32

4.6 Analysis of the TTCM data

We now apply the continuous suBART model in the cost-effectiveness setting which we introduced in Section 4.2. We analyse data from Wiertsema et al. (2019). The authors collected data on $n = 140$ patients suffering from traumatic injuries. The two treatment options are usual care and the novel transmural trauma care model (TTCM), denoted by $t = 0$ and $t = 1$, respectively. The treatment assignment was not randomised. The outcomes we use, for c and q respectively, are the costs from the healthcare perspective and generic healthcare-related quality of life.

As is usual in CEAs, the cost outcome is an aggregate measure: c comprises the total costs acquired from hospital records as well as several questionnaires conducted over the course of nine months following treatment, in which patients were surveyed on their use of various healthcare resources. The responses — examples of which relate to issues such as hospital stays, medication use, and surgeries — were then converted to costs. Conversely, the effectiveness outcome q was calculated from one single survey administered nine months after treatment using the EQ-5D-3L instrument (Lamers et al., 2006). Additional details on the data collection process can be found in Wiertsema et al. (2019). The data also

includes $p = 11$ baseline covariates, with respective sample sizes of 83 and 57 in the two treatment groups. The ratio of covariates to observations is thus reasonably large. We reproduce the table of baseline variables in Table 4.6.1. We account for the categorical predictors in the tree-based models using the method of ordering categories proposed by Breiman et al. (1984).

Table 4.6.1: Baseline data from Wiertsema et al. (2019).

Characteristics	**Mean (SD) *or* frequency (%)**	
	Intervention group ($t = 1$)	Control group ($t = 0$)
n	83	57
Age	43.4 (15.6)	50.5 (17.9)
Gender (M/F)	39/44 (47/53%)	26/31 (46/54%)
Education level		
Low	7 (8.4%)	6 (11.1%)
Middle	19 (22.9%)	16 (29.6%)
High	57 (68.7%)	32 (59.3%)
Medical history		
None	53 (63.9%)	30 (52.6%)
Chronic	14 (16.9%)	13 (22.8%)
Musculoskeletal	16 (19.3%)	14 (24.6%)
Trauma type		
Traffic	44 (53.0%)	25 (43.9%)
Work related	0 (0.0%)	2 (3.5%)
Fall	27 (32.5%)	17 (29.8%)
Sports	11 (13.3%)	9 (15.8%)
Other	1 (1.2%)	4 (7.0%)
Fracture region		
Upper extremity	31 (37.3%)	25 (43.9%)
Lower extremity	41 (49.4%)	19 (33.0%)
Vertebral	7 (8.4%)	1 (1.8%)
Multitrauma	4 (4.8%)	12 (21.1%)
Injury severity score	7.9 (4.4)	8.6 (6.3)
Hospital admission	62 (75%)	29 (51%)
Length of hospital stay (days)	7.1 (6.1)	10.0 (11.4)
Surgery	53 (64%)	21 (37%)
TTO[b]	24.3 (14.3)	14.6 (14.7)

[b] Days between trauma and first outpatient consultation.

The original dataset had some missing observations — for survey items related to the outcome variables only — which Wiertsema et al. (2019) dealt with through multiple imputation. 17% of patients did not complete any follow-up questionnaires, and hence were missing all information on q and some survey information on c (though hospital records were available for all patients). Additionally, 39% and 7% of respondents were missing some (but not all) survey items related to c and q respectively. As missing data is not the subject of this chapter, we will avoid this complication by simply working with *one* imputed dataset, obtained through predictive mean matching (Vink et al., 2014), and treating that as complete data. Specifically, the imputation is applied to the missing survey items prior to the calculation of c and q. It follows that the analysis given here is not directly comparable to the original one, and we do not claim that it is more valid in this regard. We discuss this issue further in Section 4.7.

We will compare three methods for estimating the treatment effects: (1) suBART, (2) mvBART as implemented in `skewBART`, and (3) Bayesian linear SUR, with default priors as provided in the Stan user's guide (Stan Development Team, 2024b). We use $m = 100$ trees for the tree-based methods. It is worth noting that we do not impose any restrictions on the sets of covariates associated with each response, for any of these methods. All covariates in Table 4.6.1 are used. This means that all trees are *allowed to* form splitting rules using all covariates for suBART, the single set of multivariate trees are allowed to split on all covariates for mvBART, and all linear regressions for BayesSUR also share all covariates. Following Wiertsema et al. (2019), we do not specify any interaction effects or non-linear terms in the linear predictors for BayesSUR, owing to the difficulty of pre-specifying appropriate functional forms in the presence of a large amount of candidate interactions and the associated challenges in terms of model selection. In any case, Dorie et al. (2019) found that linear models perform poorly in causal settings even when also including interactions, polynomial terms, and regularisation to avoid overfitting. Conversely, BART-based methods are well-equipped to automatically capture low-order interactions and non-linearities (Linero and Yang, 2018; Ročková and Van der Pas, 2020).

In addition, we evaluate each method again with the set of predictors augmented using propensity scores estimated via probit BART. This procedure is inspired by the univariate ps-BART method proposed by Hahn et al. (2020), which induces a covariate-dependent prior on the regression function and can substantially reduce bias due to regularisation-induced confounding. We first estimate propensity scores for each patient through probit BART, which give the probability that a patient received treatment 1, conditional on their baseline characteristics, and then add the posterior mean propensity score estimates to the set of predictors $\mathbf{x}_i$ used to estimate the conditional expectations $\mathbb{E}[c \mid t, \mathbf{x}_i]$ and $\mathbb{E}[q \mid t, \mathbf{x}_i]$ via the chosen model. Thus, we expand the comparison to include what we refer to as ps-suBART, ps-mvBART, and ps-BayesSUR, which are straightforward adaptations of the univariate ps-BART method. Although the use of BART-based propensity scores in conjunction with linear SUR deviates from usual practice in the applied health economics literature, we nonetheless use the same set of propensity scores estimated via probit BART for each method, in order to ensure the comparison is fair in this regard. Expanding the comparison to include versions of each method with and without propensity scores will help to establish the extent to which differences in results are attributable to differences in model specification or due to the inclusion of propensity scores.

4.6.1 Results of the TTCM data analysis

Figure 4.6.1 shows the distribution of the estimated propensity scores. Despite some overlap, it is evident that the treatment groups are quite imbalanced with respect to their baseline characteristics and that some form of covariate adjustment is necessary to avoid biased results. As already described, we reevaluate each model with the estimated propensity scores included as an additional covariate. Following Li et al. (2023), who characterise this as an "approximately Bayesian" procedure, we expect this to lead to results which are more robust to model misspecification. Furthermore, this step makes the models less prone to attributing the effect of confounders to the treatment variable Hahn et al. (2020).

In Figure 4.6.2, we show highest density regions of kernel density estimates of the posterior distributions of Δ_c and Δ_q — obtained through suBART, mvBART,

and BayesSUR, as well as their counterpart models which also incorporate the estimated propensity scores as additional predictors — in the form of a cost-effectiveness plane (CEP; see Gabrio et al. (2019) for more details). In general, for both Δ_c and Δ_q, the centers of the distributions are shifted further away from 0 and posterior uncertainty is greater when the propensity scores are incorporated. These differences are most pronounced between ps-BayesSUR and BayesSUR and the posterior mean is furthest from the origin under ps-suBART. A notable distinction between suBART and BayesSUR is that the former appears to be more uncertain about Δ_c while the latter appears to be more uncertain about Δ_q. Furthermore, the same applies to the comparison between ps-suBART and ps-BayesSUR.

In Table 4.6.2, we further provide some summary statistics for Δ_c, Δ_q, and the INB at a representative value of $\lambda = 20000$. We stress that this quantity is not indicative of the cost of the TTCM intervention, which averages €272 per patient (Wiertsema et al., 2019). Rather, this value is indicative of a situation in which decision-makers would be willing to pay €20,000 *per additional unit of healthcare-related quality of life*, which is much greater as the treatment effect of Δ_q is near-zero. In any case, €20,000 is a commonly used threshold for determining whether an intervention represents value for money in the CEA literature (see e.g., Drummond et al. (2015) and Gabrio et al. (2019)). At this chosen λ value, we see some considerable differences between ps-suBART and the other methods. This suggests that there may be strong non-linear functional relationships between the covariates and the outcomes (such that suBART benefits from relaxing the linearity assumption of BayesSUR, without requiring pre-specification of the functional forms) and that those relationships may differ for the two outcomes c and q (such that suBART benefits from relaxing the mvBART assumption of a common tree structure). Notably, the estimated treatment effects are markedly smaller in absolute value for each method when propensity scores are excluded and only ps-suBART yields a 95% CI for INB_{20000} which excludes zero.

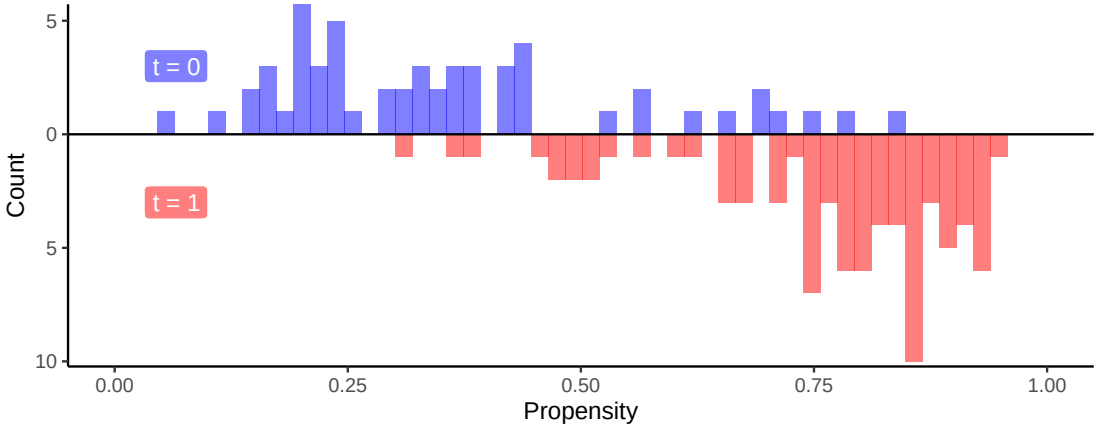


Figure 4.6.1: Propensity scores by treatment arm, estimated via probit BART, where $t = 0$ corresponds to the control group.

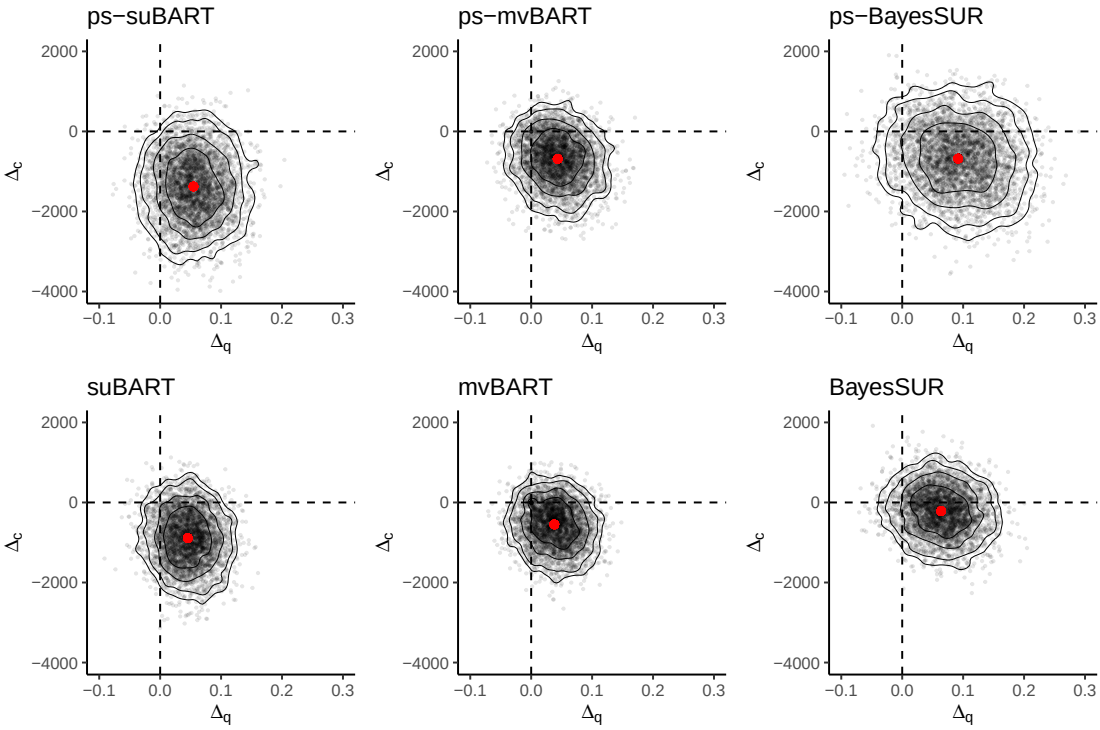


Figure 4.6.2: CEPs showing highest density regions of kernel density estimates of the posterior distributions of Δ_c and Δ_q according to each model, with and without the propensity scores. The posterior means are indicated by a red dot in each case, the individual draws of Δ_c and Δ_q are shown via grey points , and the contour lines correspond to probability levels of 0.5, 0.75, 0.9, and 0.95.

Table 4.6.2: Posterior means with 95% credible intervals for Δ_c, Δ_q, and INB_λ at a representative value of $\lambda = 20000$ for each model, with and without propensity scores.

Model	**Mean and 95% CI**					
	Δ_c		Δ_q		INB_{20000}	
ps-suBART	-1371	$[-2895, 189]$	0.054	$[-0.020, 0.127]$	2449	$[210, 4635]$
suBART	-880	$[-2126, 345]$	0.045	$[-0.019, 0.110]$	1776	$[-65, 3585]$
ps-mvBART	-674	$[-1859, 362]$	0.042	$[-0.025, 0.112]$	1516	$[-282, 3415]$
mvBART	-587	$[-1678, 477]$	0.038	$[-0.024, 0.104]$	1353	$[-373, 3137]$
ps-BayesSUR	-698	$[-2166, 771]$	0.093	$[-0.012, 0.203]$	2560	$[-228, 5395]$
BayesSUR	-191	$[-1285, 916]$	0.063	$[-0.016, 0.144]$	1448	$[-660, 3556]$

Rather than relying on a single λ value, we also show the probability of cost-effectiveness as a function of the willingness-to-pay λ in Figure 4.6.3. This plot, called a cost-effectiveness acceptability curve (CEAC; Löthgren and Zethraeus, 2000), is a highly-important tool in guiding the decision of which medical intervention to implement. These probabilities are simply estimated by counting all posterior draws for which $\text{INB}_\lambda > 0$ and dividing this count by the total number of posterior draws. We see some remarkable differences, depending on which type of model is used and whether or not the estimated propensity scores are incorporated. Notably, the estimated probability of cost-effectiveness only exceeds the typical 95% reference level at any λ under ps-suBART, suBART, and ps-BayesSUR. Moreover, this probability is consistently larger at all λ values for all methods when propensity scores are included. Given that we have reason to believe that ps-suBART, ps-mvBART, and ps-BayesSUR are more accurate than their counterparts, we henceforth discuss only these methods.

A particularly striking aspect of Figure 4.6.3 is that ps-suBART is the only method for which the estimated probability of cost-effectiveness is well above 95% for all λ values. This threshold is typically regarded as a reasonably high probability of cost-effectiveness, so we would be quite content in asserting that the TTCM is cost-effective, regardless of the value of λ. In fact, even if decision-makers are unwilling to pay anything (i.e., $\lambda = 0$) per unit of effect gained, ps-suBART still reports a probability of 0.97 of being cost-effective compared to regular care. For ps-BayesSUR, the probabilities are notably lower for small λ values, even when

including propensity scores. Based on the ps-BayesSUR results, we likely would want to collect more data before committing to a final decision. At higher λ, the probabilities of cost-effectiveness approach the estimates obtained by ps-suBART. For ps-mvBART, the probabilities follow a similar pattern as the ps-suBART estimates, but remain lower across the full range of λ values and never reach the 0.95 reference level. We are thus lead to a substantially different conclusion, depending on the method used: ps-suBART finds strong evidence for TTCM being cost-effective, while the results from the other two models are less conclusive. As previously alluded to, the results for each method are even less conclusive when the propensity scores are omitted from the set of predictors.

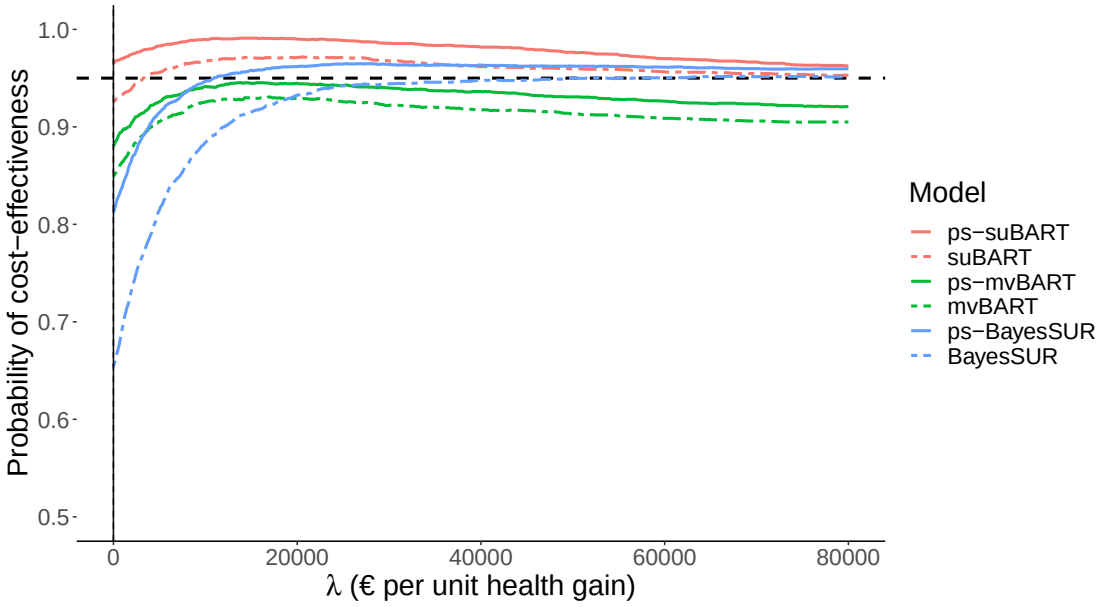


Figure 4.6.3: CEACs showing probabilities of cost-effectiveness as a function of λ for each model, with and without propensity scores, with a horizontal dotted line at the 0.95 reference level.

4.7 Discussion

In this chapter, we introduce the suBART model for multivariate outcomes both as means of accounting for non-linearities and interactions in the seemingly unrelated regression framework and as a means of addressing the key limitation of existing multivariate BART approaches which assume a single set of trees, such that the

entire response vector is partitioned in the same way by the splitting rules in the ensemble. By modelling each component of the outcome using a univariate BART, suBART captures non-linearities in the relationship between the response and predictors while allowing for and detecting the different subsets of covariates — along with the interactions between them — associated with each response without enforcing common tree structures. We further develop the model to handle multivariate binary outcomes.

The effectiveness of suBART is demonstrated through extensive simulation studies, in which it is shown that the model adequately captures non-linear responses, accurately estimates the covariance structure for multivariate responses, and generally outperforms its main competitors — including other tree-based alternatives and the Bayesian linear SUR — from the points of view of both predictive accuracy and uncertainty quantification. Notably, suBART is consistently superior to the strategy of applying the standard BART model independently to each response, which demonstrates the benefits of modelling the covariance between multivariate response under our suBART framework. Furthermore, the results show that the model exhibits enhanced flexibility compared to its direct multivariate counterpart, mvBART. This flexibility stems from suBART permitting variation in splitting rules across each response, by allowing the trees for each outcome component to differ rather than imposing common tree structures. This enables a more accurate representation, especially when different outcome components depend on distinct sets of predictors.

The main focus of this chapter is the application of suBART within the context of cost-effectiveness analysis, a setting in healthcare where it is of interest to jointly estimate the healthcare costs and the health-related quality of life associated with two or more treatment options. In our analysis, we find remarkable differences depending on the particular method used for the TTCM data. It is of course expected to see large differences between linear SUR and the two BART-based models, given the very different model assumptions. The even larger differences between suBART and mvBART are arguably more interesting. To reiterate, suBART assigns each outcome its own tree ensemble, while the trees for all outcomes are the same for mvBART. The large differences in results suggest to us that the

assumption of a shared tree structure is a significant one, which may have a strong impact on the results. Unlike mvBART, suBART can accommodate data where the dependence on $\mathbf{X}$ is very different for different components of the outcome vector. In the CEA context, this applies particularly in situations where factors which govern the costs are unrelated to the quality of life, and *vice versa*. Such situations do occur in practise; for example, when investigating the effect of total knee replacement, Dakin et al. (2012) found that the patients' sex was a strong predictor of healthcare costs yet had no measurable relationship with quality of life, while the exact opposite was true for the patients' age. On the other hand, a shared tree structure may lead to estimates which are more precise, without necessarily being more accurate, since there are more data available to inform the tree structure. Um et al. (2023) claim this as an advantage of their method. We do indeed see somewhat smaller posterior variance in the mvBART estimates. We also find that suBART further benefits from the inclusion of estimated propensity scores as an additional predictor. Overall, we consider the model we refer to as ps-suBART to be a natural adaptation of the univariate BART approach and expect it to be a very useful tool in the analysis of observational cost-effectiveness data.

Despite the extensive array of comparisons and scenarios and the additional insights gleaned by suBART in the CEA setting, there remains ample opportunity for further exploration of various extensions to suBART. We delineate some of these possibilities below in light of limitations identified in the simulation studies and real data application.

- Although suBART models non-linear responses with considerable flexibility, the model may still lack the desirable smoothness in certain scenarios; as it is based on the standard BART model, its construction relies on sums of piecewise-constant functions. However, several methods have been proposed to address the inherent lack of smoothness in BART (Linero and Yang, 2018; Prado et al., 2021; Maia et al., 2024) and these approaches could potentially be adapted to suBART as well.

- The suBART framework can be contrasted with the traditional SUR framework in that the conditional expectations are all modelled either via nonpara-

metric univariate BART models or via parametric linear models. In the simulation studies, suBART's superiority in capturing non-linear responses was comprehensively demonstrated, although BayesSUR was preferable when the response was generated by a simple linear function. A semi-parametric model in which some outcomes are modelled by BART and some are modelled by linear regressions could be of interest in cases where practitioners have strong prior belief about the complexity or lack thereof of one or more responses. Alternatively, such scenarios could be handled by varying the number of trees assigned to each component, though further experiments will be required to verify this.

- In the classic SUR model context, accounting for heteroscedasticity has been a common challenge in the literature Afolayan and Adeleke (2018). As suBART represents an effective alternative to the traditional linear SUR, the extension proposed by Pratola et al. (2020) to accommodate heteroscedasticity within BART could be adapted to the suBART setting to address cases where the homoscedasticity assumption is invalid. However, it is important to note that the approach of Pratola et al. (2020) has yet to be extended beyond scalar variance estimation.

- The proposed suBART accommodates two types of multivariate outcomes: all continuous and all binary. We stress however that this flexibility is exclusive and does not apply to both types simultaneously. Previous works in the literature, such as those by Papageorgiou et al. (2014), Zhang et al. (2015), and Pourmohamad and Lee (2016), have adapted a Bayesian framework for handling responses of mixed type. These approaches could potentially be adapted to the suBART framework, providing greater generalisation of the approach, particularly given that trading off treatment costs against binary outcomes (e.g., cancer remission) is often of interest in other healthcare applications.

- The ps-suBART model we used in our analysis of the TTCM data builds on the ps-BART model (Hahn et al., 2020). Hahn et al. (2020) also propose a another related method, the Bayesian causal forest (BCF). The authors and

their discussants found the performance of ps-BART and BCF to be similar overall, with each method improving on the other in some settings. One of the advantages of BCF is the ability to directly specify a prior on the amount of treatment effect heterogeneity, while this prior is left implicit in ps-BART. The drawback of this flexibility is that the prior specification is more complex in BCF, and it is harder to find reasonable default choices which are appropriate in a variety of settings. Additionally, the computational demands of BCF are greater than for standard BART. Both of these issues would become even more challenging in the SUR setting with multivariate outcomes. However, given that a multivariate generalisation of BCF has recently been proposed by McJames et al. (2023), which is analogous to the multivariate generalisation of BART developed by Um et al. (2023), there remains scope for embedding BCF in the seemingly unrelated framework as an alternative multivariate approach for conducting cost-effectiveness analyses.

- CEAs are sometimes conducted with more than two treatment options of interest. Conceptually this is similar to the two-treatments setting presented in the TTCM application. For details, we refer interested readers to Drummond et al. (2015), Section 4.4. The ps-suBART method could be easily extended to settings with more than two treatments, by estimating propensity scores using multinomial probit BART (Kindo et al., 2016). With more than one treatment effect for each outcome, and more than one INB, each corresponding to a pairwise comparison of two treatments, the treatment which has a positive INB in all comparisons would be considered cost-effective.

- When presenting the TTCM data, we noted that the original dataset had missing responses for survey questions related to the outcome variables. We bypassed this by working instead with an artificial complete dataset. Ideally, we would prefer to keep the missing values and incorporate the imputation directly into the posterior computation, in order to get coherent estimates of posterior uncertainty. Assuming that the observations are missing at random (the probability of missingness does not depend on the true values of the missing observations, conditional on the observed data), it is straightforward to incorporate imputation into our Gibbs sampler: we would simply draw

the missing $y_i^{(j)}$ values from their conditional distribution, given in Equation (4.9). Imputing outcomes with the probit suBART model could be done in a similar manner. However, imputation of missing outcomes for the present application would be further complicated by the fact that the outcomes were calculated after imputing the constituent survey responses by Wiertsema et al. (2019). Imputing missing *covariate* values, on the other hand, is a totally different matter: our models are formulated conditionally on $\mathbf{X}$, and hence impose none of the distributional assumptions on $\mathbf{X}$ that would be required for imputation tasks. That being said, missing covariate values could be handled by simply adapting the approach of Kapelner and Bleich (2016).

We hope to incorporate these advancements into our forthcoming research plans and anticipate suBART's adoption in other CEA settings to inspire further developments.

Appendix

4.A Performance evaluation on simulation experiments

In order to evaluate the performance of suBART against its competitors, we conducted experiments replicating those outlined in Section 4.5, where we varied $n_{\text{train}} = n_{\text{test}} = \{250, 500, 1000\}$ and $d = \{2, 3\}$. This appendix summarises the remaining results omitted from the main chapter. The findings illustrated in the boxplots below align with the conclusions drawn in Section 4.5. Overall, suBART exhibits reasonable performance metrics for both continuous and binary outcome scenarios, either matching or surpassing its tree-based model counterparts. Notably, suBART outperforms the Bayesian linear SUR across all scenarios, with the exception of the responses where the response with the predictors is exclusively linear. Additionally, we provide a summary of the estimation of correlation parameters through tables displaying the RMSE and 50% CI coverage for all ρ_{kj} where $j \neq k$. Across all scenarios, suBART consistently approaches the true values and provides credible intervals with proper coverage ratios.

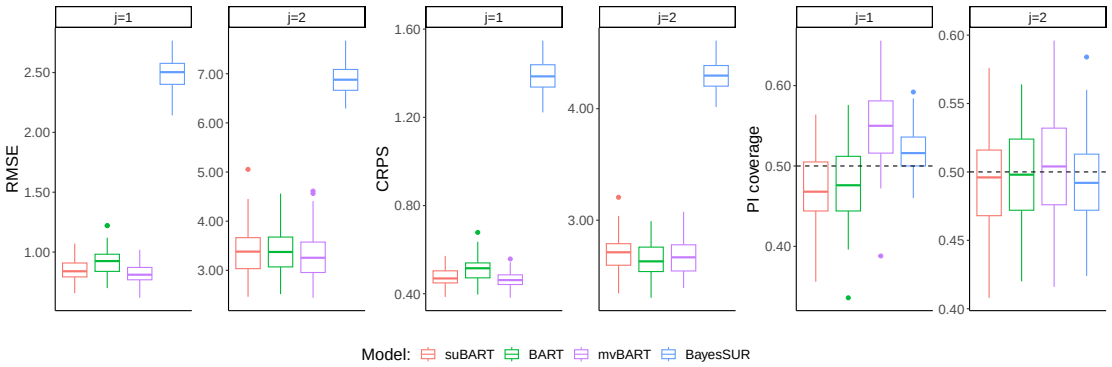


Figure 4.A.1: Simulation results for continuous outcomes for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 250$ and $d = 2$.

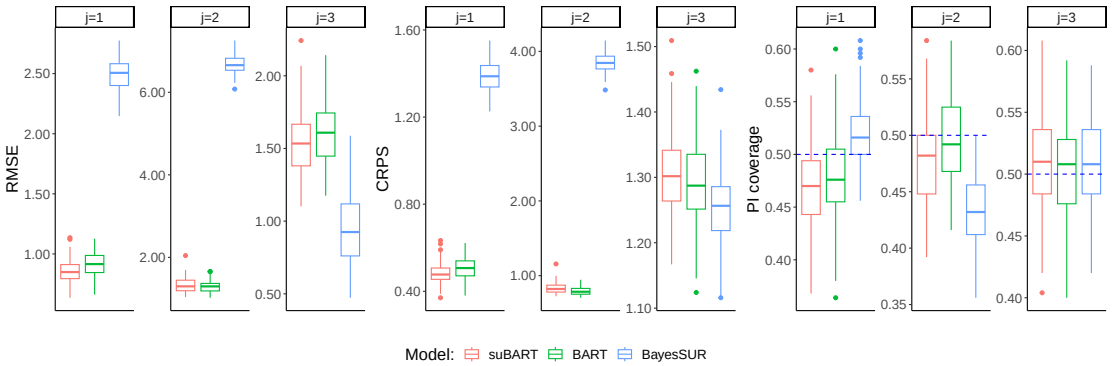


Figure 4.A.2: Simulation results for continuous outcomes for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 250$ and $d = 3$.

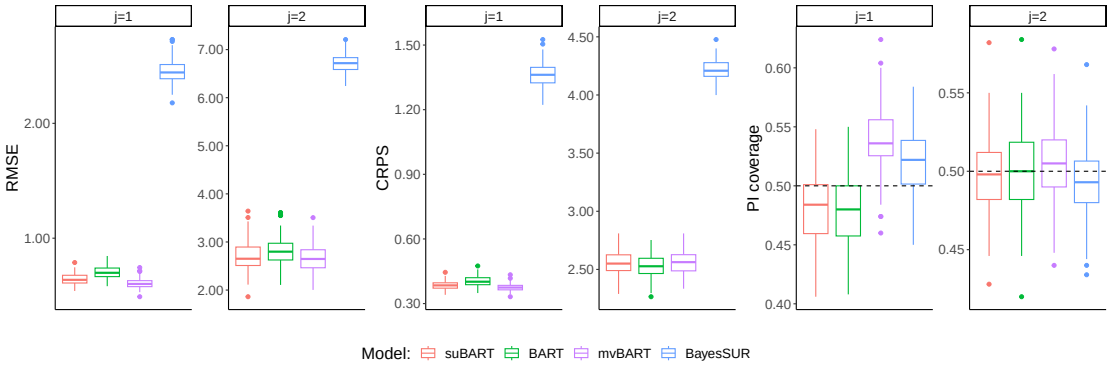


Figure 4.A.3: Simulation results for continuous outcomes for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 500$ and $d = 2$.

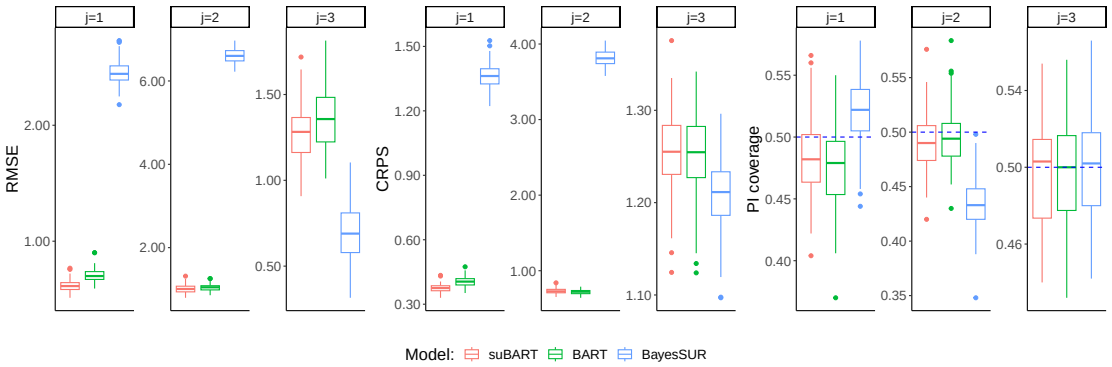


Figure 4.A.4: Simulation results for continuous outcomes for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 500$ and $d = 3$.

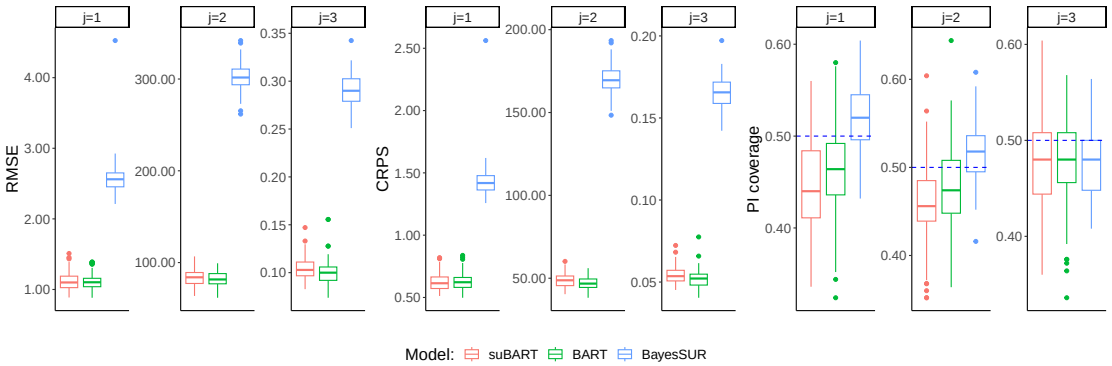


Figure 4.A.5: Simulation results for continuous outcomes for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 250$ and $d = 2$.

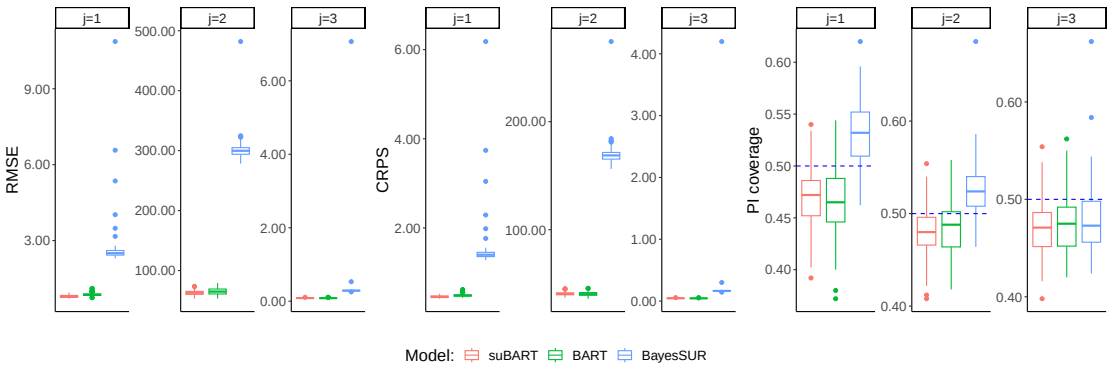


Figure 4.A.6: Simulation results for continuous outcomes for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 500$ and $d = 2$.

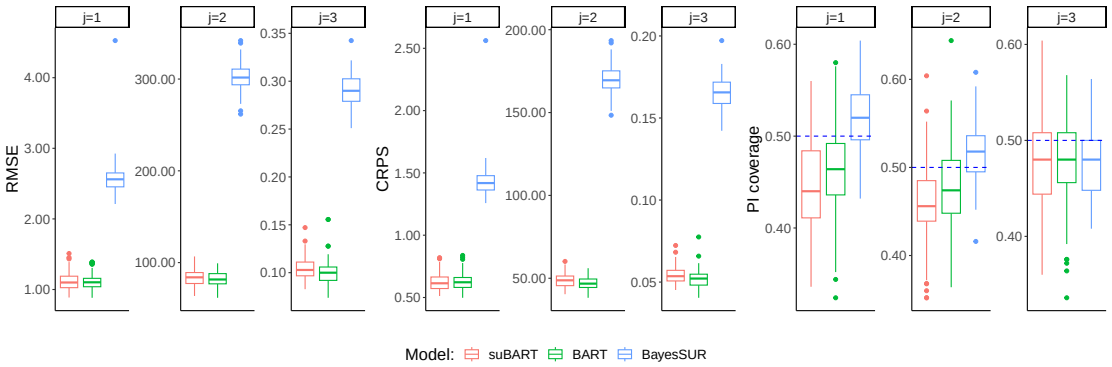


Figure 4.A.7: Simulation results for continuous outcomes for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 3$.

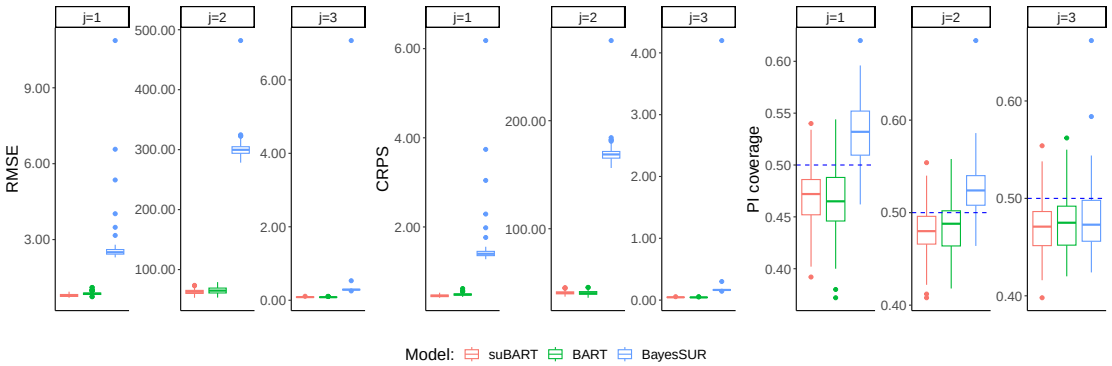


Figure 4.A.8: Simulation results for continuous outcomes for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 500$ and $d = 3$.

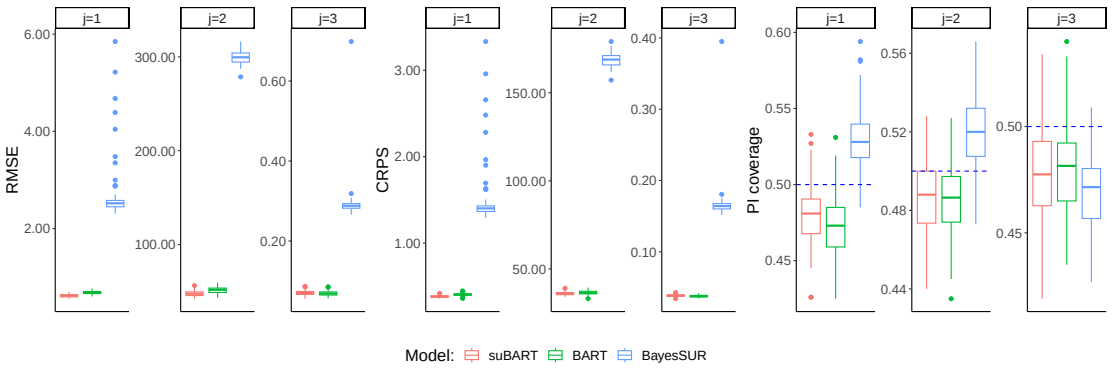


Figure 4.A.9: Simulation results for continuous outcomes for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 1000$ and $d = 3$.

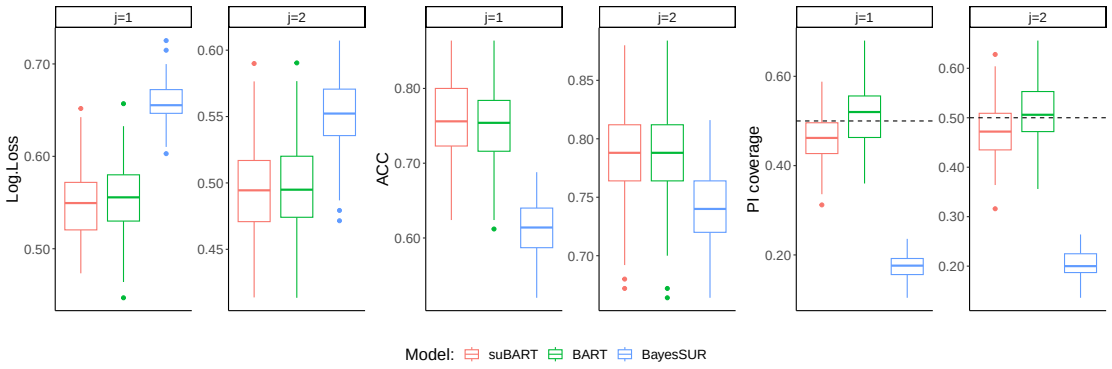


Figure 4.A.10: Simulation results for binary outcomes for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 250$ and $d = 2$.

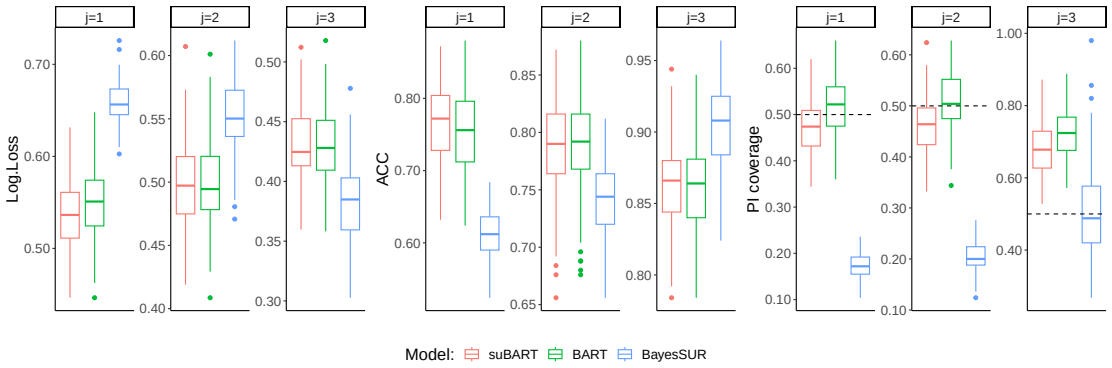


Figure 4.A.11: Simulation results for binary outcomes for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 250$ and $d = 3$.

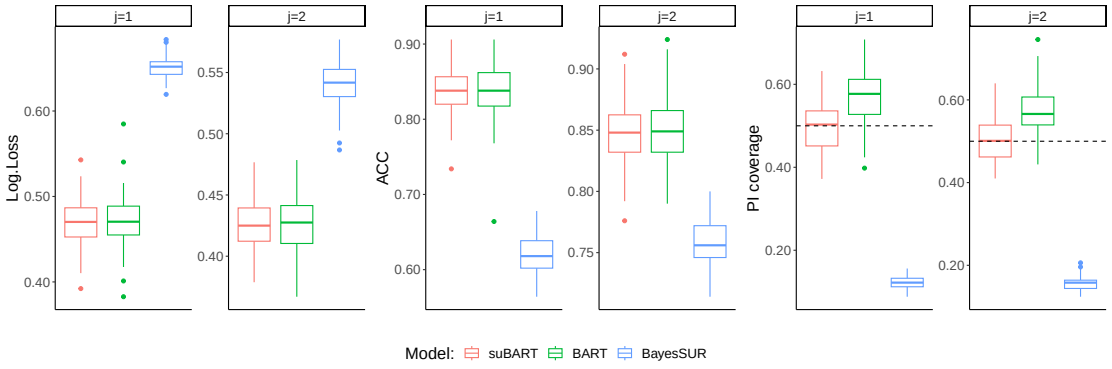


Figure 4.A.12: Simulation results for binary outcomes for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 500$ and $d = 2$.

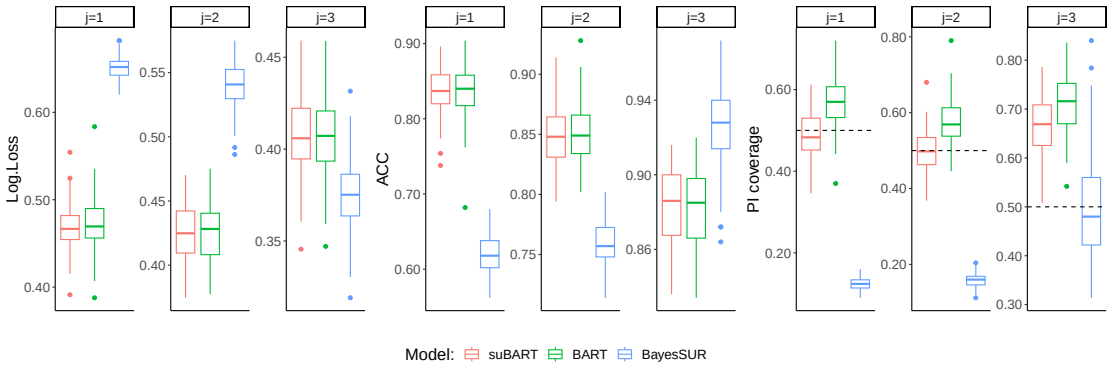


Figure 4.A.13: Simulation results for binary outcomes for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 500$ and $d = 3$.

Table 4.A.1: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 250$ for continuous outcomes.

	RMSE			**CI coverage**		
	suBART	**mvBART**	**BayesSUR**	**suBART**	**mvBART**	**BayesSUR**
$d = 2$						
ρ_{12}	0.05	0.09	0.34	0.33	0.17	0.00
$d = 3$						
ρ_{12}	0.04	—	0.39	0.30	—	0.00
ρ_{13}	0.07	—	0.32	0.39	—	0.00
ρ_{23}	0.08	—	0.18	0.39	—	0.02

Table 4.A.2: RMSE coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #1 with $n_{\text{train}} = n_{\text{test}} = 500$ for continuous outcomes.

	RMSE			**CI coverage**		
	suBART	**mvBART**	**BayesSUR**	**suBART**	**mvBART**	**BayesSUR**
$d = 2$						
ρ_{12}	0.03	0.07	0.34	0.47	0.07	0.00
$d = 3$						
ρ_{12}	0.02	—	0.38	0.48	—	0.00
ρ_{13}	0.04	—	0.32	0.39	—	0.00
ρ_{23}	0.05	—	0.17	0.45	—	0.00

Table 4.A.3: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 250$ for continuous outcomes.

	RMSE			**CI coverage**		
	suBART	**mvBART**	**BayesSUR**	**suBART**	**mvBART**	**BayesSUR**
$d = 2$						
ρ_{12}	0.05	0.258	0.34	0.33	0.00	0.00
$d = 3$						
ρ_{12}	0.04	—	0.39	0.30	—	0.00
ρ_{13}	0.07	—	0.32	0.39	—	0.00
ρ_{23}	0.08	—	0.18	0.39	—	0.02

Table 4.A.4: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 500$ for continuous outcomes.

	RMSE			**CI coverage**		
	suBART	**mvBART**	**BayesSUR**	**suBART**	**mvBART**	**BayesSUR**
$d = 2$						
ρ_{12}	0.04	0.21	0.50	0.32	0.00	0.00
$d = 3$						
ρ_{12}	0.05	—	0.56	0.24	—	0.00
ρ_{13}	0.07	—	0.49	0.33	—	0.01
ρ_{23}	0.06	—	0.71	0.55	—	0.00

Table 4.A.5: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #2 with $n_{\text{train}} = n_{\text{test}} = 1000$ for continuous outcomes.

	RMSE			**CI coverage**		
	suBART	**mvBART**	**BayesSUR**	**suBART**	**mvBART**	**BayesSUR**
$d = 2$						
ρ_{12}	0.06	0.17	0.49	0.02	0.00	0.00
$d = 3$						
ρ_{12}	0.05	—	0.58	0.02	—	0.00
ρ_{13}	0.05	—	0.49	0.18	—	0.01
ρ_{23}	0.03	—	0.72	0.51	—	0.00

Table 4.A.6: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 250$ for binary outcomes.

	RMSE		**CI coverage**	
	suBART	**BayesSUR**	**suBART**	**BayesSUR**
$d = 2$				
ρ_{12}	0.11	0.29	0.40	0.00
$d = 3$				
ρ_{12}	0.11	0.31	0.41	0.00
ρ_{13}	0.11	0.11	0.53	0.21
ρ_{23}	0.13	0.12	0.44	0.37

Table 4.A.7: RMSE and coverage of a 50% CI for $\bar{\rho}_{jk}$ from the posterior samples for Friedman #3 with $n_{\text{train}} = n_{\text{test}} = 500$ for binary outcomes.

	RMSE		**CI coverage**	
	suBART	**BayesSUR**	**suBART**	**BayesSUR**
$d = 2$				
ρ_{12}	0.06	0.26	0.50	0.00
$d = 3$				
ρ_{12}	0.05	0.28	0.50	0.00
ρ_{13}	0.08	0.13	0.51	0.12
ρ_{23}	0.09	0.10	0.48	0.40

CHAPTER 5

Incorporating smoothness in Bayesian additive regression trees via penalised splines

Bayesian additive regression trees (BART) have emerged as a prominent ensemble method across a diverse range of predictive tasks. Its appeal lies in its consistent ability to provide accurate predictions while simultaneously offering robust measures of uncertainty. This is achieved by aggregating a set of 'weak' tree models, each contributing a small share to explain the expected conditional mean of the response variable through carefully chosen prior settings. However, due to the inherent additive piecewise-constant nature of its base learners, BART may suffer from an assumption of lack of smoothness, which can be violated in various contexts. In this study, we propose a novel extension to the BART algorithm by incorporating Bayesian penalised splines within terminal nodes, thereby introducing greater flexibility to approximate smooth functions. This approach also facilitates flexible variable selection within the additive model framework, providing an alternative for determining which basis functions should be included in the model. Further flexibility is achieved by accounting for smooth interactions, beyond the interactions already handled by the standard BART model. We evaluate the performance of our proposed method using both simulated examples and one real-data benchmark. We compare our novel approach to related competitors, both tree-based and additive, and provide comprehensive insights into its effectiveness and applicability.

5.1 Introduction

Bayesian additive regression trees (BART) are a prominent ensemble method composed of Bayesian decision trees proposed by Chipman et al. (2010). BART methodology is increasingly recognised for its great predictive performance and its capability to provide accurate uncertainty quantification, a trait that distinguishes it from other statistical learning models (Hill et al., 2020), without the need for strong parametric assumptions about the model. One of BART's main features is the principled approach to regularisation of the trees that compose the aggregation, being addressed through the prior specification of the tree structure and its parameters. The flexibility from the ensemble of trees allows BART models to account for non-linearities and low-order interactions. Examples of successful applications of BART exist across various domains, including predicting daily global and diffuse solar radiation (Wu et al., 2021), competing risk analysis (Sparapani et al., 2020), spatial data analysis (Kim, 2022), and environmental modelling (Cao et al., 2023). Despite its effectiveness and flexibility, the original BART model is subject to certain assumptions and limitations inherent to its initial formulation. A notable one is the lack of smoothness in the model, a characteristic intrinsic to the decision tree-based approach from which BART constructs its estimations. These predicted values are represented by step-wise functions, delineated by the specifications of mean parameters from the terminal nodes. Consequently, even if the additive component from BART aids to improve the generalisation of the estimated function, the resulting model fundamentally retains a non-smooth essence.

Addressing this challenge, Linero and Yang (2018) introduced an enhancement to the BART methodology by incorporating so-called "soft-trees". This approach enables predictions for an observation within a terminal node to be derived not solely from the parameter associated to the current leaf but from a weighted combination of parameters across various terminal nodes within a tree. Furthermore, Prado et al. (2021) proposed another extension aimed at mitigating BART inherent lack of smoothness by integrating linear models within the terminal nodes of the trees. Maia et al. (2024) (corresponding to Chapter 3 of this thesis) proposed a novel method employing intrinsically smooth Gaussian process (GP) priors over the parameters in the leaves. In the innovative extension presented within this chapter, we propose to enhance the BART framework by incorporating additive

models, specifically penalised splines, into the tree structure. Consequently, we have designated this method as spBART.

Initially introduced by Friedman and Stuetzle (1981) and further developed in Buja et al. (1989), the concept of using additive models as the sum of smooth linear functions has been established as an effective method for capturing non-linearity within a multiple linear regression framework. Splines, a class of these functions, have been extensively employed in statistical models to add flexibility to the estimation of the model, as noted by Smith (1979) and Wegman and Wright (1983). Although splines relax the assumption of linearity regarding a set of predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$, they have the drawback of requiring the user to include extra parameters, including the number and position of knots which form the basis functions over which the function is approximated. Previous work from Friedman and Silverman (1989) and Kooperberg and Stone (1992) developed schemes to optimise this choice. A divergent approach was later proposed by other researchers whereby, instead of selecting an optimal number of knots, the strategy involved using a large number of knots and an associated parameter to limit the flexibility of the fitted curve. This was achieved by penalising the second derivative of the adjusted curve, as introduced by Reinsch (1967). This method, referred to as smoothing splines, has since become a common practice in various instances within the spline literature (Gu, 2013).

Eilers and Marx (1996) suggested applying a difference of the r-th order between adjacent coefficients of B-splines (De Boor, 1972), of any degree, as an alternative of this previous approach, thus introducing penalised splines (P-splines). The Bayesian framework for P-splines, foundational to the additive models used in this chapter, was formalised by Lang and Brezger (2004). Despite the novelty of the approach we propose, previous work has already attempted to integrate Bayesian penalised splines with tree-based models; for instance, Low-Kam et al. (2015) used Bayesian trees to model threshold effects and interactions, in conjunction with penalised B-splines for smoothing dose-time response surfaces. Our model, spBART, integrates this Bayesian P-splines framework with BART, aiming to achieve a predictive model that is smoother than the original Bayesian ensemble. Additionally, it enhances model flexibility as the set of basis functions effectively

used in the model are determined by the tree sampling, when compared with the traditional Bayesian P-splines which require the user to pre-specify which covariates are going to be modelled via the basis functions.

Within the context of spline regression and additive models, it is pertinent to acknowledge the development of multivariate adaptive regression splines (MARS), as proposed by Friedman (1991). MARS facilitates the identification and adaptation of basis functions through a recursive partitioning approach within regression models. This method offers flexibility in selecting knots and yields an interpretable model capable of selecting and identifying significant variables and interactions. A Bayesian framework for MARS was further developed by Denison et al. (1998), wherein a probability distribution over the space of possible MARS models is explored. This exploration incurs the use of the reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm, introduced by Green (1995), to accommodate changes in the size of the parameter space. Denison et al. (1998) positions Bayesian MARS as inherently related to Bayesian CART algorithm, as discussed by both Denison et al. (1998) and Chipman et al. (1998), since both involve a Bayesian setting for a partition model. Our proposed spBART is closely related to these models; however, unlike Bayesian MARS, which relies on the RJ-MCMC sampler to exclusively define the basis functions and interactions, spBART employs penalised splines over a fixed set of basis functions. Additionally, the trees aid the identification of any existing partitions within the feature space, and determine the subset of important variables and interactions.

The remainder of this chapter is organised as follows: Section 5.2 describes the methodology underpinning spBART along with the mathematical foundation of the proposed model and its specifications. Section 5.3 gives more details of the sampling algorithm. Section 5.4 presents two simulation examples to highlight the features of this novel ensemble model and to facilitate comparison with existing competitors. Section 5.5 provides a brief application to a real dataset from a meteorological application, followed by a concise final discussion in Section 5.6 on the model's conclusions and limitations. An implementation of our method is available in the R package `spBART`[2], which was used to obtain all presented results.

[2]Accessible at https://github.com/MateusMaiaDS/spBART.

5.2 spBART: penalised splines Bayesian additive regression trees

5.2.1 Bayesian P-splines

Initially, it is essential to outline the setting of the Bayesian splines used throughout this chapter. The setup largely aligns with the the Bayesian P-splines proposed by Lang and Brezger (2004), to which we refer the reader for more detail. In the specified model $y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n$, with ε_i following a normal distribution $\mathrm{N}(0, \tau^{-1})$ with residual precision τ, the estimation of $f(\mathbf{X})$ is given by additive functions $\hat{f}(\mathbf{X}) = \gamma + \sum_{j=1}^{p} s(\mathbf{x}^{(j)})$. Here $s(\mathbf{x}^{(j)}) = \mathbf{B}^{(j)}\boldsymbol{\theta}^{(j)}$ corresponds to the linear combination of the basis functions from the predictor $\mathbf{x}^{(j)} \in \mathbb{R}^n$, and γ is an intercept. These bases $\mathbf{B}^{(j)} \in \mathbb{R}^{n \times K}$ are constructed using cubic B-spline basis functions, where K denotes the number of basis functions, corresponding to the number of internal knots. To prevent over-fitting associated with a larger number of knots, as discussed by Eilers and Marx (1996), an r-th order penalty is applied to the coefficient estimates. This is facilitated by the prior specification of the parameters for each basis function:

$$\begin{aligned} \boldsymbol{\theta}^{(j)} &\sim \mathrm{N}\left(0, (\tau\lambda_j \mathbf{P}_r)^{-1}\right) \\ \lambda_j &\sim \mathrm{G}\left(a_{\lambda_j}, d_{\lambda_j}\right), \end{aligned}$$

where $\mathbf{P}_r = \mathbf{D}_r^\top \mathbf{D}_r$ is the penalty matrix generated from the difference matrix of the r-th order $\mathbf{D}_r$. To illustrate, assuming a second-order penalty, $\mathbf{D}_2$ is given by:

$$\mathbf{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{bmatrix}_{(K-r)\times K} .$$

It is observed that the the prior for $\boldsymbol{\theta}^{(j)}$ is improper due to rank deficiency at the r-th level, which can potentially lead to numerical and computational instabilities in subsequent calculations. To address this, in line with the transformation proposed by Eilers (1999), we employ a reparameterisation of the basis functions by setting $\mathbf{C}^{(j)} = \mathbf{B}^{(j)}\mathbf{D}_r(\mathbf{D}_r^\top \mathbf{D}_r)^{-1}$, whereby the basis function $\mathbf{C}^{(j)}$ inherently accounts for

the difference penalty. Consequently, $\mathbf{P}_r = \mathbf{I}_K \; \forall \, r$. Another important aspect is regarding the number of knots that are used. As a default, we set $K = 20$ as being sufficiently high enough for a second-order penalty. Lastly, due to identifiability and numerically stability, using an approach similar to Durbán and Currie (2003) and Lang and Brezger (2004), we transform to $\mathbf{C}^{(j)}$ to be a centered penalised basis function. The centered B-spline implies a rank reduction; according to Wood (2017), it is recommended to set the K-th column to zero and delete it. Hence, we obtain $\mathbf{C}^{(j)} \in \mathbb{R}^{n \times (K-1)}$. This centering has also been justified under the Bayesian perspective from the point of view of model selection (George and McCulloch, 1993); it amounts to initially integrating out the intercept parameter, if one is included in the model.

These basis functions, although being flexible enough to model non-linear effects from any predictor $\mathbf{x}^{(j)}$, do not cover the effect of interactions. In the literature, it is common to use the tensor product between two one-dimensional B-splines in order to model interactions. For instance, for $j \neq k$, we have that $s(\mathbf{x}^{(j)}, \mathbf{x}^{(k)}) = (\mathbf{B}^{(j)} \otimes \mathbf{B}^{(k)})\boldsymbol{\theta}^{(j,k)}$, where $\otimes$ denotes the tensor product. The resulting interaction basis would have its own penalty matrix; Besag and Kooperberg (1995) suggest using a prior specification to penalise the coefficients from the interaction basis based on the four nearest neighbour terms. Another choice for the prior is based on the Kronecker product of the penalty of the main effects $j \neq k$; see Clayton (1995) for further details. However, as here we use the reparameterisation $\mathbf{C}^{(j)}$ of the B-splines, the interaction bases are given simply by the tensor product $\mathbf{C}^{(j,k)} = \mathbf{C}^{(j)} \otimes \mathbf{C}^{(k)}$. Consequently, the penalty matrix for these basis still would be given by the identity matrix $\mathbf{I}_{(K-1)^2}$. For notational brevity, and because we consider only two-way interactions in the terminal nodes (other low-order interactions can be accounted for by the trees), in subsequent sections of this chapter the total number of basis functions for a given set of p predictors is denoted as $d = p + \binom{p}{2}$, where $\mathbf{C}^{(j)} \in \mathbb{R}^{n \times K^\star}$. The number of columns $K^\star$ depends on whether $\mathbf{C}^{(j)}$ refers to a set of basis functions related to a main effect or an interaction, such that

$$K^\star = \begin{cases} K - 1 & \text{if } j \leq p \\ (K-1)^2 & \text{otherwise.} \end{cases} \tag{5.1}$$

This is valid for all $j = 1, \ldots, d$, where $\mathbf{C}^{(j)}$ represents all reparameterised centered penalised bases. Recall that the indices $j \leq p$ refer to main effects, while higher indices indicate two-way interactions. Three-way interactions in the spline basis functions are not considered, as they quickly become computationally infeasible.

5.2.2 spBART

Suppose that for a given set of observed predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$, the goal of the model is to estimate the parameters of the conditional distribution of a dependent variable $\mathbf{y} \mid \mathbf{X}$, which is assumed to follow a normal distribution so that:

$$y_i \mid \mathbf{x}_i \sim \mathrm{N}\left(\sum_{t=1}^{T} g(\mathbf{x}_i, \mathcal{T}_t, \boldsymbol{\Theta}_t), \tau^{-1}\right),$$

where $\mathcal{T}_t$ refers to the tree structure with the partitions conditioned on the feature space of $\mathbf{X}$. The parameter set $\boldsymbol{\Theta} = (\{\gamma_{t1}, \mathbf{z}_{t1}, \boldsymbol{\theta}_{t1}\}, \ldots, \{\gamma_{tb_t}, \mathbf{z}_{tb_t}, \boldsymbol{\theta}_{tb_t}\})$ encompasses all parameters associated the additive models within the b_t terminal nodes of each tree t. A distinguishing feature of this model, compared to BART, is the specification of priors within terminal nodes which is determined by additive models rather than only a mean parameter. Consequently, the value assigned to an observation $\mathbf{x}_i$ within a terminal node is given by the sum of the intercept $\gamma_{t\ell}$ and the d additive components $\sum_{j=1}^{d} z_{t\ell}^{(j)} \mathbf{C}_{t\ell}^{(j)} \boldsymbol{\theta}_{t\ell}^{(j)}$. Here, the vector $\mathbf{z}_{t\ell}^{(j)} = \{z_{t\ell}^{(1)}, \ldots, z_{t\ell}^{(d)}\} \in \mathbb{R}^d$ works as an indicator vector to define whether the j-th set of basis functions and its coefficients are included in the model of the terminal node. The term $\mathbf{C}_{t\ell}^{(j)}$ corresponds to the subset of all rows from the set of basis functions $\mathbf{C}^{(j)}$ — constructed from the entire $\mathbf{x}^{(j)}$ — associated with the split rules that lead to node ℓ of the tree t. Finally, $\boldsymbol{\theta}_{t\ell}^{(j)}$ represents the vector of coefficients associated with each set of basis functions for variable j inside a given leaf.

For clarity, Figure 5.2.1 illustrates an example of spBART with two trees. Given predictors $\mathbf{X} \in \mathbb{R}^{n \times 3}$ and considering all possible two-way interactions, the model constructs a total of $d = 6$ sets of basis functions to account for both main effects and interactions. For instance, the indicator vector of the first terminal node in $\mathcal{T}_1$ is denoted by $\mathbf{z}_{11} = \{1, 0, 0, 0, 0, 0\}$ and the corresponding basis function is $\mathbf{C}_{11}^{(1)} \in \mathbb{R}^{n_{11} \times K^\star}$, where n_{11} is the number of observations where $x_i^{(1)} \leq 0.3$. For the second tree $\mathcal{T}_2$, a similar structure is observed; the leaf indicators $\mathbf{z}_{21} = \mathbf{z}_{22} =$

$\{0, 0, 1, 0, 0, 0\}$ imply that the models for each leaf solely incorporate the bases related to $\mathbf{x}^{(3)}$.

The role of $\mathbf{z}_{t\ell}$ extends beyond indicating the bases and parameters of the model; it ensures a fixed number of parameters across different nodes. Despite not having all d basis functions and vectors of predictors effectively contributing to the model, they are still present but can have their effect diminished, if necessary, by the λ_j parameter, thereby eliminating the need for techniques such as RJ-MCMC during the sampling algorithm. It is also crucial to note that the tree split rules are on $\mathbf{X}$ rather than the re-parameterised space of the basis functions, meaning that the basis functions within each leaf are essentially a subset of the original set of basis functions $\mathbf{C}^{(j)}$ and are not regenerated for the data assigned to each terminal node.

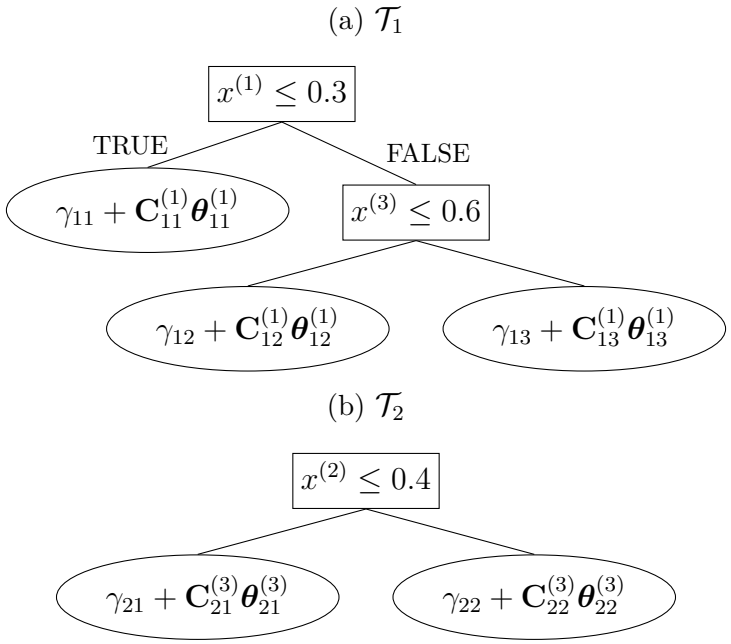


Figure 5.2.1: An example of two trees generated by spBART; the split rules are subject to the original feature space $\mathbf{X}$, and terminal node values are determined by an intercept and a subset of additive functions per leaf.

A key aspect of the model is that each tree is limited to a single main effect basis. The interaction basis functions, if included, are related solely to this primary predictor within a given tree. This constraint is crucial for several reasons: first,

by allocating only one main effect per tree, marginal effects can be easily recovered; the opposite would not be possible. By including all predictors within nodes, any tree split would represent an interaction of all basis functions with the splitting variable. Second, it narrows the search space for each tree, facilitating better convergence. For instance, consider a dataset with $p = 10$ predictors, where the true function $f(\mathbf{x_i})$ depends only on one variable. Permitting the use of all available sets of basis functions would result in a total of fifty-five potential terms (10 main effects and 45 interactions) to be included in the model, a considerably larger sample than that generated by the true process. By limiting the model to one main effect per tree, the model space is reduced to ten possible models (each capturing one main effect and its interactions bases), resulting in a more feasible search. Moreover, this restriction does not compromise the final model representation, as the aggregation of trees enables the recovery of a model encompassing all main effects and interactions, if necessary. In line with the standard BART, we can define the prior distribution for the tree structure and the terminal node parameters to be independent. Therefore, the complete prior is described by

$$\pi\left(\left(\mathcal{T}_1, \boldsymbol{\Theta}_1\right), \ldots, \left(\mathcal{T}_t, \boldsymbol{\Theta}_t\right), \lambda_1, \ldots, \lambda_d, \tau\right) = \pi\left(\tau\right) \times \prod_{j=1}^{d} \pi\left(\lambda_j\right) \times \prod_{t=1}^{T} \pi\left(\mathcal{T}_t, \boldsymbol{\Theta}_t \mid \boldsymbol{\lambda}, \tau\right),$$

where

$$\begin{aligned}\pi\left(\mathcal{T}_t, \boldsymbol{\Theta}_t \mid \boldsymbol{\lambda}, \tau\right) &= \pi\left(\boldsymbol{\Theta}_t \mid \mathcal{T}_t, \boldsymbol{\lambda}, \tau\right) \times \pi\left(\mathcal{T}_t\right) \\ &= \left[\prod_{\ell=1}^{b_t} \pi\left(\boldsymbol{\theta}_{t\ell} \mid \mathbf{z}_{t\ell}, \boldsymbol{\lambda}, \tau\right) \times \pi\left(\mathbf{z}_{t\ell}\right) \times \pi\left(\gamma_{t\ell}\right)\right] \times \pi\left(\mathcal{T}_t\right).\end{aligned}$$

We define the priors for $\mathcal{T}_t$ and $\gamma_{t\ell}$ following Chipman et al. (2010), despite adopting a different setting for the tree hyperparameters. For τ and λ_j, a data-driven approach is employed to establish their prior distributions. Hyperparameters associated with these distributions are omitted in the preceding equations for brevity but are comprehensively specified subsequently, along with the reasoning for each component of the prior.

5.2.3 The tree structure

The tree prior, in alignment with the Bayesian CART algorithm (Chipman et al., 1998), assumes the probability of a node being non-terminal is given by $\alpha(1+\nu_{t\ell})^{-\beta}$,

where $\nu_{t\ell}$ denotes the depth of node ℓ in the tree t, where α and β are hyperparameters penalising the node depth. As a distinction from the standard Bayesian CART setting, taking account that the presence of additive functions within a terminal node can sufficiently model the main effect of a predictor, we adjust α to 0.5, since the original prior excessively favours stumps being non-terminal. The distribution regarding the splitting variables and splitting rules remains the same as per Chipman et al. (1998), with a uniform prior of selecting any of the available $j = 1, \ldots, p$ variables, and uniformly selecting available values from the discrete set of cut-points of the selected j-th variable. The proposals specific to the tree structure (i.e., modifications to the number of nodes, split variables, and cut-points of a tree $\mathcal{T}_t$), are confined to grow, prune, and change movements. This approach aligns with the strategies suggested by Kapelner and Bleich (2016) to diminish the computational complexity inherent in the calculations for the back-fitting MCMC algorithm, as we elaborate later in Section 5.3.

5.2.4 The prior on terminal node parameters

As mentioned above, the model described within a terminal node is given by

$$\gamma_{t\ell} + \sum_{j=1}^{d} z_{t\ell}^{(j)} \mathbf{C}_{t\ell}^{(j)} \boldsymbol{\theta}_{t\ell}^{(j)},$$

where $\gamma_{t\ell}$ denotes the intercept of the model, $z_{t\ell}^{(j)}$ is the indicator of the contribution of the coefficients $\boldsymbol{\theta}_{t\ell}^{(j)} \in \mathbb{R}^{K^\star}$ associated to the centered penalised basis functions $\mathbf{C}_{t\ell}^{(j)} \in \mathbb{R}^{n_{t\ell} \times K^\star}$. Before the basis generation, all predictors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}$ are normalised between $[0, 1]$ and the B-splines for the main effects are generated. The K knots are equally spaced between the interval $[-\omega_{x^{(j)}}, 1+\omega_{x^{(j)}}]$ where $\omega_x^{(j)}$ is the standard deviation obtained for each component j; this strategy avoids extrapolation problems from the basis functions (see Eilers et al., 2015, for further details about the P-splines design). For numerical stability and to aid prior elicitation, the $\mathbf{y}$ is scaled to be between $[-0.5, 0.5]$.

5.2.4.1 The prior on the intercept parameter

For the intercept parameter $\gamma_{t\ell}$, the prior is defined as $\gamma_{t\ell} \sim \mathrm{N}(0, \tau_\gamma^{-1})$, where $\tau_\gamma = 4\kappa^2 T$. This hyperparameter choice for the intercept aligns with the BART prior,

positioning spBART as a generalisation of BART with the presence of additive models in the terminal nodes. The intercept plays a crucial role in modelling breakpoints or discontinuities within contributions for any set of basis functions, thereby enhancing the model's flexibility beyond what could be achieved with only a Bayesian P-splines model.

5.2.4.2 The prior on basis indicator parameters $\mathbf{z}_{t\ell}$

For the d-dimensional indicator vector $\mathbf{z}_{t\ell}$ of the basis functions $j = 1, \ldots, d$, d independent Bernoulli prior distributions are assigned, denoted by $z_{t\ell}^{(j)} \overset{i.i.d.}{\sim}$ Bernoulli(0.5). Specifying all success probabilities as 0.5 reflects the absence of prior knowledge regarding the importance of variables and interactions, implying that the model should be equally likely to include or exclude a set of basis functions. Although, in principle, these priors could be calibrated to favour certain interactions or main effects by adjusting the probabilities for different sets of basis functions j, the default approach remains non-informative in terms of variable selection. Furthermore, it simplifies the MH step as the priors cancel out in the acceptance ratio as a result. However, we emphasise the distinction between the prior and proposal distributions and stress that we allow $z_{t\ell}^{(j)} = 1$ for only one main effect at most where $j \leq p$. Otherwise, the indicator takes the value 0 for all $p-1$ remaining main effects. The indicator for the interaction effects can only take the value 1 for the set of $p-1$ candidate interactions which comprise the included main effect, or the last included main effect, and take the value 0 otherwise. We achieve these restrictions in practice by never proposing invalid trees which are in violation of these conditions when performing the MH step.

As the $\mathbf{z}_{t\ell}$ vector indicates which set of bases are effectively included in the model, it has a direct role in terms of variable importance. Indeed, an additional feature of this vector, besides avoiding the need for a RJ-MCMC sampler, is that its posterior samples can be used to construct an interpretable measure of variable importance. To summarise $\mathbf{z}_{t\ell}$ among all trees, we compute the following proportion for each posterior sample (where we omit the iteration index for notational clarity):

$$\Delta_j = \frac{\sum_t^T \sum_{\ell=1}^{b_t} z_{t\ell}^{(j)}}{\sum_{j=1}^d \sum_t^T \sum_{\ell=1}^{b_t} z_{t\ell}^{(j)}}. \tag{5.2}$$

5.2.4.3 The prior on the number of sets of basis functions $m_{t\ell}$

In the BART model, the prior on the tree structure $\mathcal{T}_t$ indirectly governs the quantity of leaf parameters. Owing to the inclusion of a novel set of parameters in the terminal nodes, it is imperative to regularise the number of sets of basis functions $m_{t\ell}$ now used in the leaves. To this end, a prior is imposed on each terminal node concerning this count, without distinguishing between interactions and main effects. To encourage parsimony in the model, we adopt a zero-truncated Poisson prior $\pi(m_{t\ell}) \sim \text{ZTP}(\psi_m = 0.1)$, with $m_{t\ell} = 1, 2, 3, \ldots$, as we only consider a terminal node to be valid if it has at least one set of basis functions of any type. This rate choice concentrates most of the density on one or two sets of basis functions, such that settings with an excess of bases incur substantial penalisation.

5.2.4.4 The prior on the basis coefficients

The prior for the coefficients $\boldsymbol{\theta}_{t\ell}$ is established within a hierarchical framework, accounting for penalization within the basis functions:

$$\begin{aligned}
\boldsymbol{\theta}_{t\ell}^{(j)} \mid \lambda_j, \tau &\sim \text{N}\left(\mathbf{0}, (\tau\lambda_j)^{-1}\,\mathbf{I}_{K^\star}\right) \\
\lambda_j &\sim \text{Gamma}\left(a_{\lambda_j}, d_{\lambda_j}\right),
\end{aligned}$$

where $K^\star$ is given by Equation (5.1). A critical aspect of the model is the interpretation of the prior for the λ_j parameter. Unlike standard BART, where the distribution of the induced prior is readily retrievable through the prior distribution of the leaf parameters, the contribution of each set of basis functions here may differ across trees, making it challenging to pre-specify a range of values for $\boldsymbol{\theta}_{t\ell}^{(j)}$ considering a fixed number of trees. Our strategy relies on the assumption that unimportant values of $\boldsymbol{\theta}_{t\ell}^{(j)}$ typically hover around zero, as the basis functions are centered. Consequently, to presume that a variable is important, a larger *a priori* variance is preferred, anticipating that the coefficients will deviate from zero. This is achieved by defining the prior for λ_j such that $\mathbb{E}[\lambda_j]$ is small.

However, defining a reasonable range of values for the coefficients $\boldsymbol{\theta}_{t\ell}^{(j)}$, and consequently their precision, is not as straightforward as prior specification for linear regression coefficients, given the additive nature of the ensemble and particularly

the structure of the basis functions. Once we have scaled $\mathbf{y}$ to fall within $[-0.5, 0.5]$, our goal is to maximise the sample variance $w(\widehat{\boldsymbol{\theta}}_{t\ell}^{(j)})$ from the coefficients from the linear combination given by $\mathbf{y}^{(j)} = \mathbf{C}^{(j)}\boldsymbol{\theta}^{(j)}$ with respect to $\boldsymbol{\theta}$, constrained on $\mathbf{y}^{(j)}$ being kept within the scaled interval. This problem is addressed through numerical optimisation, resulting in a vector of optimal values $\widehat{\boldsymbol{\theta}}^{(j)}$. Following this, the estimation for the maximum variance — or alternatively, the minimum precision — of the coefficients is determined by ${\lambda_j}^{\min} = 1/w(\widehat{\boldsymbol{\theta}}_{t\ell}^{(j)})$ where $w(\cdot)$ represents the sample variance. Given that $\mathbf{X}$ and the basis functions are scaled, this process is required to be performed only once each for both all main effect and all interaction bases to determine their respective minimum λ_k values. Subsequently, we establish the prior with default settings $a_{\lambda_k} = \lambda_k^{\min}$ and $d_{\lambda_k} = 1$, where $k \in \{1, 2\}$ indicates the type of basis; i.e., whether it is an interaction or not.

Notably, λ_j is assigned at the predictor level, meaning it holds a constant value across all trees and leaf nodes. This uniformity facilitates interpreting the value of the parameter directly with the set of bases function associated with the $\mathbf{x}^{(j)}$ predictor. Specifically, λ_j can be seen a measure of the smoothness from the associated set of basis functions, whereby smaller values yield smoother functions. Additionally, its posterior samples can be interpreted to evaluate variable importance. A variable or interaction whose marginal effect contributes insignificantly to the model is expected to be represented by a constant function, implying zero smoothness and thus larger posterior means for λ_j. Nonetheless, λ_j should not be exclusively used for variable selection, as its values may merely indicate the smoothness level; i.e, a large value of λ_j does not strictly imply that an effect is irrelevant. For a complete picture of the relevance of a given main or interaction effect, λ_j and Δ_j from Equation (5.2) should be considered jointly. Conclusions regarding the strength of a variables contribution can be further supported by visual inspection of the marginal effect surfaces.

5.2.5 The prior on the residual precision

The residual precision parameter prior is set via a conjugate gamma distribution $\tau \sim Ga(a_\tau, d_\tau)$. Given the superior flexibility of additive models for capturing non-linear relationships in comparison to linear models, there is an underlying

assumption that the residual precision of such models exceeds the one derived from a linear framework. In alignment with Chipman et al. (2010), the selection of the shape a_τ and rate d_τ hyperparameters is guided by the objective to ensure a high probability, typically $\nu_\tau = 0.9$, that τ surpasses the precision $\hat{\tau}_{OLS}$ estimated from ordinary linear regression on the predictors $\mathbf{X}$.

5.3 Posterior inference

To design the sampler for the posterior joint distribution $\pi(\mathcal{T}_1, \boldsymbol{\Theta}_1), \ldots, (\mathcal{T}_t, \boldsymbol{\Theta}_t), \ldots, \boldsymbol{\lambda}, \tau | \mathbf{y})$, the back-fitting algorithm of Hastie and Tibshirani (2000) is employed. For simplicity, $\boldsymbol{\mathcal{T}}_{(-t)}$ denotes all trees except specifically tree t, and $\boldsymbol{\Theta}_{(-t)}$ represents all parameters except those of tree t. This notation facilitates the expression of the full conditional distribution $\pi(\mathcal{T}_t, \boldsymbol{\Theta}_t \,|\, \mathbf{y}, \boldsymbol{\mathcal{T}}_{(-t)}, \boldsymbol{\Theta}_{(-t)}, \boldsymbol{\lambda}, \tau)$, which, in the back-fitting context, is characterised by the partial residuals $\mathbf{R}_t \equiv \mathbf{y} - \sum_{k \neq t}^{T} g(\mathbf{X}, \mathcal{T}_k, \boldsymbol{\Theta}_k)$, thereby being represented as $\pi(\mathcal{T}_t, \boldsymbol{\Theta}_t \mid \mathbf{R}_t, \boldsymbol{\lambda}, \tau)$.

Hence, the sampling scheme for $(\mathcal{T}_t, \boldsymbol{\Theta})$ consists of sequentially drawing from

$$\mathcal{T}_t \mid \mathbf{R}_t, \mathcal{Z}_t, \boldsymbol{\lambda}_t, \tau \tag{5.3}$$

$$\mathcal{G}_t \mid \mathbf{R}_t, \mathcal{B}_t, \mathcal{Z}_t, \boldsymbol{\lambda}, \tau \tag{5.4}$$

$$\mathcal{B}_t \mid \mathbf{R}_t, \mathcal{G}_t, \mathcal{Z}_t, \boldsymbol{\lambda}, \tau, \tag{5.5}$$

where $\mathcal{Z}_t, \mathcal{G}_t$, and $\mathcal{B}_t$ denote the collections of $\mathbf{z}_{t\ell}$, $\gamma_{t\ell}$, and $(\boldsymbol{\theta}_{t\ell}^{(1)}, \ldots, \boldsymbol{\theta}_{t\ell}^{(d)})$ parameters across all terminal nodes, respectively. The sampling strategy for Equations (5.4) and (5.5) adopts a Gibbs sampling arrangement, facilitated by the choice of conjugate priors. Analogously, the samplers for $\boldsymbol{\lambda}$ and τ employ the same approach. The sampler for $\mathcal{T}_t$ uses a Metropolis-Hastings (MH) algorithm as per Chipman et al. (2010), although some modifications are required, which we describe below for incorporating the $m_{t\ell}$, $\mathbf{z}_{t\ell}$, and $\boldsymbol{\theta}_{t\ell}^{(j)}$ parameters.

5.3.1 Metropolis-Hastings step

Omitting substantial cancellation for notational brevity, the MH-step of Equation (5.3) for a new tree proposal has an acceptance ratio given by

$$\min \left\{ 1, \frac{\pi\left(\mathbf{R}_t \mid \mathcal{T}_t^\star, \mathcal{Z}_t^\star, \boldsymbol{\lambda}, \tau\right) \pi\left(m_{t\ell}^\star\right) \pi\left(\mathcal{T}_t^\star\right) q\left(\mathcal{T}_t^\star \to \mathcal{T}_t\right)}{\pi\left(\mathbf{R}_t \mid \mathcal{T}_t, \mathcal{Z}_t, \boldsymbol{\lambda}, \tau\right) \pi\left(m_{t\ell}\right) \pi\left(\mathcal{T}_t\right) q\left(\mathcal{T}_t \to \mathcal{T}_t^\star\right)} \right\}, \tag{5.6}$$

where the likelihood term from Equation (5.6) is obtained by marginalising out the $\boldsymbol{\gamma}_{t\ell}$, $\boldsymbol{\theta}_{t\ell}^{(j)}$. The remaining quantities refer to other prior components, such as the prior distribution on the number of set of bases $\mathcal{M}_t = (m_{t1}, \ldots, m_{tb_t})$, the tree structure $\mathcal{T}_t$, as well as the associated transition probabilities $q(\cdot)$. The ratio regarding the number of sets of basis functions $\pi(m_{t\ell}^{\star})/\pi(m_{t\ell})$ is considered at the tree node level, given that MH proposals in this context exclusively alter a single terminal node at a time. Additionally, the ratio $\pi(\mathcal{Z}_t^{\star})/\pi(\mathcal{Z}_t)$ is always equal to 1, as discussed in Section 5.2.4.2.

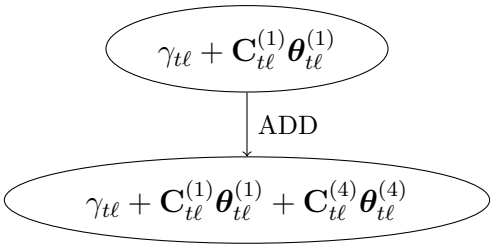
Since the tree changes are made at the level of the leaf, we are interested in the marginal distribution of a terminal node which is given by

$$\mathbf{R}_{t\ell} \mid \mathcal{T}_t, \mathbf{z}_{t\ell}, \boldsymbol{\lambda}, \tau \sim \mathrm{N}\left(\mathbf{0}, \tau^{-1}\left(\mathbf{I}_{n_{t\ell}} + \frac{\tau}{\tau_\gamma}\mathbf{1}_{n_{t\ell}}\mathbf{1}_{n_{t\ell}}^{\top} + \tau\sum_{j}^{d} z_{t\ell}^{(j)}\lambda_j^{-1}\mathbf{C}_{t\ell}^{(j)}\mathbf{C}_{t\ell}^{(j)^{\top}}\right)\right).$$

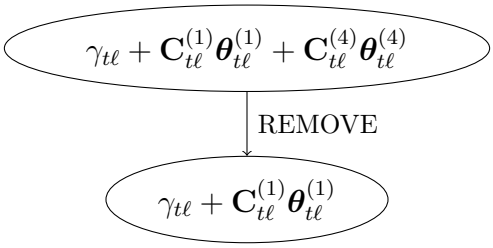
The proposals for the MH-step fall into two categories: those altering the tree structure and those affecting which basis functions contribute within a terminal node. The former includes the conventional 'grow', 'prune', and 'change' moves proposed by Chipman et al. (1998), where $\mathcal{T}_t^{\star} \neq \mathcal{T}_t$ and $\mathcal{Z}_t^{\star} = \mathcal{Z}_t$. Additionally, we introduce the 'add', 'remove', and 'modify' moves to adjust the indicator vector $\mathbf{z}_{t\ell}$ for the set of bases within terminal node ℓ of tree t, leading to $\mathcal{T}_t^{\star} = \mathcal{T}_t$ and $\mathcal{Z}_t^{\star} \neq \mathcal{Z}_t$. We also note that 'add' and 'remove' are the only moves which can lead to $m_{t\ell}^{\star} \neq m_{t\ell}$; otherwise, the associated ratio cancels in Equation (5.6) for the modify move.

To illustrate, Figure 5.3.1 depicts these proposals on a generic node with $\mathbf{z}_{t\ell} = \{1, 0, 0, 0, 0, 0\}$ for $p = 3$ and $d = 6$. In panel (a), the add move samples uniformly from a discrete grid of available two-way interactions related to $\mathbf{x}^{(1)}$, resulting in a new proposal $z_{t\ell}^{\star} = \{1, 0, 0, 1, 0, 0\}$. In panel (b), we present an analogous scenario for the remove move, sampling with equal probability a set of basis functions included in the node, leading to $\mathbf{z}_{t\ell}^{\star} = \{1, 0, 0, 0, 0, 0\}$. We note that this move is not restricted solely to interaction bases; a main effect can be removed as long there is also an interaction included in the terminal node. Lastly, in panel (c), we display the result of a modify operation; this operation randomly selects a j where $z_{t\ell}^{(j)} = 1$ within a terminal node. Subsequently, this $z_{t\ell}^{(j)}$ is set to zero, and another randomly selected valid coordinate is set to one. One of the consequences of this

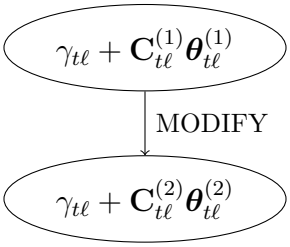
move is to replace one main effect by another, as shown in the last panel of Figure 5.3.1, where $\mathbf{z}^{\star}_{t\ell} = \{0, 1, 0, 0, 0, 0\}$ is accepted. We note that the modify move is only allowed to replace a main effect by another when the selected terminal node is a stump with no interactions. These procedures ensure the creation of only valid terminal nodes. Recall, specifically, that each leaf must be restricted to at most one main effect and any included set of interaction bases must be associated with the incorporated main effect. If a main basis is absent, the interaction should relate to the last main effect that was present. Moreover, each terminal node must hold a minimum of 25 observations, an increase from the minimum of 5 recommended by Chipman et al. (2010), to be considered valid.



(a) Add a set of basis functions.



(b) Remove a set of basis functions.



(c) Modify a set of basis functions.

Figure 5.3.1: Illustration of the novel tree-editing operations governing the basis functions in the terminal nodes. Panel (a) shows the add move; adding a set of basis functions. Panel (b) shows the remove move; removing a set of basis functions. Panel (c) shows the modify move; essentially swapping two sets of basis functions.

We set the probability of selecting the grow, prune, change, add, remove, and modify moves as $\{0.15, 0.15, 0.2, 0.15, 0.15, 0.2\}$, respectively. See the supplementary material Linero and Yang (2018) for additional details regarding the transition probabilities of a new tree. The remaining posterior samples of parameters from a leaf-node level are obtained from the full conditionals

$$\gamma_{t\ell} \mid \ldots \sim \mathrm{N}\left(s_{\gamma_{t\ell}}^{-1}\left(\sum_{i=1}^{n_{t\ell}} R_i - \mathbf{1}_{n_{t\ell}}^{\top}\sum_{j=1}^{d} z_{t\ell}^{(j)}\mathbf{C}_{t\ell}^{(j)}\boldsymbol{\theta}_{t\ell}^{(j)}\right), \tau^{-1}s_{\gamma_{t\ell}}^{-1}\right), \tag{5.7}$$

$$\boldsymbol{\theta}_{t\ell}^{(j)} \mid \ldots \sim \mathrm{N}\left(\mathbf{S}_{\theta_{t\ell}^{(j)}}^{-1}\mathbf{C}_{t\ell}^{(j)\top}\left(\mathbf{R}_{t\ell} - \left(\gamma_{t\ell}\mathbf{1}_{n_{t\ell}} + \sum_{k\neq j}^{d} z_{t\ell}^{(k)}\mathbf{C}_{t\ell}^{(k)}\boldsymbol{\theta}_{t\ell}^{(k)}\right)\right), \tau^{-1}\mathbf{S}_{\theta_{t\ell}^{(j)}}^{-1}\right), \tag{5.8}$$

where $s_{\gamma_{t\ell}} = n_{t\ell} + \frac{\tau_\gamma}{\tau}$ and $\mathbf{S}_{\theta_{t\ell}^{(j)}} = \mathbf{C}_{t\ell}^{(j)\top}\mathbf{C}_{t\ell}^{(j)} + \lambda_j\mathbf{I}_{K^\star}$.

Lastly, it is also necessary to update the parameters of the precision of the basis coefficients and the precision of the residuals. The full conditional distributions are given by

$$\lambda_j \mid \ldots \sim \mathrm{Gamma}\left(\frac{1}{2}\sum_{t=1}^{T}\sum_{\ell=1}^{b_t} z_{t\ell}^{(j)}K^\star + a_{\lambda_j}, \frac{1}{2}\sum_{t=1}^{T}\sum_{\ell=1}^{b_t} z_{t\ell}^{(j)}\boldsymbol{\theta}_{t\ell}^{(j)\top}\boldsymbol{\theta}_{t\ell}^{(j)} + d_{\lambda_j}\right) \tag{5.9}$$

$$\tau \mid \ldots \sim \mathrm{Gamma}\left(\frac{n}{2} + \frac{1}{2}\sum_{j=1}^{d}\sum_{t=1}^{T}\sum_{\ell=1}^{b_t} z_{t\ell}^{(j)}K^\star + a_\tau, \right.$$
$$\left. \frac{1}{2}(\mathbf{y}-\hat{\mathbf{y}})^{\top}(\mathbf{y}-\hat{\mathbf{y}}) + \frac{1}{2}\sum_{j=1}^{d}\sum_{t=1}^{T}\sum_{\ell=1}^{b_t}\lambda_j z_{t\ell}^{(j)}\boldsymbol{\theta}_{t\ell}^{(j)\top}\boldsymbol{\theta}_{t\ell}^{(j)} + d_\tau\right). \tag{5.10}$$

The algorithm initialises with a set of p trees, wherein each $\mathbf{z}_{t1}$ is configured as a zero vector, except for $z_{t1}^{(t)} = 1$, for all trees $t = 1, \ldots, p$. By having the number of trees equal to the number of main predictors, each tree $t = 1, \ldots, p$ is initiated with one corresponding main effect $j = 1, \ldots, p$, thereby avoiding any initial bias towards specific covariates, and allowing a complete model to be specified if all predictors are important. Although we adopt $T = p$ as a default setting, it is possible to initialise fewer (or more) trees than there are main effects. However, if the number of trees in the model is smaller than the number of important variables, the model will be misspecified, as only one main effect is allowed by each tree at most. Subsequently, the algorithm proceeds to sample all outlined parameters over $N_{\text{MCMC}} = 5000$ iterations, of which $N_{\text{burn-in}} = 3000$ are discarded as burn-in

samples. The MCMC default setting was determined based on the convergence analysis of posterior samples on the experiments applied herein. The complete process is summarised in Algorithm 5.1.

Algorithm 5.1: spBART sampling algorithm.

Input: $\mathbf{X}$, $\mathbf{y}$, T, N_{MCMC}, $N_{\text{burn-in}}$, and all hyperparameters of the priors.

Initialise: $\boldsymbol{\mathcal{T}}^{(t)}$ tree stumps where only $z_{t1}^{(t)} = 1$ with $\gamma_{t\ell} = \boldsymbol{\theta}_{t\ell}^{(j)} = 0 \,\forall\, (t, j, \ell)$, and $\tau = 1$.

for *iterations* h *from* 1 *to* N_{MCMC} **do**
  for *trees* t *from* 1 *to* T **do**
    Calculate the partial residuals $\mathbf{R}_t$;
    Propose a new tree $\mathcal{T}_t^{(j)\star}$ by a grow, prune, or change move[a], or a new indicator vector $\mathbf{z}_{t\ell}^{\star}$ by an add, remove, or modify move;
    Accept and update $\mathcal{T}_t^{(j)} = \mathcal{T}_t^{(j)\star}$ and $\mathcal{Z}_t^{(j)} = \mathcal{Z}_t^{(j)\star}$ with probability

$$\gamma^{\star}\left(\mathcal{T}_t^{(j)}, \mathcal{T}_t^{(j)\star}\right) = \min\left\{1, \frac{\pi\left(\mathbf{R}_t \mid \mathcal{T}_t^{\star}, \mathcal{Z}_t^{\star}, \boldsymbol{\lambda}, \tau\right) \pi\left(\mathcal{T}_t^{\star}\right) q\left(\mathcal{T}_t^{\star} \to \mathcal{T}_t\right)}{\pi\left(\mathbf{R}_t \mid \mathcal{T}_t, \mathcal{Z}_t, \boldsymbol{\lambda}, \tau\right) \pi\left(\mathcal{T}_t\right) q\left(\mathcal{T}_t \to \mathcal{T}_t^{\star}\right)}\right\}.$$

    for *terminal nodes* ℓ *from* 1 *to* b_t **do**
      Update $\gamma_{t\ell} \mid \mathbf{R}_t, \boldsymbol{\theta}_{t\ell}^{(j)}, \mathbf{z}_{t\ell}, \boldsymbol{\lambda}, \tau$ using Equation (5.7).
      for j *from* 1 *to* d **do**
        Update $\boldsymbol{\theta}_{t\ell}^{(j)} \mid \mathbf{R}_t, \gamma_{t\ell}, \mathbf{z}_{t\ell}, \boldsymbol{\lambda}, \tau$ using Equation (5.8).
      end
    end
  end
  for j *from* 1 *to* d **do**
    Update $\lambda_j \mid \mathbf{y}, \boldsymbol{\Theta}$ using Equation (5.9).
  end
  Update $\tau \mid \mathbf{y}, \Theta$ using Equation (5.10).
end

[a] See Kapelner and Bleich (2016) for further details on these tree proposal steps and transition probabilities $q(\cdot)$.

5.4 Simulation experiments

The simulation studies are conducted in order to highlight the promising features of spBART and showcase tasks where the model can be reasonably anticipated to perform well. First, the model should be able to provide smooth estimations, as it uses splines as a building-block component. Secondly, due to the tree-based approach, it should be possible to identify any existing change-points within the main effects and/or interactions. These aspects can be readily observed through marginal effects plots, which are easily derived from the additive model setting, thus avoiding the need for partial dependence plots — commonly used within BART literature (Kapelner and Bleich, 2016; Chipman et al., 2010) — which can be computationally intensive. Furthermore, the model also should be able to perform, at some level, an automatic model specification for the additive functions. This is achieved through the add, remove, and modify moves, which determine which set of basis functions are relevant to estimate the conditional expectation of $\mathbf{y}\,|\,\mathbf{X}$, as well as the λ_j parameter which controls the smoothness of each set of basis functions. To assess these aspects of spBART, two versions of the Friedman equation (Friedman, 1991) are presented. The first one follows the original formulation, while the second — which we refer to as the 'Friedman break' data — introduces change-points to the marginal effects of $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$. Both are explained in further detail on Section 5.4.1 and Section 5.4.2, respectively. Though we display the marginal effects for certain relevant main and interaction effects which effectively contribute to the model in both simulations, we note that we provide a complete overview of all main marginal effect estimates $\forall\, j = 1, \ldots, p$ in each case in Appendix 5.A.

In evaluating predictive performance, the model is compared using ten replications of the data, with training and test dataset sizes varying among $n_{\text{train}} = n_{\text{test}} = \{250, 500, 1000\}$. To measure the accuracy of predictions and the quality of uncertainty quantification, we use the root mean squared error (RMSE) and the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007) over the test dataset, respectively. Both metrics have the property of lower scores indicating superior model fits, with RMSE focusing on mean predictive performance while CRPS covers uncertainty calibration.

Our model is evaluated in comparison with tree-based models such as BART and some of its variants, including SoftBART (Linero and Yang, 2018) and model trees BART (MOTR-BART; Prado et al., 2021), which claim to mitigate the assumption of non-smoothness. The experiments further extend to models which perform an adaptive selection of basis functions. Aiming to assess their performance relative to some of the most advanced techniques currently employed in the domain of additive models, we used models such as MARS (Friedman, 1991) and Bayesian MARS (Denison et al., 1998). These models are fitted, respectively, using `R` packages `dbarts` (Dorie et al., 2024), `softBART` (Linero, 2022b), `MOTRbart` (Prado et al., 2021), `earth` (Milborrow, 2024), and `BASS` (Francom and Sansó, 2020) with their respective default settings, except for the `earth` package where we modify the main function to include the pairwise interactions.

5.4.1 Friedman data

The Friedman function is described by

$$y_i = 10 \sin\left(\pi x_i^{(1)} x_i^{(2)}\right) + 20\left(x_i^{(3)} - 0.5\right)^2 + 10x_i^{(4)} + 5x_i^{(5)} + \epsilon_i,\ i = 1, \ldots, n, \quad (5.11)$$

where $x_i^{(j)}$ follows a uniform distribution, $x_i^{(j)} \sim \text{Unif}(0, 1)$ for all $j = 1, \ldots, p$, and the error term ϵ_i is normally distributed $\epsilon_i \sim \text{N}(0, \tau^{-1})$. Notably, the predictors $\mathbf{x}^{(6)}$ through $\mathbf{x}^{(10)}$ are noise variables, being independent of the response variable $\mathbf{y}$. A key aspect of this simulation includes the smooth interaction effect between the pair $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, which spBART is expected to identify effectively. Additionally, given that the first two additive terms from Equation (5.11) exhibit non-linearity, they are presumed to be more accurately represented using smooth functions.

Figures 5.4.1 and 5.4.2 summarise the result for RMSE and CRPS over the test samples for all different sample sizes. Among all evaluated models, BART exhibits the least favorable performance which is likely due to its assumptions around lack of smoothness. The results indicate that spBART generally surpasses all its competitors in performance, achieving lower values of RMSE and CRPS, with the exception of SoftBART, which exhibits a performance equivalent to that of spBART. The performance of MARS and BASS notably deteriorates relative to spBART at larger sample sizes.

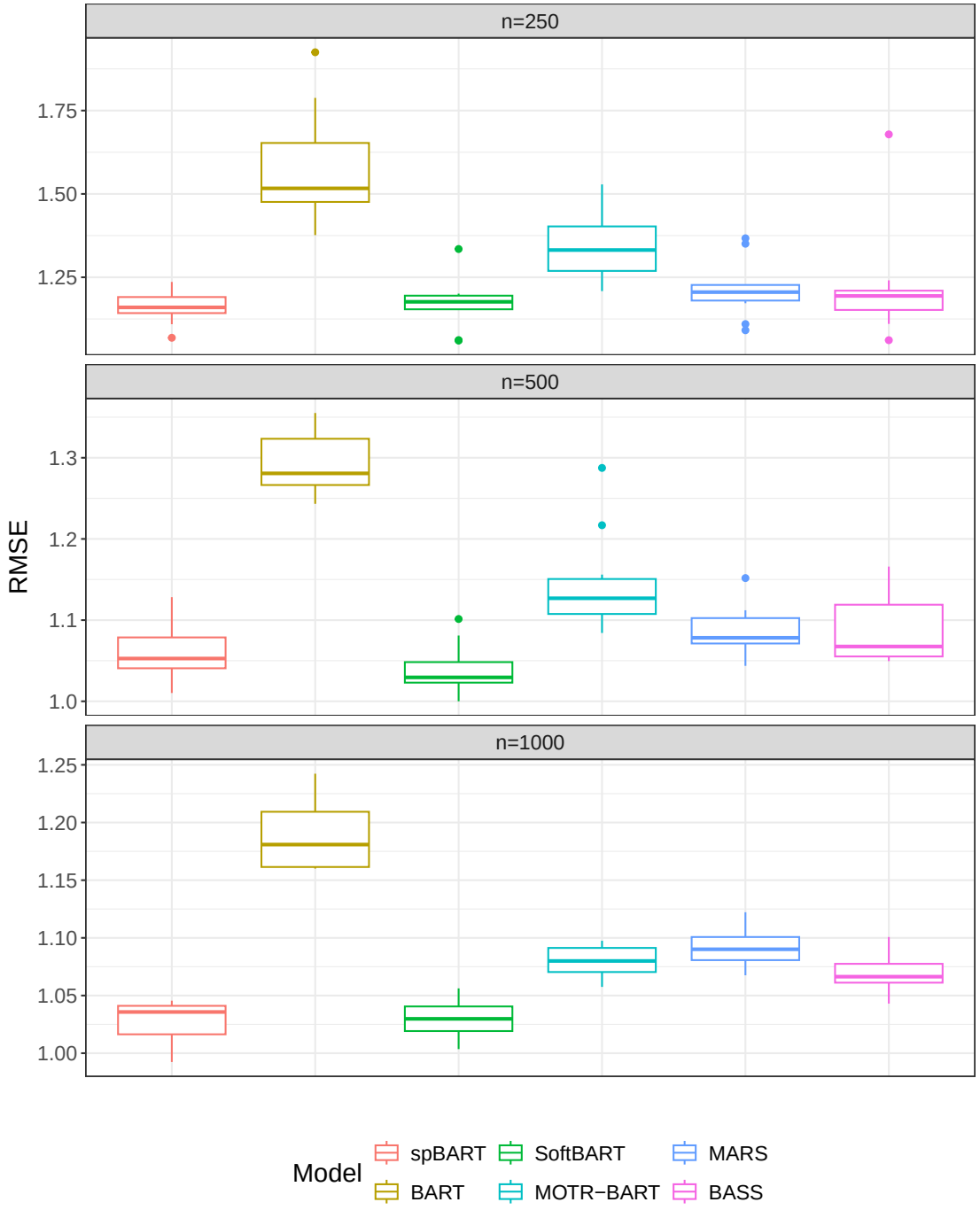


Figure 5.4.1: Comparisons between the RMSE calculated on the test samples by the competing models for the Friedman data using ten replications over different sample sizes $n_{\text{train}} = n_{\text{text}} = \{250, 500, 1000\}$.

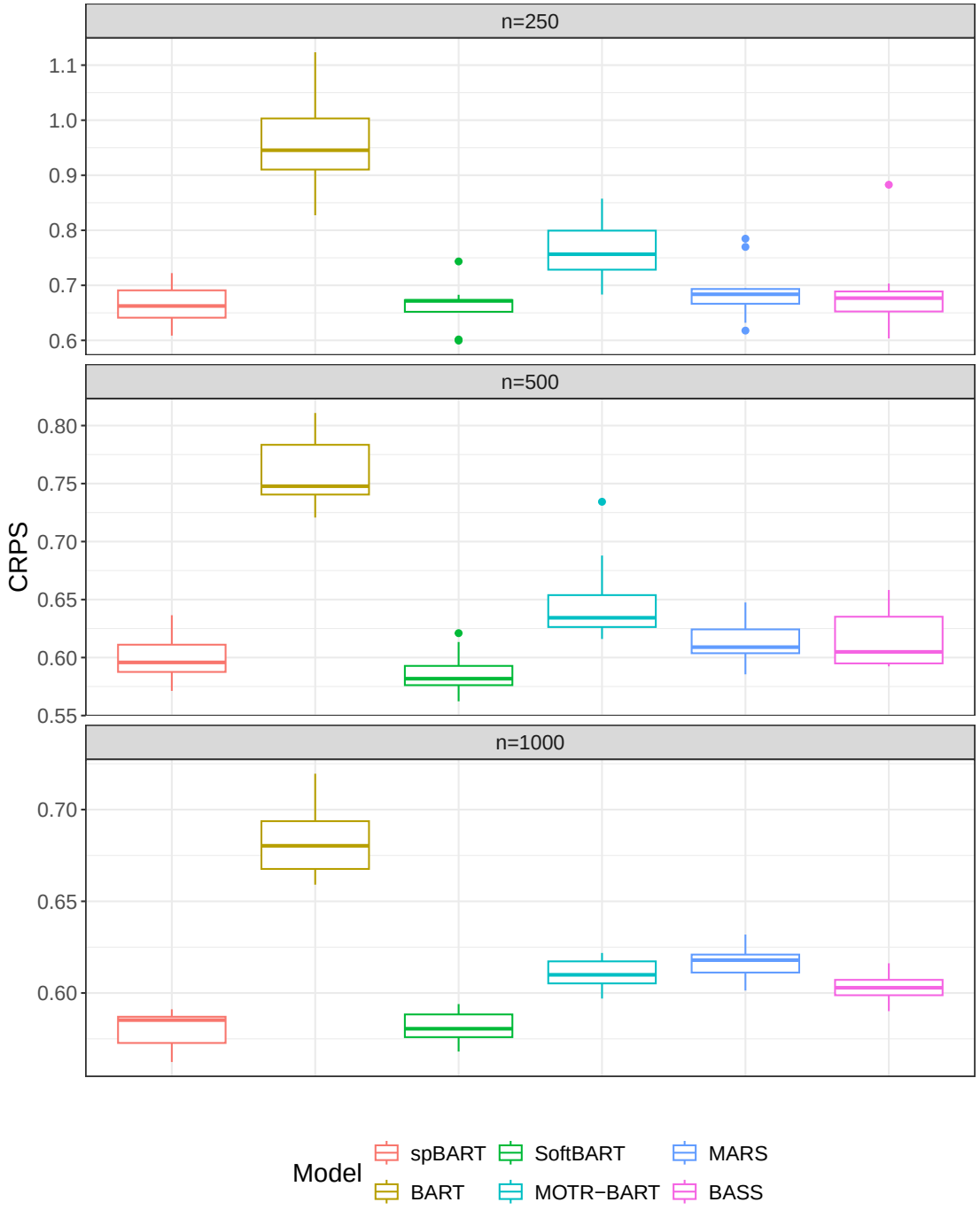


Figure 5.4.2: Comparisons between the CRPS calculated on the test samples by the competing models for the Friedman data using ten replications over different sample sizes $n_{\text{train}} = n_{\text{text}} = \{250, 500, 1000\}$.

The capacity of spBART to automatically identify the important set of basis functions can also be assessed through simulation. Figure 5.4.3 summarises the means of Δ_j and λ_j $j = 1, \ldots, d$ calculated over the posterior samples. We highlight, in blue, the set of variables on which $\mathbf{y}$ depends according to Equation (5.11). However, we note that the first two variables do not contribute directly; only a smooth interaction between $x_i^{(1)}$ and $x_i^{(2)}$ appears in Equation (5.11); thus, this is only interaction effect displayed. Conversely, the crosses denote the remaining uninformative main effects and interactions which do not contribute at all.

Initially, by calculating proportions Δ_j from Equation (5.2), we interpret that only bases with averages $\bar{\Delta}_j$ significantly exceeding zero are effectively selected by the model. We note that spBART removes $\mathbf{x}^{(1)}$ as it contributes solely through the interaction term. Subsequently, when analysing the $\bar{\lambda}_j$ averages, two points can be elucidated: the unselected j sets share approximately the same value, mirroring the prior settings for main effects and interactions. Furthermore, for effects with $\bar{\Delta}_j > 0$, the $\bar{\lambda}_j$ values can indicate the smoothness level of the marginal effect. For example, the lower value of $\bar{\lambda}_3$ among the main effect bases is expected, given the quadratic function that determines the marginal effect of $\mathbf{x}^{(3)}$ in Equation (5.11).

Another important aspect of λ_j is its ability to adapt sets of bases that are sporadically included in the trees. As previously mentioned, given the centralisation of the basis functions and their coefficients being centered at zero *a priori*, it is anticipated that less important bases will exhibit relatively larger λ_j values. Hence, when a group of bases, despite being selected for inclusion in the model by the MH step, exhibit no correlation with $\mathbf{y}$, the prediction surface is approximately constant around zero. This is grounded in the relationship detailed in Equation (5.9); i.e., the posterior rate for λ_j is proportional to the sum of the square of $\boldsymbol{\theta}_{t\ell}^{(j)}$ throughout the terminal nodes of the trees in which they are included. Therefore, the λ_j values associated with irrelevant bases tend to be larger than those for the other main effects. This can be seen by examining $\bar{\lambda}_2$ in Figure 5.4.3.

Finally, Figure 5.4.4 displays the marginal effects obtained from the spBART model which effectively contribute to $\mathbf{y}$ according to Equation (5.11). This demonstrates that the model produces predicted curves that closely match the true generating functions, ensuring the necessary smoothness in its estimates.

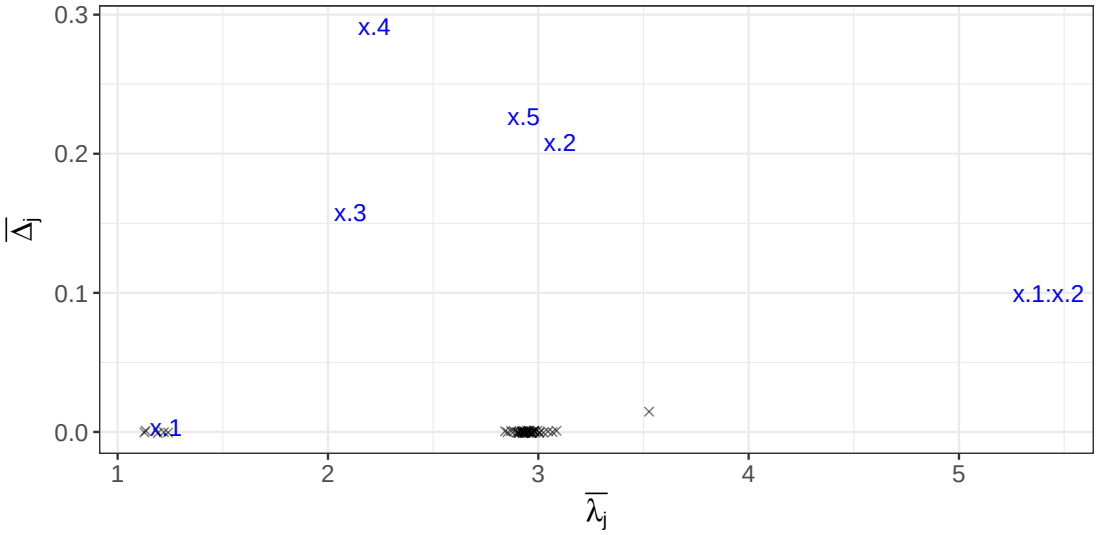


Figure 5.4.3: Posterior averages for $\bar{\Delta}_j$ and $\bar{\lambda}_j$ for one of the replications of spBART on Friedman data. The set of variables and the corresponding interactions that contribute to Equation (5.11) are highlighted. The remaining uninformative main effects and interactions which do not contribute are represented by crosses.

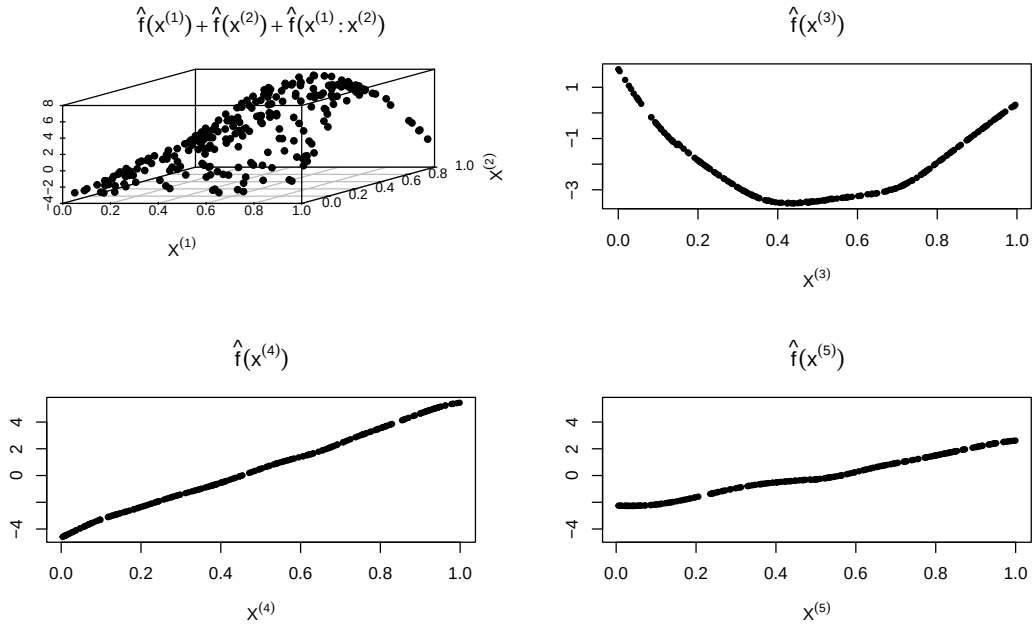


Figure 5.4.4: The panels depict the marginal effects derived from the set of basis functions most frequently selected by one of the replications of the spBART model applied to the Friedman dataset where $n_{\text{train}} = 250$.

5.4.2 Friedman break data

To evaluate the anticipated ability of spBART to detect change-points within the marginal effects of a predictor, we conducted a modified experiment based on the Friedman equation (5.11) by introducing a discontinuity in $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$. The data-generating process is described by

$$\begin{aligned} y_i =& 10 \sin\left(\pi x_i^{(1)} x_i^{(2)}\right) + \mathbb{1}_{(x_i^{(3)} \leq 0.5)} \left(20\left(x_i^{(3)} - 0.5\right)^2 + 5\right) \\ & - \mathbb{1}_{(x_i^{(3)} > 0.5)} \left(20\left(x_i^{(3)} - 0.5\right)^2 + 5\right) - \mathbb{1}_{(x_i^{(4)} \leq 0.3)} \left(15 x_i^{(4)} + 5\right) \\ & + \mathbb{1}_{(x_i^{(4)} > 0.3)} \left(10 x_i^{(4)} + 5\right) + 5 x_i^{(5)} + \epsilon_i, \quad i = 1, \ldots, n, \end{aligned} \tag{5.12}$$

where $\mathbb{1}(\cdot)$ is the usual indicator function, $x_i^{(j)} \sim \text{Unif}(0,1)\ \forall\, j = 1, \ldots, p$, and $\epsilon_i \sim \mathcal{N}(0, \tau^{-1})$. It is expected that spBART will be able to model the smooth curves, taking into account their inflections and discontinuities existent in the marginal effects. Furthermore, as in the previous example, it should effectively handle variable selection, detecting the most important set of basis functions, being interactions or not.

The outcomes are shown in Figures 5.4.5 and 5.4.6, presenting boxplots of the RMSE and CRPS values computed across the sample sets for varying sample sizes. From the results, we can observe that spBART consistently ranks among the methods producing the lowest values for both metrics. Unlike the previous Friedman simulation scenario, MARS appears to exhibit the weakest performance in this context, revealing a limitation in handling change-point scenarios. Conversely, we can notice that alternative models to BART demonstrate improved performance, suggesting that the lack of smoothness from the original formulation may still adversely affect estimations.

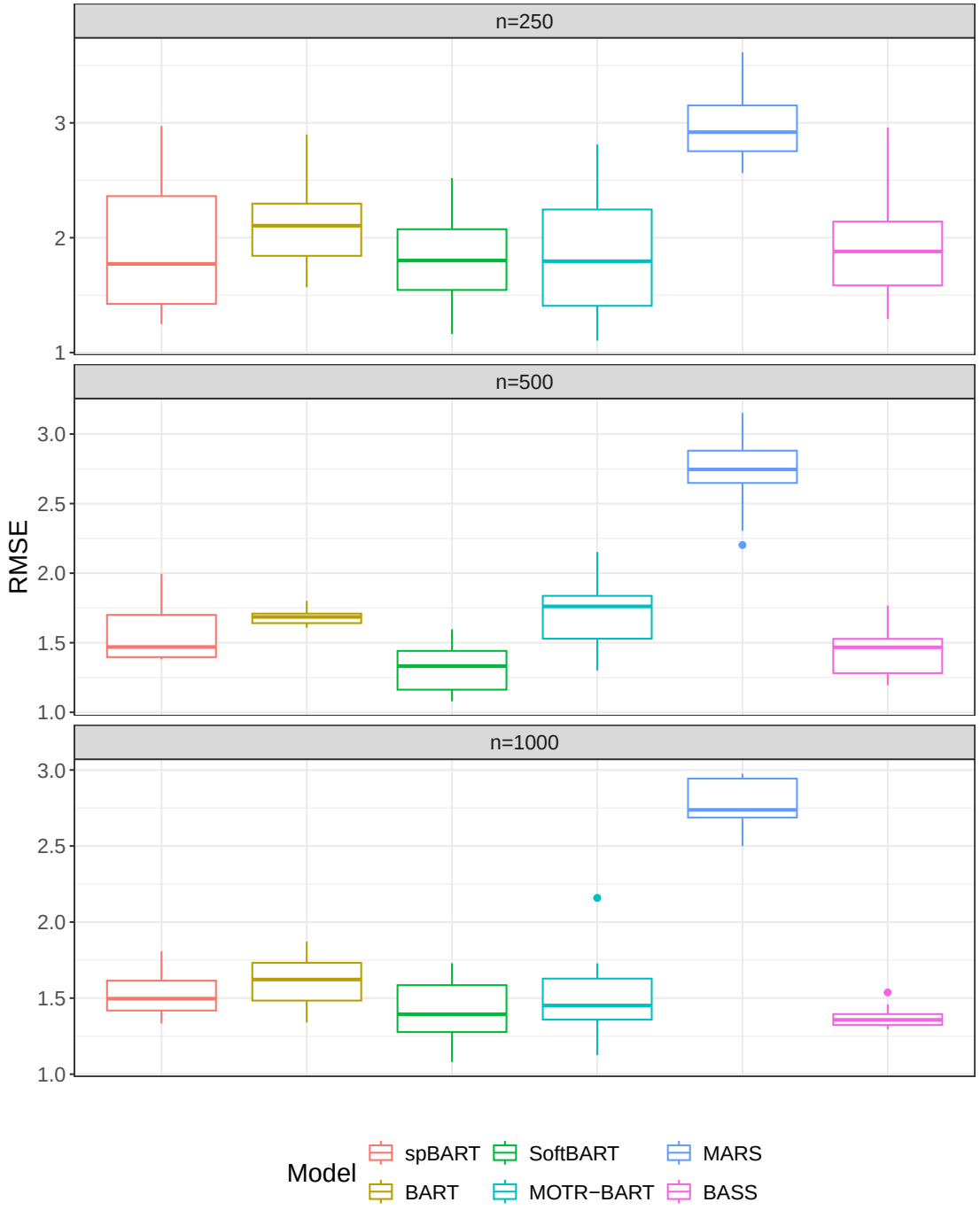


Figure 5.4.5: Comparisons between the RMSE calculated on the test samples by the competing models for the Friedman break data using ten replications over different sample sizes $n_{\text{train}} = n_{\text{text}} = \{250, 500, 1000\}$.

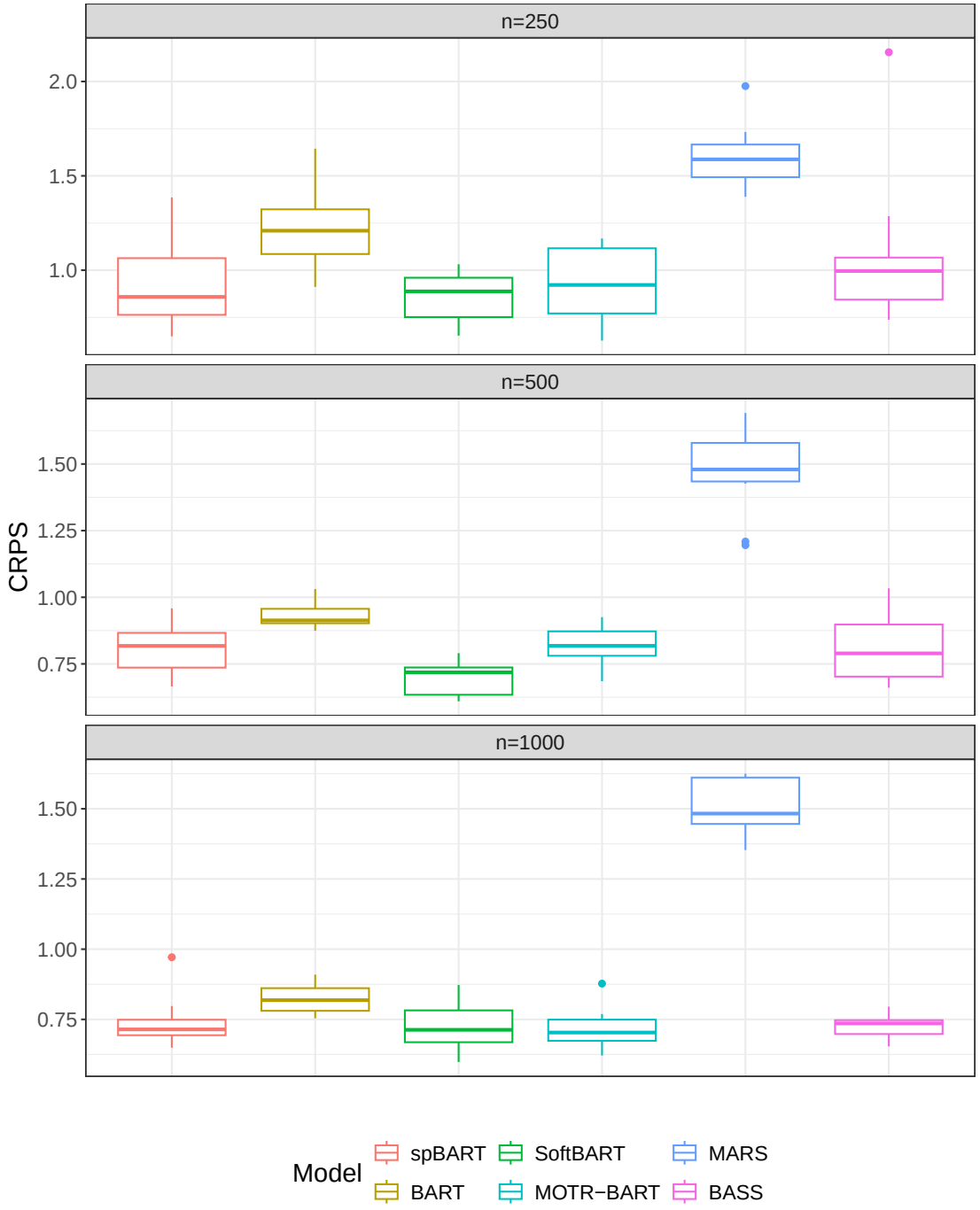


Figure 5.4.6: Comparisons between the CRPS calculated on the test samples by the competing models for the Friedman break data using ten replications over different sample sizes $n_{\text{train}} = n_{\text{text}} = \{250, 500, 1000\}$.

The result for the variable selection for the Friedman break example is illustrated by Figure 5.4.7. Analogously to the interpretation on the original Friedman data, we can observe that most of the sets of basis functions which are effectively contributing to the model have a higher proportion than those remaining. Their respective degrees of smoothness are also represented by their $\bar{\lambda}_j$ values where, as expected, among the main effects both $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$ are the ones with a smaller average. When a set of irrelevant basis functions is selected, λ_j can correct the influence of that marginal effect. This is exemplified by the unimportant $j = 2$, which has a moderate proportion $\bar{\Delta}_2$ and a large $\bar{\lambda}_2$ value.

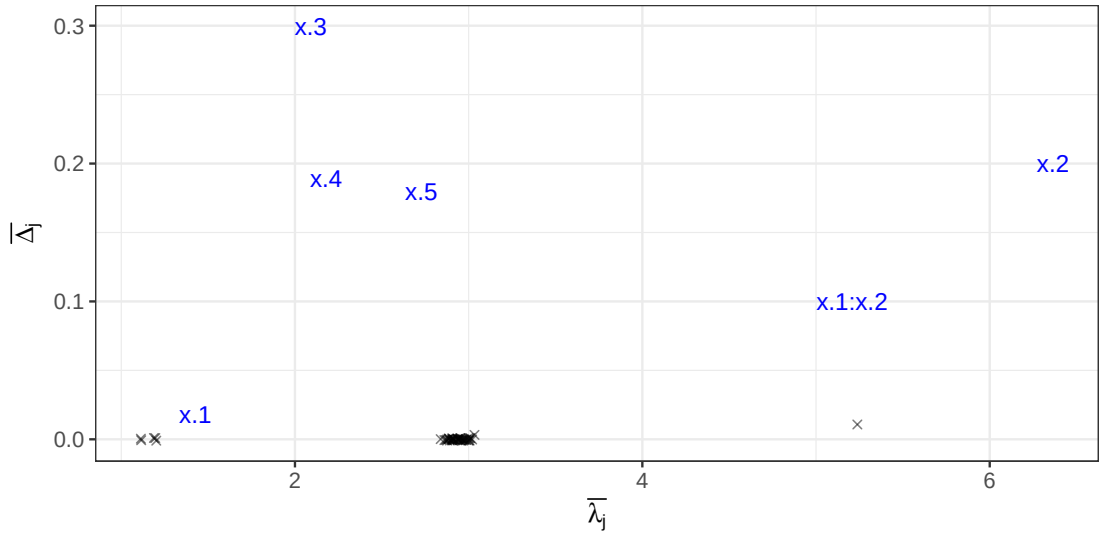


Figure 5.4.7: Posterior averages for $\bar{\Delta}_j$ and $\bar{\lambda}_j$ for one of the replications of spBART on Friedman break data. The set of variables and the corresponding interactions that contribute to (5.12) are highlighted. The remaining uninformative main effects and interactions which not contribute are represented by crosses.

Figure 5.4.8 shows the marginal effect of the main variables on the training data, derived from Equation (5.12), in one of the replications of spBART applied to the Friedman break data with $n_{\text{train}} = 250$. It is important to observe that the model successfully identified the change-points for both main effects and accurately approximated the generating functions within each partition. Furthermore, these panels underscore that spBART avoids unnecessary splits unless a change-point genuinely exists within the marginal effect, thereby facilitating insightful interpretations of the data.

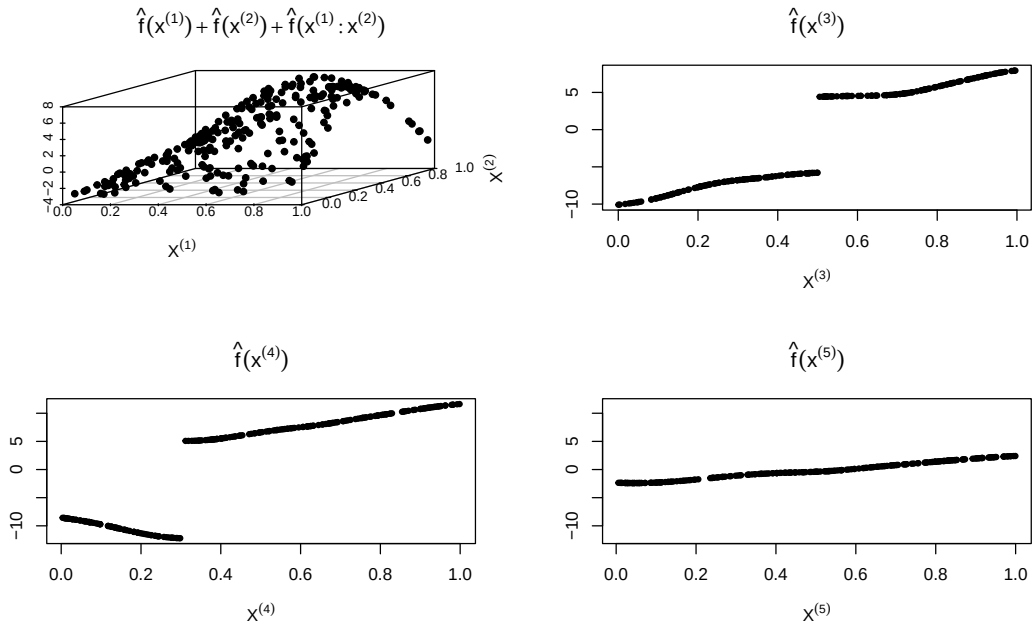


Figure 5.4.8: The panels depict the marginal effects derived from the set of basis functions most frequently selected by one of the replications of the spBART model applied to the Friedman break dataset where $n_{\text{train}} = 250$.

5.5 Real data benchmarking

We now evaluate spBART on real data. This dataset, previously used to demonstrate Bayesian MARS in Denison et al. (1998), comes from research by Bruntz et al. (1974) which investigates how ozone levels depend on specific meteorological factors over 153 days from May to September 1973 in the New York metropolitan area. The cube root of the ozone levels is the target variable, in accordance with Yu and Jones (1998). The dataset includes three predictors: solar radiation, temperature, and wind speed. The data is available in base `R` under the name `airquality`. Small proportions of the ozone and solar radiation variables are missing; in accordance with Denison et al. (1998), we present result using only the complete cases, resulting in a total of 111 observations.

To evaluate the model performance, 10-fold cross-validation was performed and the RMSE and CRPS were calculated over the test fold partitions. The spBART model was compared under its default setting, along with the same competitors

presented in Section 5.4. The results are summarised in Figure 5.5.1 which displays the RMSE and CRPS metrics resulting from the comparative analysis. spBART demonstrated the lowest median values for both indicators, reflecting higher predictive precision and improved uncertainty calibration compared to alternative models. SoftBART, emerging as the second-best performer, implies that adopting a smooth approximation may be advantageous for achieving more precise estimates of the cube root of ozone levels.

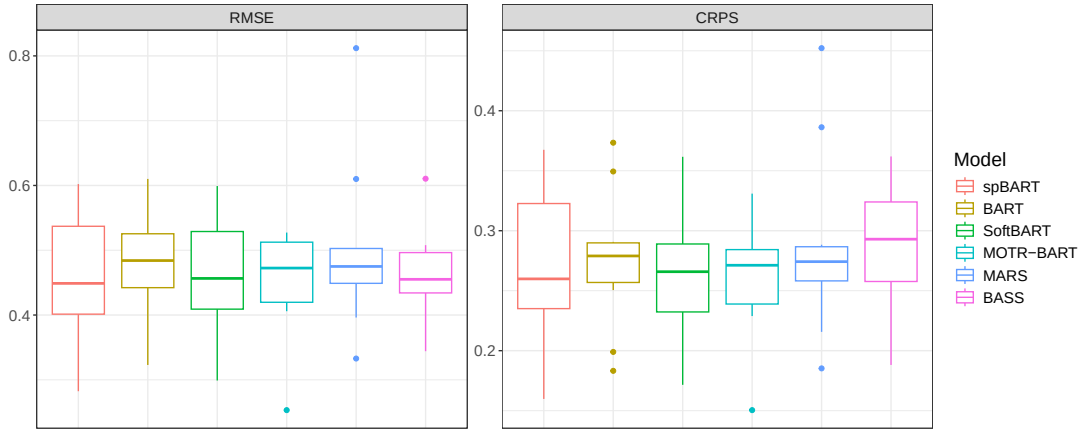


Figure 5.5.1: Comparison between the RMSE and CRPS values for the `airquality` data across the six competing methods using 10-fold cross-validation.

The analyses for variable importance were also performed and shown in a scatter plot of the posterior means $\bar{\Delta}_j$ and $\bar{\lambda}_j$ in Figure 5.5.2, using the whole dataset. The framework to interpret these results is the same as that of Figures 5.4.3 and 5.4.7. First we analyse the $\bar{\Delta}_j$ values and observe which set of bases have higher proportions. We see that *Temperature*, *Solar R.* and the interaction *Wind:Temperature* seem to be the main contributors to variability in the ozone levels. Examining their corresponding $\bar{\lambda}_j$ values, we infer that they differ from the prior expectation, thereby correctly adjusting the smoothness of the curve according to the data. On the other hand, the group of basis functions which are considered unimportant are the ones which present a low average proportion $\bar{\Delta}_j$, such as the predictor *Wind*, and the interactions *Solar.R:Temp*, and *Solar.R:Temp*; consequently, their respective $\bar{\lambda}_j$ values are close to the mean of their prior distribution.

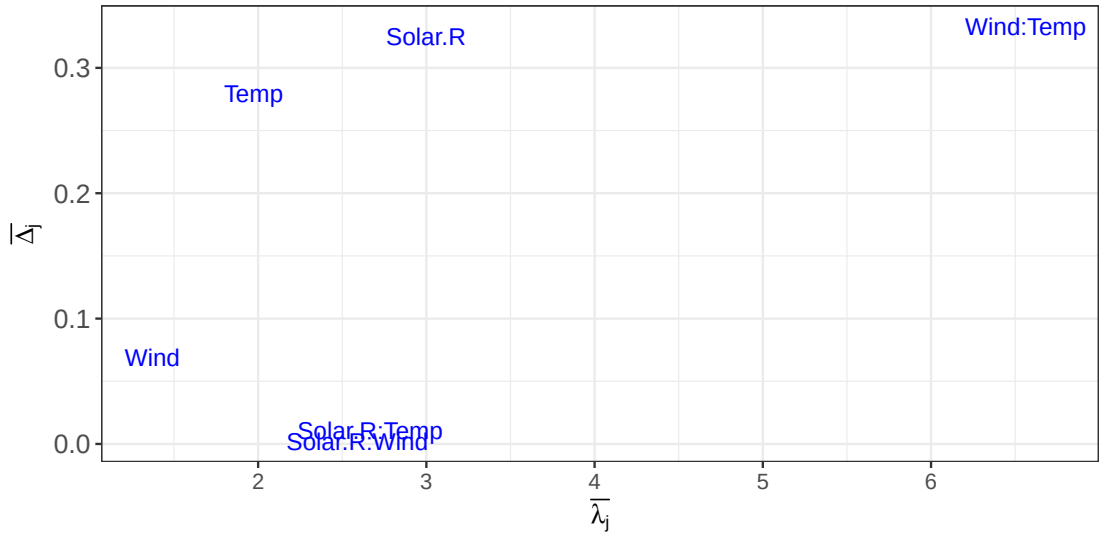


Figure 5.5.2: Posterior averages for $\bar{\Delta}_j$ and $\bar{\lambda}_j$ for the `airquality` data set, including all three main effects and all three pairwise interactions.

To substantiate the interpretations derived from Figure 5.5.2, the marginal effects for wind, temperature, solar radiation, and the *Wind:Temperature* interaction are explored in further detail. Figure 5.5.3 illustrates the three marginal main effect estimates, determined by the posterior median of each linear combination of their respective sets of basis functions and coefficients retrieved by the sampler. These results reinforce the initial conclusions, highlighting that temperature and solar radiation significantly contribute to the final prediction. Analysis of these two functions reveals their smooth characteristics, which may elucidate the relative underperformance of BART compared to other methods. As anticipated, the wind predictor exhibits a marginal effect consistently near zero and flat, which is in agreement with its small $\bar{\Delta}_j$ and $\bar{\lambda}_j$, respectively. The proportion $\bar{\Delta}_j$ is non-zero but nonetheless quite small; given that interaction bases are included only when they comprise of an interaction between the included main effect and some other variable, the frequency with which the unimportant wind variable is included could be attributable to the predominance of the *Wind:Temperature* interaction and not necessarily the importance of the main effect of wind. In other words, the frequent presence of this particular interaction modestly enhances the probability of including the main effect of wind in certain model samples compared to the other two unimportant effects, namely the two interactions involving solar radiation.

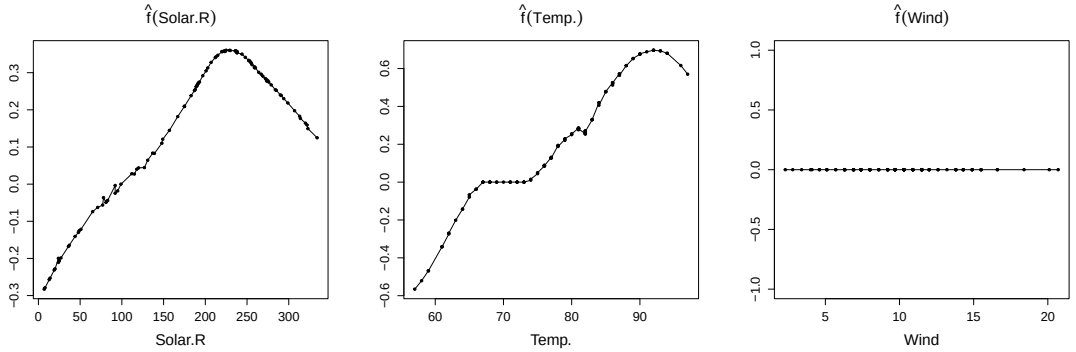


Figure 5.5.3: The marginal effect of the main effects calculated over the posterior median of predictions for the training observations across each set $j = 1, \ldots, 3$ of basis functions for the `airquality` data. This analysis reveals a non-linear relationship between the cubic root of solar radiation and temperature, while indicating an absence of the marginal effects of wind speed on ozone levels.

Figure 5.5.4 presents the estimated plane characterising the interaction between wind and temperature. In line with expectations deriving from the prominent proportion of this interaction, this term is contributing to explain the ozone levels through the presented non-linear smooth surface. In further detail, the surface exhibits a pronounced gradient with respect to changes in wind speed, particularly in areas of elevated temperatures. The intricate non-linear structure observed here underscores the benefits of employing basis functions to capture inherent flexibility and smoothness needed in the model. Notably, we omit the additional interaction surfaces — i.e., the *Solar.R:Wind* and *Solar.R:Temperature* interactions — because they were approximately flat surfaces with zero effect.

Lastly, the results regarding the variable selection herein are also aligned with the ones obtained by Denison et al. (1998). The findings from Bayesian MARS suggest that the most important basis functions in the fit were the main effect terms for radiation and temperature and the wind and temperature interaction term. This congruence reinforces the reliability of our proposed methodology, which is further complemented by the superior predictive performance and calibration obtained by spBART when compared with its competitors.

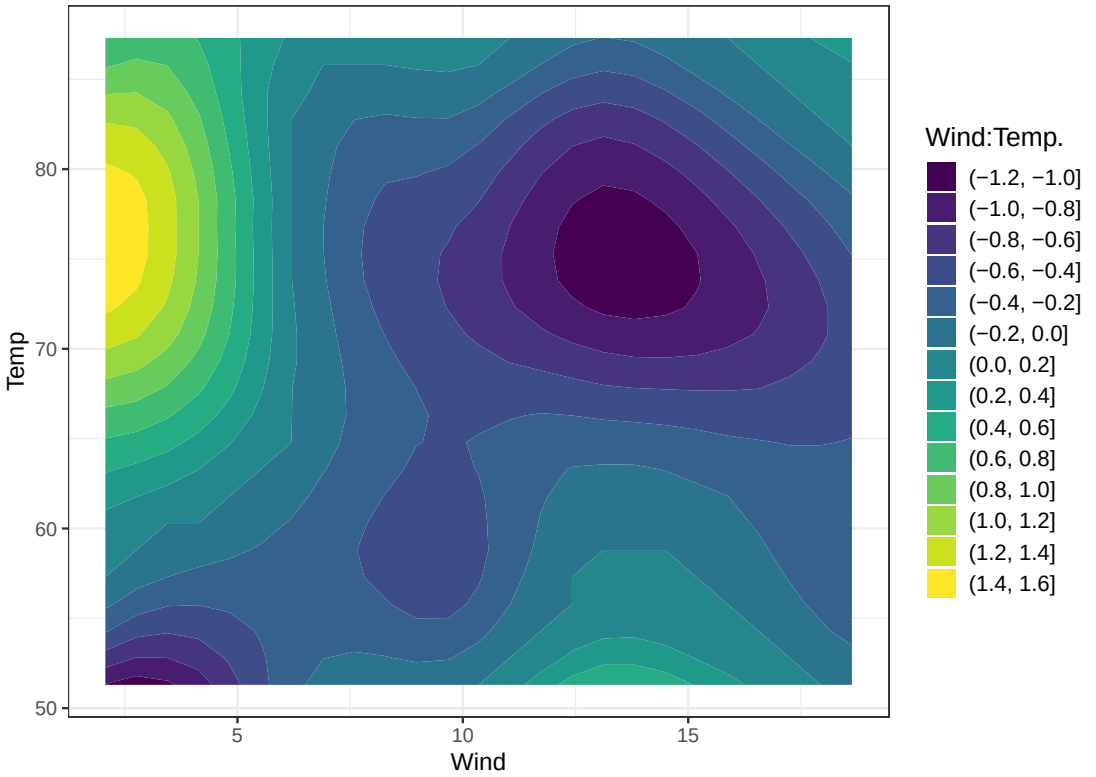


Figure 5.5.4: The marginal effect of the *Wind:Temperature* interaction, determined using the posterior median of predictions for the training observations. The surface exhibits a non-linearity and demonstrates the flexibility afforded by P-splines. Despite the relatively higher value of $\bar{\lambda}_j$, this parameter calibrates the appropriate degree of smoothness required to adjust the curve accurately, and is not merely a consequence of a constant marginal effect.

5.6 Discussion

In this work, we present spBART, an extension of BART models that integrates Bayesian P-splines into the terminal nodes, resulting in smoother predictions. Furthermore, spBART is envisioned as a tool for identifying important main effects and low-order interaction components, while adapting the degree of smoothness without the need for pre-specifying the functional form of an additive model.

The novel ensemble approach was evaluated over simulation scenarios with two primary objectives: to assess the predictive performance and uncertainty calibration of the method in comparison to its competitors, and to perform a selection

among the relevant subset of basis functions that should be included in the model. For the first task, spBART demonstrated consistent performance, out-performing most of its competitors, with the exception of SoftBART, against which it maintained competitive results. With respect to variable selection, spBART identified the important main components of the model, discarding the ones that do not contribute effectively to explaining the the variability in the response $\mathbf{y}$. Moreover, due to the additive nature of the model facilitated by splines, it was straightforward to visualise and provide interpretable marginal effects, thereby aiding in the development of more comprehensible models.

Ultimately, spBART was employed on a real dataset to estimate ozone levels based on several physical properties, including radiation level, wind speed, and temperature. The model identified a subset of significant effects and interactions, along with marginal effects plots that could offer valuable insights to domain specialists. These findings are corroborated by the outcomes of cross-validation experiments, which demonstrated that the model possesses better predictive capabilities and effective uncertainty quantification relative to closely-related alternative models.

Despite the simulations and experimentation described above, there are still some open possibilities to be explored that could improve the performance of the model:

- Setting the correct number of trees T seems to be a sensitive specification of the model. Despite the default value for total number of trees being $T = p$, such that there is initially one tree per main effect, we observed that the model performance can deteriorate when using a larger number of trees. The harm on the model performance is more pronounced regarding the variable selection, once it seems that if enough trees are dominating the modelling of the signal, the remaining trees will attempt to model noise leading to the inclusion of spurious correlation in final aggregated model. Chakraborty (2016) proposed an approach for automatic tree selection in BART which may could be adapted here at the cost of adding extra computational step, as it would be necessary to use a RJ-MCMC (Green, 1995) step as the number of trees would be changing.

- In certain scenarios, maintaining fixed probabilities for the add, remove, and modify operations, even after the model identifies the requisite subset of basis functions, can result in superfluous and inefficient MCMC iterations. Implementing techniques for diminishing adaptation, as discussed in Haario et al. (2001) and Roberts and Rosenthal (2009), for tree proposals could enhance efficiency and build a more computationally effective sampler for the trees.

- The number d of possible sets of basis functions can escalate rapidly with p and thus harm the exploration of the space of predictors. This challenge was partially addressed by limiting each tree to one main effect and its associated interactions. However, further enhancement is possible through the introduction of a Dirichlet prior on the probability of selecting set j in the add, remove, and modify operations. This strategy, which could adopt the prior configuration proposed by Linero (2018), has the potential to improve convergence and prevent the wasteful expenditure of MCMC iterations.

- The prior for the number of sets of basis functions $m_{t\ell}$ is specified as a zero-truncated Poisson distribution with a fixed rate parameter ψ_m. Assuming a hyperprior for ψ_m may allow greater flexibility for more complex models.

- Although the choice of the number of number of internal knots is facilitated by the penalised splines approach, it would still be interesting to further evaluate the sensitivity of the model for a wider range of choices for this parameter. Similarly, the effects of varying the degree of the B-splines used to generate $\mathbf{C}^{(j)}$ and the order r of the penalty applied to their differences could also be evaluated.

We plan to integrate these developments into our future research projects.

Appendix

5.A Marginal plots from the main effects from simulations

One of the main advantages from the additive functions used within the terminal nodes of spBART is the possibility to easily recover the marginal effects for each predictor $\mathbf{x}^{(j)}$ through their respective basis functions and associated parameters $z_{t\ell}^{(j)}$ and $\boldsymbol{\theta}_{t\ell}^{(j)}$. Figures 5.A.1 and 5.A.2 represent all the marginal effects for each main basis $j = 1, \ldots, p$ over the training sample, derived from a randomly selected replication within the cross-validation setting, where $n_{\text{train}} = n_{\text{test}} = 250$, for the Friedman dataset and Friedman break dataset, respectively. Though these plots include some main effects which were already depicted in Figures 5.A.1 and 5.A.2, most of the panels depict main effects which were omitted from the main section of the paper as they were deemed not to be effectively contributing to the model estimation since they are mostly flat at zero. For similar reasons, and the fact that they greatly outnumber the number of main effects, we elect not to show the remaining interaction effects either. Overall, both examples illustrate the successful variable selection performed by spBART.

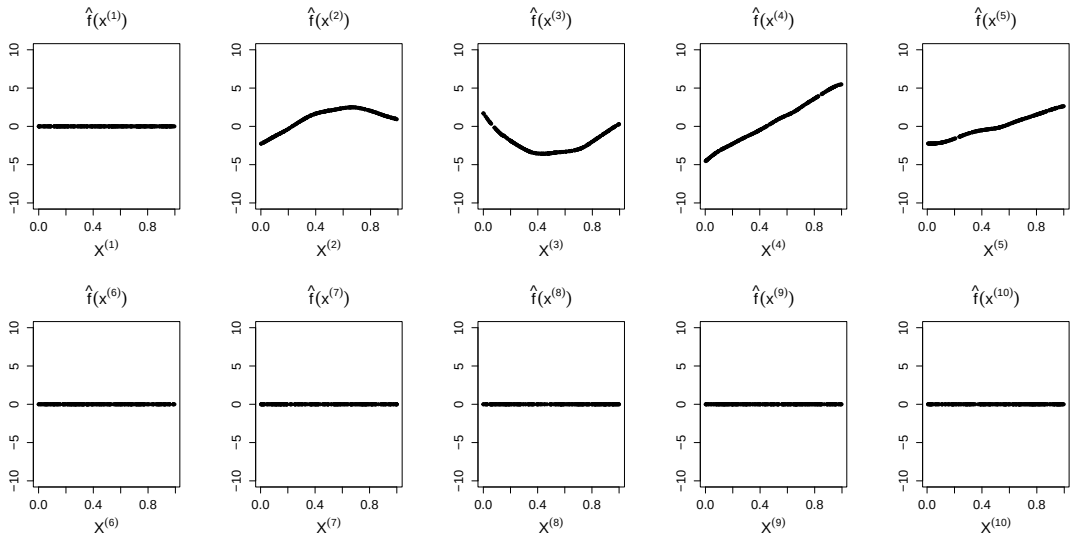


Figure 5.A.1: All marginal effect estimates for the main bases $j = 1, \ldots, p$, derived from a randomly selected replication within the cross-validation setting, where $n_{\text{train}} = n_{\text{test}} = 250$, on the training sample of Friedman data.

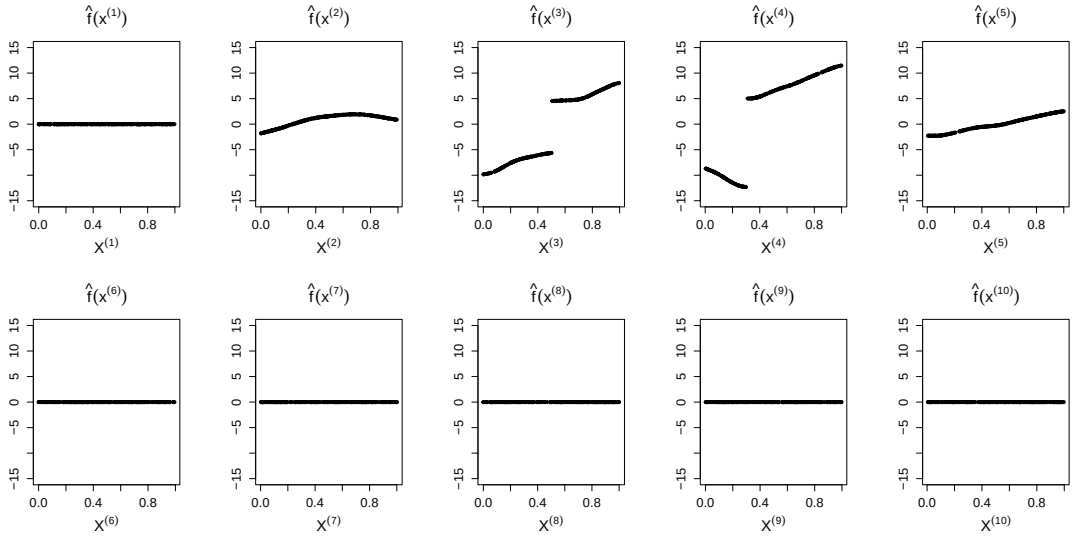


Figure 5.A.2: All marginal effect estimates for the main bases $j = 1, \ldots, p$, derived from a randomly selected replication within the cross-validation setting, where $n_{\text{train}} = n_{\text{test}} = 250$, on the training sample of Friedman break data.

CHAPTER 6

Conclusions

Bayesian additive regression trees (BART) is a non-parametric Bayesian approach which leverages an ensemble of Bayesian decision trees to approximate non-linear dynamics and capture low-order interactions for modelling univariate responses from a set of predictors with minimal assumptions and model specifications. In this thesis, we introduce a suite of extensions to address some shortcomings of BART, and to accommodate the breach of certain key assumptions, by incorporating a covariance structure within the feature space (GP-BART; Chapter 3), modelling the covariance structure of the response for multivariate outcomes (suBART; Chapter 4), and enhancing smoothness by embedding penalised splines within BART (spBART; Chapter 5). We now revisit the innovations presented in each chapter, underlining the novel contributions made, in light of their comprehensive evaluation and comparison against existing methodologies aimed at resolving similar challenges. We conclude with reflections on open research inquiries that merit further exploration to refine or broaden the enhancements discussed herein.

Initially, by incorporating Gaussian processes (GP) to account for possible correlations among observations, the model manages the covariance structure inherent in the multivariate normal prior over function spaces. Introducing GPs not only facilitates the modelling of dependencies but also grants the ability to create smooth surfaces. Furthermore, we propose new, oblique tree-splitting strategies, referred to as 'grow-rotate' and 'change-rotate', offering modifications over traditional parallel-

axis splits and thus increasing the flexibility of partitions from the model. We anticipate these innovations being particularly effective in (but not limited to) the domain of spatial data analysis. The effectiveness of GP-BART is demonstrated through extensive simulation studies and analyses of several real-world benchmark datasets, where it consistently outperforms various tree-based alternatives and spatial methods in terms of prediction accuracy and uncertainty calibration.

In developing the GP-BART approach, a significant computational challenge was encountered due to the complexity of $\mathcal{O}(n_{t\ell}^3)$ associated with the required matrix inversion operations within each terminal node. Attempts to alleviate this computational load through low-rank approximations were made, yet such strategies detrimentally impacted the model's predictive accuracy, indicating a need for a more nuanced treatment. Other strategies, including warm-up iterations aimed at reducing $n_{t\ell}$, did not yield successful outcomes as the resultant tree structures would take longer to converge. An exploration towards a more adaptable model involved varying the kernel function's length parameter within terminal nodes — that is, allowing for $\phi_{t\ell}$ rather than ϕ_{tj} as adopted throughout Chapter 3 — but it is more challenging to provide an efficient sampler for the length parameter if we consider it on the terminal node level.

Another path for investigation stems from the fact that GP-BART operates under the assumption of homoscedasticity, implying constant residual variance τ. However, this assumption often does not hold in either the GP literature or in real-world applications, presenting a challenge that has been addressed in several papers (e.g., Binois et al., 2018). Within the scope of this thesis, consideration was given to adopting the approach of Pratola et al. (2020), which incorporates multiplicative trees for modelling non-stationary variations within the BART framework. Aiming towards a more versatile ensemble of GPs in this direction, however, was deferred to future work as it was beyond the initial publication scope for GP-BART.

Moving towards the covariance structure on the outcomes, BART has the limitation that it only considers univariate responses. Consequently, in cases where it is of interest to analyse data with multiple correlated responses, it is necessary to modify the model in order to account for the dependencies between them.

Motivated by a problem from cost-effective analysis, we were interested in jointly modelling the average treatment effect with respect to two outcomes associated with healthcare cost and healthcare quality jointly. This extension facilitates the construction of a forest in which each distinct response is represented by its own ensemble of trees, while also accounting for the joint distribution of these outcomes by linking the trees through a correlated error term. In addition to the suBART model for multiple correlated continuous outcomes, we also present a probit extension of suBART tailored for multivariate binary outcomes. The evaluation of both suBART models encompasses several simulation settings. These assessments precede the application of the continuous suBART model to data derived from an observational study situated within the field of health economics.

Despite the comprehensive set of comparisons and scenarios, coupled with the nuanced insights provided by suBART in cost-effectiveness analysis, there still exists significant scope for further investigation into diverse extensions of the suBART model. One of the avenues to explore is to incorporate smoothness within the tree-based framework using probabilistic splitting rules, in a similar way to the SoftBART approach of Linero and Yang (2018). Another limitation which could be addressed is accounting for settings where the assumption of homoscedasticity is violated. This remains a prevalent issue within the conventional seemingly unrelated regression framework, as reported by Afolayan and Adeleke (2018). Given suBART's role as a robust alternative to the linear SUR methodology, the heteroscedasticity-tailored modification of BART, as proposed by Pratola et al. (2020), offers promising opportunities for adaptation in suBART applications, especially in situations where the assumption of uniform variance proves to be untenable. However, it is important to note that Pratola et al. (2020) only proposes a solution for estimating a scalar variance; adapting this to covariance structures in the presence of multiple outcomes would require additional specifications to the model beyond suBART's original scope. Furthermore, while our methodology introduces a general framework for multidimensional outcomes, the two presented versions of suBART are tailored distinctly for exclusively continuous and exclusively binary multivariate outputs. Future work aims at expanding suBART to enable the joint modelling of outcomes of mixed type, integrating both

binary and continuous responses. This advancement is particularly pertinent to data encountered in CEA settings, where such mixed outcomes are frequently observed, and given the current framework of suBART this could be easily addressed with extra adjustments for the covariance sampler.

In Chapter 5, we addressed one more way of tackling the lack-of-smoothness limitation that was first explored in the GP-BART approach. Within the domain of multiple regression, achieving smoothness can be alternatively pursued by incorporating additive functions (Friedman and Silverman, 1989). These function classes, due to their ability to reparameterise the original feature space and effectively capture non-linear relationships, are recognised as smoothers. Prior works, such as that of Prado et al. (2021), have shown that incorporating model trees, with linear regressions in the terminal nodes within the BART framework, can alleviate the inflexibility inherent in BART's piecewise-constant construction. Taking these considerations into account, we proposed the integration of penalised splines into BART, aiming to augment the versatility of the tree-based BART approach. The resulting model, termed spBART, enhances the capacity of BART to adapt to smooth predictive surfaces. Additionally, spBART can be viewed as facilitating model specification within the penalised-splines paradigm. This is because, as we demonstrate, the proposed sampling strategy is capable of automatically selecting the relevant sets of basis functions for the model, including those related to two-way interactions, thereby circumventing the need for pre-specifying the model structure. Similar to GP-BART, spBART's capabilities are showcased through comprehensive simulation studies, encompassing scenarios where the main effects exhibit smoothness and/or discontinuities. Furthermore, we demonstrate through an application that spBART outperforms or performs competitively against other established tree-based methods and spline approaches.

For spBART, there are also directions for further research to enhance the performance of the model. While the default number of trees (one per main effect) appears suitable, using an excessive number can deteriorate model performance, particularly regarding variable selection. This occurs when a subset of trees modelling the signal dominate, leaving remaining trees to capture noise and introduce spurious correlations. The automatic tree selection approach introduced by

Chakraborty (2016) for BART could potentially be adapted here, albeit at an increased computational cost due to the inclusion of reversible-jump MCMC steps required to account for varying tree numbers. Additionally, employing adaptive MCMC techniques — as explored in Haario et al. (2001) and Roberts and Rosenthal (2009) — could increase the efficiency of the sampler. Dynamically adjusting the probabilities of proposals — particularly those which modify the spline models at the terminal node level — could improve efficiency by reducing unnecessary MCMC iterations after the model identifies the relevant sets of basis functions. Furthermore, the rapid growth in the number of possible basis function sets with increasing predictors hinders exploration of the predictor space. While this is partially addressed by limiting each tree to include only the basis functions for one main effect at most and only the two-way interactions comprising this same variable, introducing a Dirichlet prior on the probability of selecting a specific set of basis functions as part of the novel 'add', 'remove', and 'modify' operations could further enhance convergence and reduce the number of wasteful iterations. This approach could potentially leverage the prior configuration proposed by Linero (2018). Finally, while the penalised splines approach relies on a choice regarding the number of internal knots, further investigation is warranted to evaluate the sensitivity of the model to a wider range of values for this parameter. Similar evaluations should be conducted for both the degree of the B-splines used to generate the basis functions and the order of the penalty applied to their differences.

Finally, there are other potential extensions that could be developed by borrowing features of the three main proposals included herein. Firstly, the rotated splitting rules which are presently exclusive to GP-BART could be incorporated into suBART and psBART — given that this innovation is not strictly limited to the spatial setting — which may further improve their performance. Secondly, the incorporation of smoothness achieved by GP-BART and spBART could be leveraged in a more flexible version of suBART; embedding GPs or splines in the terminal nodes of the trees in suBART could further enhance the suBART model and further differentiate our method from the purely parametric linear seemingly unrelated regression approach that is predominant in the cost-effectiveness literature.

Each proposed extension in this thesis has a corresponding software implementation available using the open-source statistical programming language R (R Core Team, 2024). All are distributed on GitHub repositories. These include `gpbart` (`github.com/MateusMaiaDS/gpbart`) for the GP-BART model proposed in Chapter 3, `suBART` (`github.com/MateusMaiaDS/suBART`) for the seemingly unrelated BART model proposed in Chapter 4, and `spBART` (`github.com/MateusMaiaDS/spBART`) for the penalised splines BART model proposed in Chapter 5. While the current versions are intended solely for replicating the results in this thesis, future work should focus on creating comprehensive R packages. These packages would enhance efficiency and potentially leverage C++ for faster computations across all models (not only GP-BART). Ultimately, the goal is to encourage broader user adoption of the extensions presented here.

Overall, this thesis introduces extensions to the BART ensemble method that effectively address key limitations of the standard approach. The proposed modifications demonstrate promising results in various applications, suggesting their broad applicability. As highlighted in the concluding chapter, there are many promising areas for future research. These areas hold the potential to further improve their capabilities and ultimately lead to their inclusion in formalised R packages.

Bibliography

Afolayan, R. B. and Adeleke, B. L. (2018). On the efficiency of some estimators for modeling seemingly unrelated regression with heteroscedastic disturbances. *IOSR Journal of Mathematics*, 14(4):1–13. 117, 167

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679. 21

Alvarez, I., Niemi, J., and Simpson, M. (2014). Bayesian inference for a covariance matrix. In *The 26th Annual Conference on Applied Statistics in Agriculture*, pages 71–82, Kansas State University. 91

Andugula, P., Durbha, S. S., Lokhande, A., and Suradhaniwar, S. (2017). Gaussian process based spatial modeling of soil moisture for dense soil moisture sensing network. In *6th International Conference on Agro-Geoinformatics*, pages 1–5. IEEE. 26

Baio, G. (2012). *Bayesian Methods in Health Economics*. Chapman & Hall/CRC Biostatistics Series, Boca Ration, FL, U.S.A. 79

Baldi, P. (2024). *Probability: An Introduction Through Theory and Exercises*. Universitext. Springer Nature. 95

Balog, M., Lakshminarayanan, B., Ghahramani, Z., Roy, D. M., and Teh, Y. W. (2016). The Mondrian kernel. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 32–41, Arlington, VA, U.S.A. AUAI Press. 25

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848. 26

Barnard, J., McCulloch, R. E., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311. 94

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(2):65–75. 2

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746. 133

Binois, M., Gramacy, R. B., and Ludkovski, M. (2018). Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821. 166

Bivand, R. and Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3):716–748. 51

Blaser, R. and Fryzlewicz, P. (2016). Random rotation ensembles. *The Journal of Machine Learning Research*, 17(1):126–151. 26

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140. 2, 101

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, 26(3):801–824. 2

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. 2, 22

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC Press, New York, NY, U.S.A. 2, 12, 108

Bruntz, S. M. L., Cleveland, W. S., Kleiner, B., and Warner, J. L. (1974). The dependence of ambient ozone on solar radiation, wind, temperature, and mixing height. In *Symposium on Atmospheric Diffusion and Air Pollution*, pages 125–128, Boston, MA, U.S.A. 156

Buja, A., Hastie, T. J., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510. 130

Cao, T., Lu, L., and Jiang, T. (2023). Robust regression in environmental modeling based on Bayesian additive regression trees. *Environmental Modeling & Assessment*, pages 1–13. 3, 129

Chakraborty, S. (2016). Bayesian additive regression tree for seemingly unrelated regression with automatic tree selection. In *Handbook of Statistics*, volume 35, pages 229–251. Elsevier. 81, 161, 169

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361. 59, 93, 94

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948. 2, 9, 12, 13, 15, 24, 28, 29, 80, 86, 87, 131, 136, 137, 142

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298. 3, 5, 17, 18, 23, 28, 30, 32, 33, 34, 40, 47, 80, 81, 86, 87, 88, 89, 90, 94, 129, 136, 141, 143, 146

Clark, T. E., Huber, F., Koop, G., Marcellino, M., and Pfarrhofer, M. (2023). Tail forecasting with multivariate Bayesian additive regression trees. *International Economic Review*, 64(3):979–1022. 81

Clayton, D. G. (1995). Generalized linear mixed models. In Gilks, W. R., Richardson, S. T., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, volume 1, pages 275–302. Chapman and Hall/CRC Press, New York, NY, U.S.A. 133

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297. 22

Cressie, N. (2015). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, New Jersey, NJ, U.S.A., revised edition. 41

Dakin, H., Gray, A., Fitzpatrick, R., MacLennan, G., Murray, D., KAT Trial Group, et al. (2012). Rationing of total knee replacement: a cost-effectiveness analysis on a large trial data set. *BMJ Open*, 2(1):e000332. 116

De Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1):50–62. 130

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Bayesian MARS. *Statistics and Computing*, 8:337–346. 2, 9, 12, 13, 14, 15, 20, 131, 147, 156, 159

Dorie, V., Chipman, H. A., and McCulloch, R. E. (2024). *dbarts: discrete Bayesian additive regression trees sampler*. R package version 0.9-26. https://CRAN.R-project.org/package=dbarts. 4, 20, 100, 147

Dorie, V., Hill, J. L., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68. 3, 79, 80, 82, 109

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., and Torrance, G. W. (2015). *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, Oxford, UK, 4th edition. 111, 118

Durbán, M. and Currie, I. D. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics*, 18:251–262. 133

Eilers, P. H. C. (1999). Discussion on: the analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48:307–308. 132

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121. 4, 130, 132

Eilers, P. H. C., Marx, B. D., and Durbán, M. (2015). Twenty years of P-splines. *SORT: Statistics and Operations Research Transactions*, 39(2):149–186. 137

El Alili, M., van Dongen, J. M., Esser, J. L., Heymans, M. W., van Tulder, M. W., and Bosmans, J. E. (2022). A scoping review of statistical methods for trial-based economic evaluations: the current state of play. *Health Economics*, 31(12):2680–2699. 79

Francom, D. and Sansó, B. (2020). BASS: an R package for fitting and performing sensitivity analysis of Bayesian adaptive spline surfaces. *Journal of Statistical Software*, 94(8):1–36. 147

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67. 22, 47, 76, 101, 131, 146, 147

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232. 2, 3, 17, 22, 24, 80

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21. 4, 8, 130, 168

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823. 130

Gabrio, A., Baio, G., and Manca, A. (2019). Bayesian statistical economic evaluation methods for health technology assessment. In Hamilton, J., editor, *Economic Theory and Mathematical Models*. Oxford Research Encyclopedia of Economics and Finance, Oxford, UK. 82, 84, 111

García-Pedrajas, N., García-Osorio, C., and Fyfe, C. (2007). Nonlinear boosting projections for ensemble construction. *The Journal of Machine Learning Research*, 8(1):1–33. 26

Gelfand, A. E. and Schliep, E. M. (2016). Spatial statistics and Gaussian processes: a beautiful marriage. *Spatial Statistics*, 18:86–104. 26

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889. 133

Gilley, O. W. and Pace, R. K. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31(3):403–405. 51

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 41, 99, 146

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130. 25, 26, 41

Gramacy, R. B. and Taddy, M. (2010). Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed Gaussian process models. *Journal of Statistical Software*, 33(6):1–48. 41

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. 2, 15, 25, 131, 161

Gu, C. (2013). *Smoothing Spline ANOVA Models*, volume 297 of *Springer Series in Statistics*. Springer-Verlag, New York, NY, U.S.A., 2nd edition. 130

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242. 162, 169

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056. 3, 81, 84, 110, 117

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102. 51

Hastie, T. J. and Tibshirani, R. (2000). Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science*, 15(3):196–223. 3, 19, 35, 88, 95, 141

He, J. and Hahn, P. R. (2023). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association*, 118(541):551–570. 4, 57

Hernán, M. A. and Robins, J. M. (2024). *Causal Inference: What If*. Chapman & Hall/CRC Press, Boca Raton, FL, U.S.A. 82

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240. 82

Hill, J. L., Linero, A. R., and Murray, J. (2020). Bayesian additive regression trees: a review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278. 3, 24, 129

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844. 2

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452. 91, 92, 96, 100

Huber, F. and Rossini, L. (2022). Inference in Bayesian additive vector autoregressive tree models. *The Annals of Applied Statistics*, 16(1):104–123. 80

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, U.S.A. 83

Inglis, A., Parnell, A. C., and Hurley, C. B. (2024). Visualisations for Bayesian additive regression trees. *Journal of Data Science, Statistics, and Visualisation*, 4(1). 3

Janizadeh, S., Vafakhah, M., Kapelan, Z., and Dinan, N. M. (2021). Novel Bayesian additive regression tree methodology for flood susceptibility modeling. *Water Resources Management*, 35(13):4621–4646. 24, 80

Kapelner, A. and Bleich, J. (2016). bartMachine: machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40. 4, 20, 29, 30, 88, 97, 119, 137, 145, 146

Kim, C. (2022). Bayesian additive regression trees in spatial data analysis with sparse observations. *Journal of Statistical Computation and Simulation*, 92(15):3275–3300. 3, 129

Kindo, B. P., Wang, H., and Peña, E. A. (2016). Multinomial probit Bayesian additive regression trees. *Stat*, 5(1):119–131. 118

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328. 130

Lamers, L. M., McDonnell, J., Stalmeier, P. F. M., Krabbe, P. F. M., and Busschbach, J. J. V. (2006). The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health Economics*, 15(10):1121–1132. 107

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212. 130, 132, 133

Li, F., Ding, P., and Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153. 82, 83, 110

Li, J., Potter, A., Huang, Z., Daniell, J. J., and Heap, A. D. (2011). Predicting seabed mud content across the Australian margin: comparison of statistical and mathematical techniques using a simulation experiment. Technical report, Geoscience Australia, Canberra, Australia. Record 2010/11. 51

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(1):1–25. 41

Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559. 3, 22, 25

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636. 3, 162, 169

Linero, A. R. (2022a). Generalized Bayesian additive regression trees models: beyond conditional conjugacy. *arXiv preprint: arXiv:2202.09924*. 3

Linero, A. R. (2022b). SoftBart: soft Bayesian additive regression trees. *arXiv preprint: arXiv:2210.16375*. 41, 147

Linero, A. R. and Antonelli, J. L. (2023). The how and why of Bayesian nonparametric causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1583. 3

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110. 3, 24, 37, 40, 47, 80, 109, 116, 129, 144, 147, 167

Löthgren, M. and Zethraeus, N. (2000). Definition, interpretation and calculation of cost-effectiveness acceptability curves. *Health Economics*, 9(7):623–630. 84, 113

Low-Kam, C., Telesca, D., Ji, Z., Zhang, H., Xia, T., Zink, J. I., and Nel, A. E. (2015). A Bayesian regression tree approach to identify the effect of nanoparticles' properties on toxicity profiles. *The Annals of Applied Statistics*, 9(1):383–401. 130

Maia, M., Murphy, K., and Parnell, A. C. (2024). GP-BART: a novel Bayesian additive regression trees approach using Gaussian processes. *Computational Statistics & Data Analysis*, 190:107858. 116, 129

McJames, N., Parnell, A. C., Goh, Y. C., and O'Shea, A. (2023). Bayesian causal forests for multivariate outcomes: application to Irish data from an international large scale education assessment. *arXiv preprint: arXiv:2303.04874*. 5, 81, 89, 118

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer. 26

Mersmann, O. (2021). *microbenchmark: accurate timing functions*. R package version 1.4.9. https://CRAN.R-project.org/package=microbenchmark. 50

Milborrow, S. (2024). *earth: multivariate adaptive regression splines*. R package version 5.3.3. https://CRAN.R-project.org/package=earth. 147

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434. 2

Müller, P., Shih, Y.-C. T., and Zhang, S. (2007). A spatially-adjusted Bayesian additive regression tree model to merge two datasets. *Bayesian Analysis*, 2(3):611–633. 3

Murthy, S. K. and Salzberg, S. (1995). Decision tree induction: how effective is the greedy heuristic? In *KDD*, KDD'95, pages 222–227. AAAI Press. 2

Nychka, D., Furrer, R., Paige, J., and Sain, S. (2021). *fields: tools for spatial data*. University Corporation for Atmospheric Research, Boulder, CO, U.S.A. R package version 14.1. https://github.com/dnychka/fieldsRPackage. 41

Papageorgiou, G., Richardson, S., and Best, N. (2014). Bayesian non-parametric models for spatially indexed data of mixed type. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(5):973–999. 117

Peruzzi, M. and Dunson, D. B. (2022). Spatial multivariate trees for big data Bayesian regression. *The Journal of Machine Learning Research*, 23(1):747–786. 5, 80

Pourmohamad, T. and Lee, H. K. H. (2016). Multivariate stochastic process models for correlated responses of mixed type. *Bayesian Analysis*, 11(3):797–820. 117

Prado, E. B., Moral, R. A., and Parnell, A. C. (2021). Bayesian additive regression trees with model trees. *Statistics and Computing*, 31(3):1–13. 3, 8, 24, 116, 129, 147, 168

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417. 3, 117, 166, 167

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106. 2

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, U.S.A. 2

Quiñonero-Candela, J., Rasmussen, C. E., and Williams, C. K. I. (2007). Approximation methods for Gaussian process regression. In *Large-scale Kernel Machines*, pages 203–223. MIT Press. 57

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, volume 20 of *NIPS'07*, pages 1177–1184, Red Hook, NY, USA. Curran Associates Inc. 57

Ramful, P. and Zhao, X. (2009). Participation in marijuana, cocaine and heroin consumption in Australia: a multivariate probit approach. *Applied Economics*, 41(4):481–496. 81

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 4, 170

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183. 130

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125. 98

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367. 162, 169

Rocková, V. (2020). On semi-parametric inference for BART. In *International Conference on Machine Learning*, pages 8137–8146. PMLR. 80

Ročková, V. and Saha, E. (2019). On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848. PMLR. 3, 80

Ročková, V. and Van der Pas, S. (2020). Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131. 3, 80, 109

Rudolph, K. E., Williams, N. T., Miles, C. H., Antonelli, J., and Diaz, I. (2023). All models are wrong, but which are useful? Comparing parametric and nonparametric estimation of causal effects in finite samples. *Journal of Causal Inference*, 11(1):20230022. 79, 80, 82

Sarti, D. A., Prado, E. B., Inglis, A. N., Dos Santos, A. A. L., Hurley, C. B., Moral, R. A., and Parnell, A. C. (2023). Bayesian additive regression trees for genotype by environment interaction models. *The Annals of Applied Statistics*, 17(3):1936–1957. 80

Smith, P. L. (1979). Splines as a useful and convenient statistical tool. *The American Statistician*, 33(2):57–62. 130

Sparapani, R., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2020). Nonparametric competing risks analysis using Bayesian additive regression trees. *Statistical Methods in Medical Research*, 29(1):57–77. PMID: 30612519. 3, 129

Sparapani, R., Spanbauer, C., and McCulloch, R. E. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package. *Journal of Statistical Software*, 97(1):1–66. 4, 20, 41

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753. 3

Stan Development Team (2024a). RStan: the R interface to Stan. R package version 2.32.5. https://mc-stan.org/. 100

Stan Development Team (2024b). *Stan User's Guide*. Stan Development Team. https://mc-stan.org/docs/. 100, 109

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., Scott, J. G., et al. (2020). BART with targeted smoothing: an analysis of patient-specific stillbirth risk. *The Annals of Applied Statistics*, 14(1):28–50. 24

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240. 6

Um, S., Linero, A. R., Sinha, D., and Bandyopadhyay, D. (2023). Bayesian additive regression trees for multivariate skewed responses. *Statistics in Medicine*, 42(3):246–263. 5, 81, 89, 100, 116, 118

Vink, G., Frank, L. E., Pannekoek, J., and Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90. 109

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900. 92

Wang, M., He, J., and Hahn, P. R. (2023). Local Gaussian process extrapolation for BART models with applications to causal inference. *Journal of Computational and Graphical Statistics*, 0(0):1–12. advance online publication: https://doi.org/10.1080/10618600.2023.2240384. 25

Wegman, E. J. and Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association*, 78(382):351–365. 130

Wiertsema, S. H., Van Dongen, J. M., Geleijn, E., Huijsmans, R. J., Bloemers, F. W., De Groot, V., and Ostelo, R. W. J. G. (2019). Cost-effectiveness of the transmural trauma care model (TTCM) for the rehabilitation of trauma patients. *International Journal of Technology Assessment in Health Care*, 35(4):307–316. xvii, 79, 82, 107, 108, 109, 111, 119

Willan, A. R., Briggs, A. H., and Hoch, J. S. (2004). Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics*, 13(5):461–475. 6, 79, 84

Williams, C. K. I., Rasmussen, C. E., Scwaighofer, A., and Tresp, V. (2002). Observations on the Nyström method for Gaussian process prediction. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany. 57

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. (2020). Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning*, ICML '20, pages 10292–10302. PMLR. 57

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition. 133

Wright, M. N. and König, I. R. (2019). Splitting on categorical predictors in random forests. *PeerJ*, 7:e6339. 30

Wright, M. N. and Ziegler, A. (2017). ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17. 30

Wu, W., Tang, X., Lv, J., Yang, C., and Liu, H. (2021). Potential of Bayesian additive regression trees for predicting daily global and diffuse solar radiation in arid and humid areas. *Renewable Energy*, 177:148–163. 3, 129

Wu, Y., Tjelmeland, H., and West, M. (2007). Bayesian CART: prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66. 26

Xie, G., Chen, X., and Weng, Y. (2018). An integrated Gaussian process modeling framework for residential load prediction. *IEEE Transactions on Power Systems*, 33(6):7238–7248. 26

Yee, R. and Deshpande, S. K. (2023). Evaluating plate discipline in Major League Baseball with Bayesian additive regression trees. *Journal of Quantitative Analysis in Sports*. advanced online publication. 80

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American statistical Association*, 93(441):228–237. 156

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368. 79, 81, 88

Zhang, T., Geng, G., Liu, Y., and Chang, H. H. (2020). Application of Bayesian additive regression trees for estimating daily concentrations of PM2.5 components. *Atmosphere*, 11(11):1233. 24

Zhang, X. (2020). Parameter-expanded data augmentation for analyzing correlated binary data using multivariate probit models. *Statistics in Medicine*, 39(25):3637–3652. 94, 99

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896. 100

Zhang, X., Boscardin, W. J., Belin, T. R., Wan, X., He, Y., and Zhang, K. (2015). A Bayesian method for analyzing combinations of continuous, ordinal, and nominal categorical data with missing values. *Journal of Multivariate Analysis*, 135:43–58. 100, 117

Zhao, Y., Zheng, W., Zhuo, D. Y., Lu, Y., Ma, X., Liu, H., Zeng, Z., and Laird, G. (2018). Bayesian additive decision trees of biomarker by treatment interactions for predictive biomarker detection and subgroup identification. *Journal of Biopharmaceutical Statistics*, 28(3):534–549. 24